Basic Econometrics

Riccardo (Jack) Lucchetti

21st April 2024

Foreword

This is a very basic course in econometrics, in that it only covers basic techniques, although I tried to avoid the scourge of over-simplification, so some may find it not so basic in style. What makes it perhaps a little different from others you find on the Net is that I made a few not-so-common choices.

1. Separating clearly the properties OLS has by construction from those it has when interpreted as an estimator.

2. Using matrix algebra whenever possible.

3. Using asymptotic inference only.

Point number one is modelled after the ideas in the two great masterpieces, Davidson and MacKinnon (1993) and Davidson and MacKinnon (2004). I have several reasons for this choice, but it is mainly a pedagogical one. The students I am writing for are people who often don't feel at ease with the tools of statistical inference: they have learned the properties of estimators by heart, they are not sure they can read a test, find the concept of the distribution of a statistic a little unclear (never mind asymptotic distributions), get confused between the variance of an estimator and an estimator of the variance. In the best cases. Never mind; no big deal.

There's an awful lot you can say on the base tool in econometrics (OLS) even without all this, and that's good to know. Once a student has learned how to handle OLS properly as a mere computational tool, the issues of its usage and interpretation as an estimator and of how to read the associated test statistics can be grasped more correctly. If you mix the two aspects too early, a beginner is prone to mistake properties of least squares that are true by construction for properties that depend on some probabilistic assumptions.

Point number two is motivated by laziness. In my teaching career, I have found that once students get comfortable with matrices, my workload halves. Of course, it takes some initial sunk cost to convey properly ideas such as projections and properties of quadratic forms, but the payoff is very handsome. This book contains no systematic account of matrix algebra; we're using just the basics, so anything you find on the Net by googling "matrix algebra lecture notes" is probably good enough.

As for probability and statistics, I will only assume some familiarity with the very basics: simple descriptive statistics and basic properties of probability, random variables and expectations. Chapter 2 contains a cursory treatment of the concepts I will use later, but I wouldn't recommend it as a general reference on the subject. Its purpose is mainly to make the notation explicit and clarify a few points. For example, I will avoid any kind of reference to maximum likelihood methods.

I don't think I have to justify point number three. I am writing this in 2024, when typical data sets have hundreds, if not thousands observations and nobody would ever dream of running any kind of inferential procedure with less than 50 data points. Apart from OLS, there is no econometric technique in actual use that does not depend vitally on asymptotics, so I guess that readers should get familiar with the associated concepts if there is a remote chance that this will not be put them off econometrics completely. The t test, the F tests and, in general, all kinds of degrees-of-freedom corrections are ad-hockeries of the past; unbiasedness is overrated. Get over it.

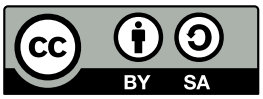
I promise I'll try to be respectful of the readers and don't treat them like idiots. I assume that if you're reading this, you want to know more than you do about econometrics, but this doesn't give me the right to assume that you need to be taken by the hand and treated like an 11-year-old.

All the examples and scripts in this book are replicable. All the material is in a zip file you can download from this link. The software I used throughout the book is gretl, so data and scripts are in gretl format, but if you insist on using inferior software (`;-)`), data are in CSV format too.

Finally, a word of gratitude. A book like this is akin to a software project, and there's always one more bug to fix. So, I'd like to thank first all my students who helped me eradicate quite a few. Then, my colleagues Allin Cottrell, Stefano Fachin, Francesca Mariani, Giulio Palomba, Luca Pedini, Matteo Picchio, Claudia Pigini, Alessandro Pionati, Alessandro Sterlacchini and Francesco Valentini for making many valuable suggestions. Needless to say, the remaining shortcomings are all mine. Claudia also allowed me to grab a few things from her slides on IV estimation, so thanks for that too. If you want to join the list, please send me bug reports and feature requests. Also, I'm not an English native speaker (I suppose it shows). So, Anglophones of the world, please correct me whenever needed.

The structure of this book is as follows: chapter 1 explores the properties of OLS as a descriptive statistic. Inference comes into play at chapter 2 with some general concepts, while their application to OLS is the object of chapter 3, with some basic ideas on diagnostic testing and heteroskedasticity in Chapter 4. Extension of basic OLS are considered in the subsequent chpater: Chapter 5 deals with dynamic models chapter 6 with instrumental variable estimation and finally, Chapter 7 considers linear models for panel data. Each chapter has an appendix, named "Assorted results", where I discuss some of the material I use during the chapter in a little more detail.

In some cases, I will use a special format for short pieces of texts, like this. They contain extra stuff that I consider interesting, but not indispensable for the overall comprehension of the main topic.



Contents

Chapter 1

OLS: algebraic and geometric properties

1.1 Models

I won't even attempt to give the reader an account of the theory of econometric modelling. For our present purposes, suffice it to say that we econometricians like to call a **model** a mathematical description of something, that doesn't aim at being 100% accurate, but still, hopefully, useful.[1]

We have a quantity of interest, also called the **dependent variable**, which we observe more than once: a collection of numbers $y_1, y_2, \ldots, y_n$, where n is the size of our data set. These numbers can be anything that can be given a coherent numerical representation; in this course, however, we will confine ourselves to the case where the i-th observation y_i is a real number. So for example, we could record the income for n individuals, the export share for n firms, the inflation rate for a given country at n points in time.

Now suppose that, for each data point, we also have a vector of k elements containing auxiliary data possibly helpful in better understanding the differences between the y_is; we call these **explanatory variables**,[2] or $\mathbf{x}_i$ in symbols.[3] To continue the previous examples, $\mathbf{x}_i$ may include a numerical description of the individuals we recorded the income of (such as age, gender, educational attainment and so on), or characteristics of the firms we want to study the export propensity for (size, turnover, R&D expenditure and so on), or the conditions of the economy at the time the inflation rate was recorded (interest rate, level of output, and so forth).

[1] "All models are wrong, but some are useful" (G. E. P. Box). In fact, one may argue that, in order to be useful, a model may *have* to be inaccurate. More on this in section 1.4.2.

[2] Terminology is very much field-specific here; statisticians traditionally tend to use the term *covariates*, while people from the machine learning community like the word *features*.

[3] I will almost always use boldface symbols to indicate vectors.

What we call a model is a formula like the following:

$$y_i \simeq m(\mathbf{x}_i)$$

where we implicitly assume that if $\mathbf{x}_i$ is not too different from $\mathbf{x}_j$, then we should expect y_i to be broadly close to y_j: if we pick two people of the same age, with the same educational level and many other characteristics in common we would expect that their income should be roughly the same. Of course this won't be true in all cases (in fact, chances are that this will *never* be true exactly), but hopefully our model won't lead us to catastrophic mistakes.

The reason why we want to build models is that, once the function $m(\cdot)$ is known, it becomes possible to ask ourselves interesting questions by inspecting the characteristics of that function. So for example, if it turned out that the export share of a firm is increasing in the expenditure in R&D, we may make conjectures about the reasons why it should be so, look for some economic theory that could explain the result, and wonder if one could improve export competitiveness by giving the firms incentives to do research.

Moreover, the door is open to forecasting: given the characteristics of a hypothetical firm or individual, the model makes it possible to guess what their export share or income (respectively) should be. I don't think I have to convince the reader of how useful this could be in practice.

Of course, we will want to build our model in the best possible way. In other words, our aim will be choosing the function $m(\cdot)$ according to some kind of optimality criterion. This is what the present course is about.

But there's more: as we will see, building an optimal model is impossible in general. At most, we may hope to build the best possible model *for the data that we have available*. Of course, there is no way of knowing if the model we built, that perhaps works rather well with our data, will keep working equally well with new data. Imagine you built a model for the inflation rate in a country with monthly data from January 2000 to December 2017. It may well be that your model performs (or, as we say, "fits the data") very well for that period, but what guarantee do you have that it will keep doing so in 2018, or in the more distant future? The answer is: you have none. But still, this is something that we'd like to do; our mind has a natural tendency to generalise, to infer, to extrapolate. And yet, there is no logical compelling basis for proving that it's a good idea to do so.[4] The way out is framing the problem in a probabilistic setting, and this is the reason why econometrics is so intimately related with probability and statistics.

For the moment, we'll start with the problem of choosing $m(\cdot)$ in a very simple case, that is when we have no extra information $\mathbf{x}_i$. In this case, the function becomes a constant:

$$y_i \simeq m(\mathbf{x}_i) = m$$

and the problem is very much simplified, because it means we have to pick a number m in some optimal way, given the data $y_1, y_2, \ldots, y_n$. In other words, we

[4] 'The philosophically inclined reader may at this point google for "Bertrand Russel's turkey".

have to find a function of the data which returns the number m. Of course, a function of the data is what we call a statistic. In the next section, I will prove that the statistic we're looking for is, in this case, the average of the y_is, that is $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

1.2 The average

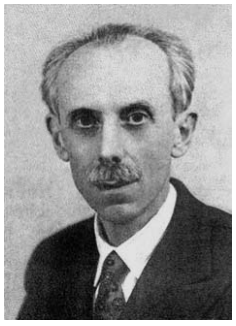
What is a descriptive statistic? It is a function of the data which synthesises a particular feature of interest of the data; of course, the more informative, the better. The idea behind descriptive statistics is more or less: we have some data on some real-world phenomenon; our data set, unfortunately, is too "large", and we don't have time/can't/don't feel like going through the whole thing. Hence, we are looking for a function of these data to tell us what we want, without being bothered with unnecessary details.

The most obvious example of a descriptive statistic is, of course, the sample average. Let's stick our observations $y_1, y_2, \ldots, y_n$ into a column vector $\mathbf{y}$; the sample average is nothing but

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\boldsymbol{\iota}'\mathbf{y}, \tag{1.1}$$

where $\boldsymbol{\iota}$ is a column vector full of ones. The "sum" notation is probably more familiar to most readers; I prefer the matrix-based one not only because I find it more elegant, but also because it's far easier to generalise. The nice feature of the vector $\boldsymbol{\iota}$ is that its inner product with any conformable vector $\mathbf{x}$ yields the sum of the elements of $\mathbf{x}$.[5]

We use averages all the time. Why is the average so popular? As I said, we're looking for a descriptive statistic m, as a synthesis of the information contained in our data set.



OSCAR CHISINI

In 1929, Oscar Chisini (pronounced kee-*zee*-nee) proposed the following definition: for a function of interest $g(\cdot)$, the mean of the vector $\mathbf{y}$ is the number m that yields the unique solution to $g(\mathbf{y}) = g(m \cdot \boldsymbol{\iota})$. Powerful idea: for example, the average is the solution of the special case when the $g(\cdot)$ function is the sum of the vector's elements, and the reader may want to spend some time with more exotic cases.

Chisini's idea may be further generalised: if our aim is to use m — that we haven't yet chosen — as an imperfect

[5] Reminder: the inner products of two vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as $\sum_i a_i b_i$. Mathematicians like the notation $\langle \mathbf{a}, \mathbf{b} \rangle$ for the inner product, on the grounds of its greater generality (google "Hilbert space" if you're curious), but we econometricians are more accustomed to the "matrix" notation $\mathbf{a}'\mathbf{b}$, where the apostrophe means "transposed".

but parsimonious description of the whole data set, the question that naturally arises is: how much information is lost?

If all we knew, for a given data set, was m, what could we say about each single observation? If we lack any more information, the most sensible thing to say is that, for a generic i, y_i should more or less be m. Consider the case of A. S. Tudent, who belongs to a class for which the "typical" grade in econometrics is 23;[6] the most sensible answer we could give to the question "What was the grade that A. S. Tudent got in econometrics?" would be "Around 23, I guess". If the actual grade that A. S. got were in fact 23, OK. Otherwise, we could measure by how much we were wrong by taking the difference between the actual grade and our guess, $e_i = y_i - m$. We call these quantities the **residuals**; the vector of residuals is, of course, $\mathbf{e} = \mathbf{y} - \boldsymbol{\iota} \cdot m$.

In the ideal case, using m to summarise the data should entail no information loss at all, and the difference between y_i and m should be 0 for all i (all students got 23). If it weren't so, we may measure how well m does its job through the size of the residuals. Let's define a function, called **loss function**, which measures the cost we incur because of the information loss.

$$L(m) = C[\mathbf{e}(m)]$$

In principle, there are not many properties such a function should be assumed to have. It seems reasonable that $C(\mathbf{0}) = 0$:[7] if all the residuals are 0, no approximation errors occur and the cost is nil. Another reasonable idea is $C(\mathbf{e}) \geq 0$: you can't gain from a mistake.[8] Apart from this, there is not very much you can say: the $L(\cdot)$ function cannot be assumed to be convex, or symmetric, or anything else. It depends on the context.

Whatever the shape of this function, however, we'll want to choose m so that is $L(m)$ as small as possible. In math-speak: for a given problem, we can write down the loss function and choose the statistic which minimises it. In formulae:

$$\hat{m} = \underset{m \in \mathbb{R}}{\operatorname{Argmin}}\ L(m) = \underset{m \in \mathbb{R}}{\operatorname{Argmin}}\ C(\mathbf{y} - \boldsymbol{\iota} \cdot m), \tag{1.2}$$

where you read the above as: m with a hat on is that number that you find if you choose, among all real numbers, the one that makes the function $L(m)$ as small as possible.

In practice, by finding the minimum of the $L(\cdot)$ function for a given problem, we can be confident that we are using our data in the best possible way. At this point, the first thing that crosses a reasonable person's mind is "How do I choose $L(\cdot)$? I mean, what should it look like?". Fair point. Apart from extraordinary cases when the loss function is a natural consequence of the problem itself,

[6] Note for international readers: in the Italian academic system, which is what I'm used to, grades go from 18 (barely pass) to 30 (full marks).

[7] I use a boldface 0 to indicate a vector full of zeros, as in $0 \cdot \boldsymbol{\iota} = \mathbf{0}$.

[8] Warning: the converse is not necessarily true. It's possible that the cost is nil even with non-zero errors. For example, in some contexts "small" error may be irrelevant.

writing down its exact mathematical form may be complicated. What does the $L(m)$ function look like for the grades in econometrics of our hypothetical class? Hard to say.

Moreover, we often must come up with a summary statistic without knowing in advance what it will be used for. Obviously, in these cases finding a one-size-fits-all optimal solution is downright impossible. We have to make do with something that is not too misleading. A possible choice is

$$L(m) = \sum_{i=1}^{n} (y_i - m)^2 = (\mathbf{y} - \boldsymbol{\iota} \cdot m)'(\mathbf{y} - \boldsymbol{\iota} \cdot m) = \mathbf{e}'\mathbf{e} \tag{1.3}$$

The above criterion is a function of m based on the sum of squared residuals, that enjoys several desirable properties. Not only it's simple to manipulate algebraically: it's symmetric and convex, so that positive and negative residuals are penalised equally, and large errors are more costly than small ones. It's not unreasonable to take this loss function as an acceptable approximation. Moreover, this choice makes it extremely easy to solve the associated minimisation problem.

Minimising $L(m)$ with respect to m leads to the so-called **least squares** problem. All is needed to find the minimum in (1.3) is taking the derivative of $L(m)$ with respect to m;

$$\frac{\mathrm{d}L(m)}{\mathrm{d}m} = \sum_{i=1}^{n} \frac{\mathrm{d}\left(y_i - m\right)^2}{\mathrm{d}m} = -2 \sum_{i=1}^{n} \left(y_i - m\right)$$

The derivative must be 0 for a minimum, so that

$$\sum_{i=1}^{n} \left(y_i - \hat{m}\right) = 0$$

which in turn implies

$$n \cdot \hat{m} = \sum_{i=1}^{n} y_i$$

and therefore $\hat{m} = n^{-1} \sum_{i=1}^{n} y_i = \bar{Y}$. The reader is invited to verify that $\hat{m}$ is indeed a minimum, by checking that the second derivative $\frac{\mathrm{d}^2 L(m)}{\mathrm{d}m^2}$ is positive. Things are even smoother in matrix notation:

$$L(m) = (\mathbf{y} - \boldsymbol{\iota} m)'(\mathbf{y} - \boldsymbol{\iota} m) = \mathbf{y}'\mathbf{y} - 2m \cdot \boldsymbol{\iota}'\mathbf{y} + m^2 \boldsymbol{\iota}'\boldsymbol{\iota},$$

so the derivative is

$$\frac{\mathrm{d}L(m)}{\mathrm{d}m} = -2\boldsymbol{\iota}'\mathbf{y} + 2m \cdot \boldsymbol{\iota}'\boldsymbol{\iota} = -2\boldsymbol{\iota}'(\mathbf{y} - \boldsymbol{\iota} m) = 0$$

whence

$$\boldsymbol{\iota}'\mathbf{y} = (\boldsymbol{\iota}'\boldsymbol{\iota}) \cdot \hat{m} \Longrightarrow \hat{m} = (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \boldsymbol{\iota}'\mathbf{y} = \bar{Y}$$

because of course $\boldsymbol{\iota}'\boldsymbol{\iota} = n$. The value of $L(m)$ at the minimum, that is $L(\hat{m}) = \mathbf{e}'\mathbf{e} = \sum_{i=1}^{n}(y_i - \bar{Y})^2$ is a quantity that in this case we call **deviance**, but that we will more often call **SSR**, as in **Sum of Squared Residuals**.

The mathematically sophisticated way to say the same, that we used a few pages back, is

$$\hat{m} = \underset{m\in\mathbb{R}}{\text{Argmin}}\ L(m);$$

where again, the hat ($\hat{}$) on m indicates that, among all possible real numbers, we are choosing the one that minimises our loss function.

The argument above, which leads to choosing the average as an optimal summary is, in fact, much more general than it may seem: many of the descriptive statistics we routinely use are special cases of the average, where the data $\mathbf{y}$ are subject to some preliminary transformation. In practice: the average of $\mathbf{z}$, where $z_i = h(y_i)$ can be very informative, if we choose the function $h(\cdot)$ wisely. The variance is the most obvious example: the sample variance[9] is just the average of $z_i = (y_i - \bar{Y})^2$, which measures how far y_i is from $\bar{Y}$.

Things get even more interesting when we express a frequency as an average: define the event $E = \{y_i \in A\}$, where A is some subset of the possible values for y_i; now define the variable $z_i = \mathbb{I}(y_i \in A)$, where $\mathbb{I}(\cdot)$ is the so-called "indicator function", that gives 1 when its argument is true and 0 when false. Evidently, the average of the z_i, $\overline{Z}$, is the relative frequency of E:

$$\overline{Z} = \frac{\sum_{i=1}^{n} z_i}{n} = K/n;$$

since z_i can only be 0 or 1, $K = \sum_{i=1}^{n} z_i$ is just the number of times the event E has occurred. I'm sure you can come up with more examples.

1.3 OLS as a descriptive statistic

1.3.1 OLS on a dummy variable

Now let's bring the explanatory variables $\mathbf{x}_i$ back in. For the moment, let's consider the special case where $\mathbf{x}_i$ is a one-element vector, that is a scalar.

A possible way to check if y_i and x_i are related to each other is to see if y_i is "large" or "small" when x_i is "large" or "small". Define

$$z_i = (y_i - \bar{Y})(x_i - \bar{X})$$

which is, in practice, a sort of indicator of "matching magnitudes": z_i is positive when $y_i > \bar{Y}$ and $x_i > \bar{X}$ (both are "large") or when $y_i < \bar{Y}$ and $x_i < \bar{X}$ (both are "small"); on the contrary, z_i is negative when magnitudes don't match. As

[9] I'm not applying the "degrees of freedom correction"; I don't see why I should, as long I'm using the variance as a descriptive statistic.

is well known, the average of z_i is known as covariance; but this is just boring elementary statistics.

The reason why I brought this up is to highlight the main problem with covariance (and correlation, that is just covariance rescaled so that it's guaranteed to be between -1 and 1): it's a symmetric concept. The variables y_i and x_i are treated equally: the covariance between y_i and x_i is by construction the same as between x_i and y_i. On the contrary, we often like to think in terms of $y_i = m(x_i)$, because what we have in mind is an interpretation where y_i "depends" on x_i, and not the other way around.[10] This is why we call y_i the **dependent variable** and x_i the **explanatory variable**. In this context, it's rather natural to see what happens if you split **y** into several sub-vectors, according to the values that x_i takes. In a probabilistic context, we'd call this **conditioning** (see section 2.2.2).

Simple example: suppose our vector **y** includes observations on n people, with n_m males and $n_f = n - n_m$ females. The information on gender is in a variable x_i, that equals 1 for males and 0 for females. As is well known, a 0/1 variable may be called "binary", "Boolean", "dichotomic", but we econometricians traditionally call it a **dummy** variable.[11]

Common sense suggests that, if we take into account the information we have on gender, the average *by gender* will give us a data description which should be slightly less concise than overall average (since we're using two numbers instead of one), but certainly not less accurate. Evidently, we can define

$$\bar{Y}_m = \frac{\sum_{x_i=1} y_i}{n_m} = \frac{S_m}{n_m} \qquad \bar{Y}_f = \frac{\sum_{x_i=0} y_i}{n_f} = \frac{S_f}{n_f}$$

where S_m and S_f are the sums of y_i for males and females, respectively.

Now, everything becomes more elegant and exciting if we formalise the problem in a similar way to what we did with the average. We would like to use in the best possible way the information (that we assume we have) on the gender of the i-th individual. So, instead of summarising the data by a *number*, we are going to use a *function*, that is something like

$$m(x_i) = m_m \cdot x_i + m_f \cdot (1 - x_i)$$

which evidently equals m_m for men (since $x_i = 1$) and m_f for women (since $x_i = 0$). Our summary will be a rule giving us 'representative' values of y_i according to x_i.

Let's go back to our definition of residuals as approximation errors: in this case, you clearly have that $e_i \equiv y_i - m(x_i)$, and therefore

$$y_i = m_m x_i + m_f (1 - x_i) + e_i \tag{1.4}$$

[10] I'm being deliberately vague here: in everyday speech, saying that A depends on B may mean many things, not necessarily consistent. For example, "dependence" may not imply a cause-effect link. This problem is much less trivial than it seems at first sight, and we'll leave it to professional epistemologists.

[11] I am aware that there are people who don't fit into the tradtional male/female distinction, and I don't mean to disrespect them. Treating gender as a binary variable just makes for a nice and simple example here, ok?

Equation (1.4) is a simple example of an econometric model. The number y_i is split into two additive components: a systematic part, that depends on the variable x_i (a linear function of x_i, to be precise), plus a remainder term, that we just call the residual for now. In this example, $m(x_i) = m_m x_i + m_f(1 - x_i)$.

It is convenient to rewrite (1.4) as

$$y_i = m_f + (m_m - m_f)x_i + e_i = \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} m_f \\ m_m - m_f \end{bmatrix} + e_i$$

so we can use matrix notation, which is much more compact and elegant

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1.5}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} m_f \\ m_m - m_f \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

and $\mathbf{X}$ is a matrix with n rows and 2 columns; the first column is $\boldsymbol{\iota}$ and the second one is $\mathbf{x}$. The i-th row of $\mathbf{X}$ is $[1,1]$ if the corresponding individual is male and $[1,0]$ otherwise. To be explicit:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n-1} \\ e_n \end{bmatrix}$$

Therefore, the problem of choosing m_m and m_f optimally is transformed into the problem of finding the vector $\boldsymbol{\beta}$ that minimises the loss function $\mathbf{e}'\mathbf{e}$. The solution is not difficult: find the solutions to[12]

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}}\mathbf{e}'\mathbf{e} = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}}(\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = 0$$

By using the well-known[13] rules for matrix differentiation, you have

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X} \cdot \hat{\boldsymbol{\beta}} \tag{1.6}$$

What we have to do now is solve equation (1.6) for $\hat{\boldsymbol{\beta}}$. The solution is unique if $\mathbf{X}'\mathbf{X}$ is invertible (if you need a refresher on matrix inversion, and related matters, subsection 1.A.3 is for you):

$$\boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.} \tag{1.7}$$

[12] Need I remind the reader of the rule for transposing a matrix product, that is $(AB)' = B'A'$? Obviously not.

[13] Not so well-known, maybe? Jump to subsection 1.A.1.

Equation (1.7) is the single most important equation in this book, and this is why I framed it into a box. The vector $\hat{\boldsymbol{\beta}}$ is defined as the vector that minimises the sum of squared residuals among all vectors with k elements (where $k = 2$ in this case):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\text{Argmin}} \; \mathbf{e}'\mathbf{e},$$

and the expression in equation (1.7) turns the implicit definition into an explicit formula that you can use to calculate $\hat{\boldsymbol{\beta}}$.

The coefficients $\hat{\boldsymbol{\beta}}$ obtained from (1.7) are known as **OLS coefficients**, or **OLS statistic**, from **Ordinary Least Squares**.[14] A very common idiom that economists use when referring to the calculation of OLS is "regressing $\mathbf{y}$ on $\mathbf{X}$". The usage of the word "regression" here might seem odd, but will be justified in chapter 3.

The "hat" symbol has exactly the same meaning as in eq. (1.2): of all the possible choices for $\boldsymbol{\beta}$, we pick the one that makes eq. (1.6) true, and therefore minimises the associated loss function $\mathbf{e}'\mathbf{e}$. The vector

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

is our approximation to $\mathbf{y}$. The elements of $\hat{\mathbf{y}}$ are customarily called the **fitted values**: the closer they are to $\mathbf{y}$, the better we say that the model fits the data.

In this example, a few simple calculations suffice to show that

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{bmatrix} n & n_m \\ n_m & n_m \end{bmatrix} \\ \mathbf{X}'\mathbf{y} &= \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{x_i=1} y_i \end{bmatrix} = \begin{bmatrix} S_m + S_f \\ S_m \end{bmatrix} \end{aligned}$$

where $S_m = \sum_{x_i=1} y_i$ and $S_f = \sum_{x_i=0} y_i$: the sums of y_i for males and females, respectively. By using the standard rule for inverting (2×2) matrices, which I will also assume known,[15]

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n_m n_f} \begin{bmatrix} n_m & -n_m \\ -n_m & n \end{bmatrix}$$

so that

$$\hat{\boldsymbol{\beta}} = \frac{1}{n_m n_f} \begin{bmatrix} n_m & -n_m \\ -n_m & n \end{bmatrix} \begin{bmatrix} S_m + S_f \\ S_m \end{bmatrix} = \frac{1}{n_m n_f} \begin{bmatrix} n_m S_f \\ n_f S_m - n_m S_f \end{bmatrix}$$

and finally

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \frac{S_f}{n_f} \\ \frac{S_m}{n_m} - \frac{S_f}{n_f} \end{bmatrix} = \begin{bmatrix} \bar{Y}_f \\ \bar{Y}_m - \bar{Y}_f \end{bmatrix}$$

[14] Why "ordinary"? Well, because there are more sophisticated variants, so we call these "ordinary" as in "not extraordinary". We'll see one of those variants in section 4.2.1.

[15] If you're in trouble, go to subsection 1.A.4.

so that our model is:

$$\hat{y}_i = \bar{Y}_f + \left(\bar{Y}_m - \bar{Y}_f\right) x_i$$

and it's easy to see that the fitted value for males ($x_i = 1$) is $\bar{Y}_m$, while the one for the females ($x_i = 0$) is $\bar{Y}_f$.

Example 1.1

Let me give you a numerical example of the above: suppose we have 80 individuals (50 males and 30 females) and that we're interested in their monthly wage. Moreover, $S_m = \sum_{x_i=1} y_i$ = € 60000 and $S_f = \sum_{x_i=0} y_i$ = € 42000: therefore, the average wage is $\bar{Y}_m$ = 1200 = 60000/50 for males and $\bar{Y}_f$ = 1400 = 42000/30 for females. After ordering observations by putting the data for males first,[16] the $\mathbf{X}$ *matrix looks like*

$$\mathbf{X} = \left[\begin{array}{cc} 1 & 1 \\ 1 & 1 \\ \vdots & \\ 1 & 1 \\ \hline 1 & 0 \\ \vdots & \\ 1 & 0 \end{array}\right]$$

where the top block of rows has 50 rows and the bottom one has 30. As the reader may easily verify,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 80 & 50 \\ 50 & 50 \end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 102000 \\ 60000 \end{bmatrix}$$

By performing the appropriate calculations, one finds that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} = \begin{bmatrix} 1400 \\ -200 \end{bmatrix}$$

and the model can be written as:

$$\hat{y}_i = 1400 - 200 x_i,$$

which reads: for females, $x_i = 0$, so their typical income is €1400; for males, instead, $x_i = 1$, so their income is given by $1400 - 200 \cdot 1$ = €1200.

Once again, opting for a quadratic loss function (and therefore minimising $\mathbf{e}'\mathbf{e}$) delivers a solution consistent with common sense, and our approximate description of the vector $\mathbf{y}$ uses a function whose parameters are the statistics we are interested in.

[16] With no loss of generality, as a mathematician would say.

1.3.2 The general case

In reading the previous subsection, the discerning reader will have noticed that, in fact, the assumption that $\mathbf{x}$ is a dummy variable plays a very marginal role. There is no reason why the equation $m(x_i) = \beta_1 + \beta_2 x_i$ should not hold when x_i contains generic numeric data. The solution to the problem remains unchanged; clearly, the vector $\hat{\boldsymbol{\beta}}$ will not contain the averages by sub-samples, but the fact that the loss function is minimised by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ keeps being true.

Example 1.2

Suppose that

$$\mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 3 \\ 0 \\ 1 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 4 \\ 3 \\ 2 \\ 5 \\ 1 \\ 1 \end{bmatrix}$$

The reader is invited to check that[17]

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 16 \\ 16 & 56 \end{bmatrix} \Rightarrow (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.7 & -0.2 \\ -0.2 & 3/40 \end{bmatrix} \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 10 \\ 33 \end{bmatrix}$$

and therefore

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0.4 \\ 0.475 \end{bmatrix} \qquad \hat{\mathbf{y}} = \begin{bmatrix} 2.3 \\ 1.825 \\ 1.345 \\ 2.775 \\ 0.875 \\ 0.875 \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} -1.3 \\ 1.175 \\ 0.65 \\ 0.225 \\ -0.875 \\ 0.125 \end{bmatrix}$$

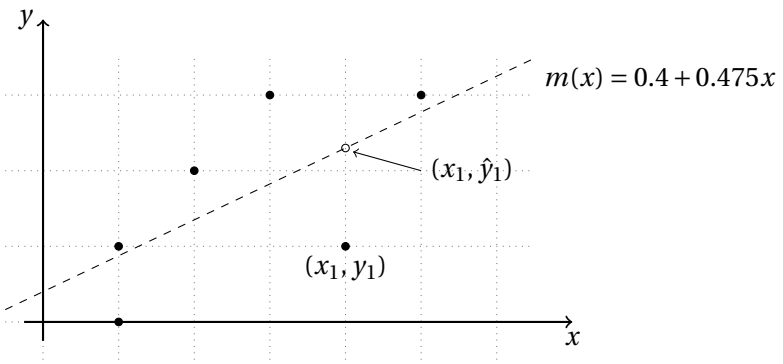
Hence, the function $m(x_i)$ minimising the sum of squared residuals is $m(x_i) = 0.4 + 0.475x_i$ and $\mathbf{e}'\mathbf{e}$ equals 4.325.

In traditional textbooks, at this point you always get a picture similar to the one in Figure 1.1, which is supposed to aid intuition; I don't like it very much, and will explain why shortly. Nevertheless, let me show it to you: in this example, we use the same data as in the present example.

In Figure 1.1, each black dot corresponds to a (x_i, y_i) pair; the dashed line plots the $m(x)$ function and the residuals are *the vertical differences between the dots and the dashed line*; the least squares criterion makes the line go through the dots in such a way that the sum of these differences (squared) is minimal. So, for example, for observation number 1 the observed value of x_i is 4 and the

[17] Before you triumphantly shout "It's wrong!", remember to stick $\boldsymbol{\iota}$ and $\mathbf{x}$ together.

Figure 1.1: OLS on six data points



observed observed value for y_i is 1; the approximation yields $\hat{y}_1 = 0.4+0.475\times 4 = 2.3$ (observe the position of the white dot). Therefore, $e_1 = y_1 - \hat{y}_1 = -1.3$ (the vertical distance between the black dot and the white dot).

The example above can be generalised by considering the case where we have more than one explanatory variable, except for the fact that producing a figure akin to Figure 1.1 becomes difficult, if not impossible. Here, the natural thing to do is expressing our approximation as a function of the vector $\mathbf{x}_i$, that is $m(\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$, or more explicitly,

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

For example, suppose we have data on each student in the class A. S. Tudent belongs to. How many hours each student spent studying econometrics, their previous grades in related subjects, and so on; these data, for the i-th student, are contained in the vector $\mathbf{x}_i'$, which brings us back to equation (1.5).

The algebraic apparatus we need for dealing with the generalised problem is, luckily, unchanged; allow me to recap it briefly. If the residual we use for minimising the loss function is $e_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, then the vector of residuals is

$$\mathbf{e}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \tag{1.8}$$

so the function to minimise is $L(\boldsymbol{\beta}) = \mathbf{e}(\boldsymbol{\beta})'\mathbf{e}(\boldsymbol{\beta})$.

Since the derivative of $\mathbf{e}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is $-\mathbf{X}$, we can use the chain rule and write

$$\mathbf{X}'\mathbf{e}(\hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{1.9}$$

(a more detailed proof, should you need it, is in subsection 1.A.5). By putting together (1.8) and (1.9) you get a system of equations sometimes referred to as **normal equations**:

$$\mathbf{X}'\mathbf{X}\cdot\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{1.10}$$

and therefore, if $\mathbf{X}'\mathbf{X}$ is invertible, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

CARL FRIEDRICH GAUSS

If you think that all this is very clever, well, you're right. The inventor of OLS is arguably the greatest mathematicians of all time: the great Carl Friedrich Gauss, also known as the *princeps mathematicorum*.[18]

Note, again, that the average can be obtained as the special case when $\mathbf{X} = \boldsymbol{\iota}$. Moreover, it's nice to observe that the above formulae make it possible to compute all the relevant quantities without necessarily observing the matrices $\mathbf{X}$ and $\mathbf{y}$; in fact, all the elements you need are the following:

1. the scalar $\mathbf{y}'\mathbf{y}$;
2. the k-element vector $\mathbf{X}'\mathbf{y}$ and
3. the $k \times k$ matrix $\mathbf{X}'\mathbf{X}$ (or equivalently, its inverse).

where k is the number of columns of $\mathbf{X}$, the number of unknown coefficients in our $m(\cdot)$ function. Given these quantities, $\hat{\boldsymbol{\beta}}$ is readily computed, but also $\mathbf{e}'\mathbf{e}$:

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}$$

and using (1.10) you have

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{y}). \tag{1.11}$$

Equation (1.11) expresses the SSR as the difference between a scalar and the inner product of two k-element vectors $\hat{\boldsymbol{\beta}}$ and $(\mathbf{X}'\mathbf{y})$. The number of *rows* of $\mathbf{y}$, that is the number of observation n, never comes into play, and could well be huge.

I guess you now understand my lack of enthusiasm for Figure 1.1: if $\mathbf{X}$ has 3 columns, drawing a similar picture is difficult. For 4 or more columns, it becomes impossible. Worse, the geometric intuition that it conveys may overlap with another geometric interpretation of OLS, which I consider more interesting and more useful, and is the object of section 1.4.

A nice feature of a linear function like (1.5) is that the coefficients $\boldsymbol{\beta}$ can be interpreted as **marginal effects**, or partial derivatives if you prefer. In the previous example, the coefficient associated to the number of hours that each student spent studying econometrics may be defined as

$$\frac{\partial m(\mathbf{x})}{\partial x_j} = \beta_j \tag{1.12}$$

and therefore can be read as the partial derivative of the $m(\cdot)$ function with respect to the number of hours. Clearly, you may attempt to interpret these magnitude by their sign (do more hours of study improve your grade?) and by their

[18] To be fair, the French mathematician Adrien-Marie Legendre rediscovered it independently a few years later.

magnitude (if so, by how much?). However, you should resist the temptation to give the coefficients a counterfactual interpretation (If A. S. Tudent had studied 2 more hours, instead of watching that football game, by how much would their mark have improved?); this is possible, in some circumstances, but not always (more on this in Section 3.6).

Focusing on marginal effects is what we do most often in econometrics, because the question of interest is not really approximating y given $\mathbf{x}$, but rather understanding what the effect of $\mathbf{x}$ on y is (and, possibly, how general and robust this effect is). In other words, the object of interest in econometrics is much more often $\boldsymbol{\beta}$, rather than $m(\mathbf{x})$. The opposite happens in a broad class of statistical methods that go, collectively, by the name of *machine learning* methods and focus much more on *prediction* than *interpretation*. In order to predict correctly, these models use much more sophisticated ways of handling the data than a simple linear function, and even writing the rule that links $\mathbf{x}$ to $\hat{y}$ is impossible.

Machine learning tools have been getting quite popular at the beginning of the XXI century, and are the tools that companies like Google and Amazon use to predict what video you'd like to see on Youtube or what book you'd like to buy when you open their website. As we all know, these models perform surprisingly well in practice, but nobody would be able to reconstruct *how* their predictions come about. The phrase some people use is that machine learning procedures are "black boxes": they work very well, but they don't provide you with an explanation of why *you* like that particular video. The pros and cons of econometric models versus machine learning tools are still under scrutiny by the scientific community, and, if you're curious, I'll just give you a pointer to Mullainathan and Spiess (2017).

1.3.3 Collinearity and the dummy trap

Of course, for solving equation (1.10), $\mathbf{X}'\mathbf{X}$ must be invertible. Now, you may ask: what if it's singular? This is an interesting case. The solution $\hat{\mathbf{y}}$ can still be found, but there is more than one vector $\hat{\boldsymbol{\beta}}$ associated with it. In fact, there are infinitely many. Let me give you an example. Suppose that $\mathbf{X}$ contain only one non-zero column, $\mathbf{x}_1$. The solution is easy to find:

$$\hat{\beta}_1 = \frac{\mathbf{x}_1'\mathbf{y}}{\mathbf{x}_1'\mathbf{x}_1},$$

so that $\hat{\mathbf{y}} = \beta_1\mathbf{x}_1$. Now, add to $\mathbf{X}$ a second column, $\mathbf{x}_2$, which happens to be equal to $\mathbf{x}_1$, so $\mathbf{x}_2 = \mathbf{x}_1$. Evidently, $\mathbf{x}_2$ adds no information to our model, because it contains exactly the same information as $\mathbf{x}_1$ so $\hat{\mathbf{y}}$ remains the same. Now, however, we can write it in infinitely may ways:

$$\hat{\mathbf{y}} = \beta_1\mathbf{x}_1 = 0.5\beta_1\mathbf{x}_1 + 0.5\beta_1\mathbf{x}_2 = 0.01\beta_1\mathbf{x}_1 + 0.99\beta_1\mathbf{x}_2 = \ldots$$

because obviously $\beta_1\mathbf{x}_2 = \beta_1\mathbf{x}_1$. In other words, there are infinitely many ways to combine $\mathbf{x}_1$ and $\mathbf{x}_2$ to obtain $\hat{\mathbf{y}}$, even though the latter is unique and the objective function has a well-defined minimum. It is rather easy to generalise the example above when $\mathbf{x}_2$ is a multiple of $\mathbf{x}_1$, that is $\mathbf{x}_2 = \alpha\mathbf{x}_1$.

We call this problem **collinearity**, or **multicollinearity**, and it can be solved quite easily: all you have to do is drop the collinear columns until $\mathbf{X}$ has full rank. Therefore, in the example above, you can choose to leave out $\mathbf{x}_1$ or $\mathbf{x}_2$; whatever your choice, problem solved.

In practice, things are not always so easy, because (as is well known) digital computers work with finite numerical precision, but in the cases we will consider we should have no problems. The interested reader may want to have a look at section 1.A.6.

A situation where this issue may arise goes commonly under the name of **dummy trap**. Suppose that you want to include in your model a qualitative variable, in which we have a conventional coding. For example, the marital status of an individual, and you conventionally code this information as 1=single, 2=married, 3=divorced, 4=it's complicated, etc.

Clearly, using this variable "as is" makes no sense: a function like $\hat{y}_i = \beta_1 + \beta_2 x_i$ would consider x_i as a proper numerical value, whereas in fact its coding is purely conventional. The solution is recoding x_i as a set of dummy variables, in which each dummy corresponds to one category: so for example the vector

$$\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

would be substituted by the matrix

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

so the first column of $\mathbf{Z}$ is a dummy variable for $x_i = 1$, the second one is a dummy variable for $x_i = 2$, and so on.

However, using the matrix $\mathbf{Z}$ unmodified leads to a collinearity problem if the model contains a constant, since the sum of the columns of the matrix $\mathbf{Z}$ is by construction equal to $\boldsymbol{\iota}$. Hence, the matrix

$$\mathbf{X} = [\boldsymbol{\iota} \quad \mathbf{Z}]$$

has not full rank, so $\mathbf{X}'\mathbf{X}$ doesn't have full rank either, and consequently is not invertible.[19]

[19] If you have problems following this argument, sections 1.A.3 and 1.A.4 may be of help.

The remedy you normally adopt is to drop one of the column of **Z**, and the corresponding category becomes the so-called "reference" category. For example, suppose you have a geographical variable x_i conventionally coded from 1 to 3 (1=North, 2=Centre, 3=South). The model $m(x_i) = \beta_1 + \beta_2 x_i$ is clearly meaningless, but one could think to set up an alternative model like

$$\hat{y}_i = \beta_1 + \beta_2 N_i + \beta_3 C_i + \beta_4 S_i,$$

where $N_i = 1$ if the i-th observation pertains to the North, and so on. This would make more sense, as all the variables in the model have a proper numerical interpretation. However, in this case we would have a collinearity problem for the reasons given above, that is $N_i + C_i + S_i = 1$ by construction for all observations i.

The solution is dropping one of the geographical dummies from the model: for example, let's say we drop the "South" dummy S_i: the model would become

$$\hat{y}_i = \beta_1 + \beta_2 N_i + \beta_3 C_i;$$

observe that with the above formulation the fitted value for a southern observation would be

$$\hat{y}_i = \beta_1 + \beta_2 \times 0 + \beta_3 \times 0 = \beta_1$$

whereas for a northern one you would have

$$\hat{y}_i = \beta_1 + \beta_2 \times 1 + \beta_3 \times 0 = \beta_1 + \beta_2,$$

so β_2 indicates the *difference* between a northern observation and a southern one, in the same way as β_3 indicates the difference between Centre and South. More in general, after dropping one of the dummies, the coefficient for each of the remaining ones indicates the difference between that category and the one you chose as a reference.

1.3.4 Nonlinearity

A further step in enhancing this setup would be allowing for the possibility that the function $m(x_i)$ is non-linear. In a traditional econometric setting this idea would take us to consider the so-called **NLS** (Nonlinear Least Squares) technique. I won't go into this either, for two reasons.

First, because minimising a loss function like $L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i - m(x_i, \boldsymbol{\beta}) \right]^2$, where $m(\cdot)$ is some kind of crazy arbitrary function may be a difficult problem: it could have more than one solution, or none, or maybe one that cannot be written in closed form.

Second, the linear model is in fact more general than it seems, since in order to use OLS it is sufficient that the model be linear *in the parameters*, not necessarily *in the variables*. For example, suppose that we have one explanatory variable; it is perfectly possible to use a model formulation like

$$m(x_i) = \beta_1 + \beta_2 x_i + \beta_3 x_i^2. \tag{1.13}$$

The equation above contains a non-linear transformation of x_i (the square), but the function itself is just a linear combination of observable data: in this case, we use a formulation that implies that the effect of x_i on $m(x_i)$ is nonlinear, but this is still achieved by employing a linear combination of observable variables. To be more explicit, the $\mathbf{X}$ matrix would be, in this case,

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \vdots & \\ 1 & x_n & x_n^2 \end{bmatrix}$$

and the algebra would proceed as usual.

This device is very common in applied econometrics, where powers of observable variables are used to accommodate nonlinear effects in the model without having to give up the computational simplicity of OLS. The parameter β_3 is also quite easy to read: if it's positive (negative), the $m(x_i)$ function is convex (concave).

The only caveat we have to be aware of is that, of course, you cannot interpret the $\boldsymbol{\beta}$ vector as marginal effects, as the right-hand side of equation (1.12) is no longer a fixed scalar. In fact, the marginal effects for each variable in the model become functions of the whole parameter vector $\boldsymbol{\beta}$ *and* of $\mathbf{x}_i$; in other terms, marginal effects may be different for each observation in our sample. For example, for the model in equation (1.13) the marginal effect of x_i would be

$$\frac{\partial m(x_i)}{\partial x_i} = \beta_2 + 2\beta_3 x_i;$$

and its sign would depend on the condition $x_i > -\frac{\beta_2}{2\beta_3}$, so it's entirely possible that the marginal effect of x_i on y_i is positive for some units in our sample and negative for others.

Example 1.3

Suppose you have the following model:

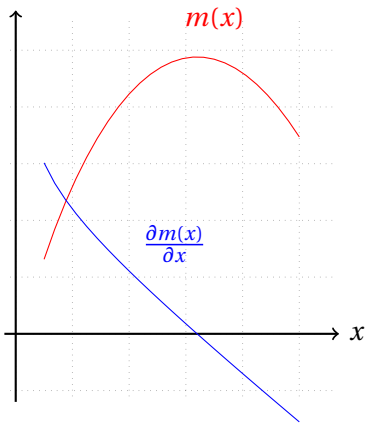
$$\hat{y}_i = m(x_i) = -1 + 2x_i - 0.4x_i^2 + 2\sqrt{x_i}$$

A plot of this function is depicted in Figure 1.2. The marginal effect of x_i on y_i is easy to find as

$$\frac{\partial m(x_i)}{\partial x_i} = 2 - 0.8x_i + \frac{1}{\sqrt{x_i}}$$

by differentiating each term. As you can see, the effect of x_i on y_i becomes individual-specific: for two individuals with a different x_i, the effect of a rise in x_i on y_i would depend on x_i, and can even change sign. So, what is a good thing for someone could be a bad thing for someone else.

Figure 1.2: $m(x_i)$ and its derivative as functions of x_i in example 1.3



More generally, what we can treat via OLS is the class of models that can be written as

$$m(\mathbf{x}_i) = \sum_{j=1}^{k} \beta_j g_j(\mathbf{x}_i),$$

where $\mathbf{x}_i$ are our "base" explanatory variables and $g_j(\cdot)$ is a sequence of arbitrary transformations, no matter how crazy. Each element of this sequence becomes a column of the $\mathbf{X}$ matrix. Clearly, once you have computed the $\hat{\boldsymbol{\beta}}$ vector, the marginal effects are easy to calculate (of course, as long as the $g_j(\cdot)$ functions are differentiable):

$$\frac{\partial m(\mathbf{x}_i)}{\partial \mathbf{x}_i} = \sum_{j=1}^{k} \hat{\beta}_j \frac{\partial g_j(\mathbf{x}_i)}{\partial \mathbf{x}_i}.$$

1.4 The geometry of OLS

The OLS statistic and associated concepts can be given an interpretation that has very little to do with statistics; instead, it's a geometrical interpretation. Given the typical audience of this book, a few preliminaries may be in order here.

The first concept we'll want to define is the concept of **distance** (also known as *metric*). Given two objects a and b, their distance is a function that should enjoy four properties:

1. $d(a,b) = d(b,a)$
2. $d(a,b) \geq 0$
3. $d(a,b) = 0 \Leftrightarrow a = b$
4. $d(a,b) + d(b,c) \geq d(a,c)$

The first three are obvious; as for the last one, called *triangle inequality*, it just means that the shortest way is the straight one. The objects in question may be of various sorts, but we will only consider the case when they are vectors. The distance of a vector from zero is its **norm**, written as $\|\mathbf{x}\| = d(\mathbf{x}, \mathbf{0})$.

Many functions $d(\cdot)$ enjoy the four properties above, but the concept of distance we use in everyday life is the so-called **Euclidean distance**, defined as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$$

and the reader may verify that the four properties are satisfied by this definition. Obviously, the formula for the Euclidean norm is $\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}}$.

The second concept I will use is the idea of a **vector space**. If you're not familiar with vector spaces, linear combinations and the rank of a matrix, then sections 1.A.2 and 1.A.3 are for you.[20] In brief, I use the expression $\text{Sp}(\mathbf{X})$ to indicate the set of all vectors that can be obtained as a linear combination of the columns of $\mathbf{X}$.

Consider the space $\mathbb{R}^n$, where you have a vector $\mathbf{y}$ and a few vectors $\mathbf{x}_j$, with $j = 1 \dots k$ and $k < n$, all packed in a matrix $\mathbf{X}$. What we want to find is the element of $\text{Sp}(\mathbf{X})$ which is closest to $\mathbf{y}$. In formulae:

$$\hat{\mathbf{y}} = \underset{\mathbf{x} \in \text{Sp}(\mathbf{X})}{\text{Argmin}} \ \|\mathbf{y} - \mathbf{x}\|;$$

since the optimal point must belong to $\text{Sp}(\mathbf{X})$, the problem can be rephrased as: find the vector $\boldsymbol{\beta}$ such that $\mathbf{X}\boldsymbol{\beta}$ (that belongs to $\text{Sp}(\mathbf{X})$ by construction) is closest to $\mathbf{y}$:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\text{Argmin}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|. \tag{1.14}$$

If we decide to adopt the Euclidean definition of distance, then the solution is exactly the same as the one to the statistical problem of Section 1.3.2: since the "square root" function is monotone, the minimum of $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|$ is the same as the minimum of $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, and therefore

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\text{Argmin}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\| = \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

from which

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Note that $\hat{\mathbf{y}}$ is a linear transform of $\mathbf{y}$: you obtain $\hat{\mathbf{y}}$ by premultiplying $\mathbf{y}$ by the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$; this kind of transformation is called a "projection".

[20] If, on the other hand, you find the topic intriguing and want a rigorous yet very readable book on this subject, check out Axler (2015).

1.4.1 Projection matrices

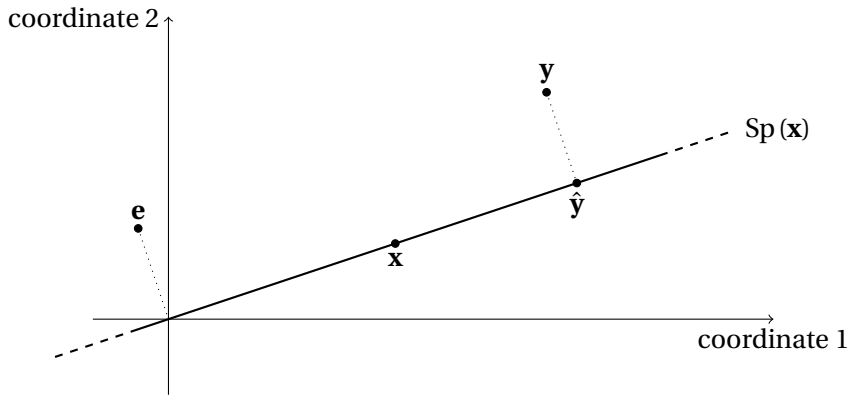
In the previous subsection I pointed out that $\hat{\mathbf{y}}$ is a linear transform of $\mathbf{y}$. The matrix that operates the transform is said to be a **projection matrix**.[21] To see why, there's an example I always use: the fly in the cinema. Imagine you're sitting in a cinema, and there's a fly somewhere in the room. You see a dot on the screen: the fly's shadow. The position of the fly is $\mathbf{y}$, the space spanned by $\mathbf{X}$ is the screen and the shadow of the fly is $\hat{\mathbf{y}}$.

The matrix that turns the position of the fly into the position of its shadow is $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. To be more precise, this matrix projects onto $\text{Sp}(\mathbf{X})$ any vector it premultiplies, and it's such a handy tool that it has its own abbreviation: $\mathbf{P}_\mathbf{X}$.

$$\mathbf{P}_\mathbf{X} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and $\hat{\mathbf{y}}$ can be written as $\hat{\mathbf{y}} = \mathbf{P}_\mathbf{X}\mathbf{y}$. The reader may find it amusing that in the econometrics jargon the $\mathbf{P}_\mathbf{X}$ matrix is sometimes referred to as the "hat" matrix, because $\mathbf{P}_\mathbf{X}$ "puts a hat on $\mathbf{y}$".

Figure 1.3: Example: projection of a vector on another one



In this simple example, $\mathbf{x} = (3,1)$ *and* $\mathbf{y} = (5,3)$*; the reader may want to check that* $\hat{\mathbf{y}} = (5.4, 1.8)$ *and* $\mathbf{e} = (-0.4, 1.2)$.

The base property of $\mathbf{P}_\mathbf{X}$ is that, by construction, $\mathbf{P}_\mathbf{X}\mathbf{X} = \mathbf{X}$, as you can easily check. Moreover, it's symmetric and idempotent.

$$\mathbf{P}_\mathbf{X} = \mathbf{P}_\mathbf{X}' \qquad\qquad \mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X} = \mathbf{P}_\mathbf{X}.$$

We call **idempotent** something that does not change when multiplied by itself; for example, the real numbers 0 and 1 are idempotent. A nice way to understand the meaning of idempotency is by reflecting on its geometrical implication: the

[21] If I were pedantic, I'd have to say *orthogonal* projection, because you also get a tool called *oblique* projection. We'll never use it in this book, apart from a passing reference in chapter 6.

matrix $\mathbf{P_X}$ takes a vector from wherever it is and moves it onto the closest point of $\text{Sp}(\mathbf{X})$; but if the starting point already belongs to $\text{Sp}(\mathbf{X})$, obviously no movement takes place at all, so applying $\mathbf{P_X}$ to a vector more than once produces no extra effects ($\mathbf{P_X y} = \mathbf{P_X P_X y} = \mathbf{P_X P_X} \cdots \mathbf{P_X y}$).

It can also be proven that $\mathbf{P_X}$ is singular;[22] again, this algebraic property can be given a nice intuitive geometric interpretation: a projection entails a loss of information, because some of the original coordinates get "squashed" onto $\text{Sp}(\mathbf{X})$: in the fly example, it's impossible to know the exact position of the fly from its shadow, because one of the coordinates (the distance from the screen) is lost. In formulae, the implication of $\mathbf{P_X}$ being singular is that no matrix $\mathbf{A}$ exists such that $\mathbf{A} \cdot \mathbf{P_X} = \mathbf{I}$, and therefore no matrix exists such that $\mathbf{A}\hat{\mathbf{y}} = \mathbf{y}$, which means that $\mathbf{y}$ is impossible to reconstruct from its projection.

In practice, when you regress $\mathbf{y}$ on $\mathbf{X}$, you are performing exactly the same calculations that are necessary to find the projection of $\mathbf{y}$ onto $\text{Sp}(\mathbf{X})$, and the vector $\hat{\beta}$ contains the coordinates for locating $\hat{\mathbf{y}}$ in that space.

There is another interesting matrix we'll be using often:

$$\mathbf{M_X} = \mathbf{I} - \mathbf{P_X}.$$

By definition, therefore, $\mathbf{M_X y} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}$. The $\mathbf{M_X}$ matrix performs a complementary task to that of $\mathbf{P_X}$: when you apply $\mathbf{M_X}$ to a vector, it returns the difference between the original point and its projection. We may say that $\mathbf{e} = \mathbf{M_X y}$ contains the information that is lost in the projection. It is easily checked that $\mathbf{M_X X} = [0]$ and as a consequence,

$$\mathbf{M_X P_X} = \mathbf{P_X M_X} = [0],$$

where I'm using the notation $[0]$ for "a matrix full of zeros".

Some more noteworthy properties: $\mathbf{M_X}$ is symmetric, idempotent and singular, just like $\mathbf{P_X}$.[23] As for is rank, it can be proven that its rank equals $n - k$, where n is the number of rows of $\mathbf{X}$ and $r = \text{rk}(\mathbf{X})$.

A fundamental property this matrix enjoys is that every vector of the type $\mathbf{M_X y}$ is orthogonal to $\text{Sp}(\mathbf{X})$, so it forms a 90° angle with any vector that can be written as $\mathbf{X}\lambda$.[24] These properties are very convenient in many cases; a notable one is the possibility of rewriting the SSR as a quadratic form:[25].

$$\mathbf{e}'\mathbf{e} = (\mathbf{M_X y})'(\mathbf{M_X y}) = \mathbf{y}'\mathbf{M_X M_X y} = \mathbf{y}'\mathbf{M_X y}$$

[22] To be specific: it can be proven that $\text{rk}(\mathbf{P_X}) = \text{rk}(\mathbf{X})$, so $\mathbf{P_X}$ is a $n \times n$ matrix with rank k; evidently, in the situation we're considering here, $n > k$. Actually, it can be proven that no idempotent matrix is invertible, the identity matrix being the only exception.

[23] In fact, $\mathbf{M_X}$ is itself a projection matrix, but let's not get into this, ok?

[24] Let me remind the reader that two vectors are said to be **orthogonal** if their inner product is 0. In formulae: $\mathbf{x} \perp \mathbf{y} \Leftrightarrow \mathbf{x}'\mathbf{y} = 0$. A vector is orthogonal to a space if it's orthogonal to all the points that belong to that space: $\mathbf{y} \perp \text{Sp}(\mathbf{X}) \Leftrightarrow \mathbf{y}'\mathbf{X} = \mathbf{0}$, so $\mathbf{y} \perp \mathbf{X}\lambda$ for any λ.

[25] A **quadratic form** is an expression like $\mathbf{x}'A\mathbf{x}$, where $\mathbf{x}$ is a vector and A is a square matrix, usually symmetric. I sometimes use the metaphor of a sandwich and call $\mathbf{x}$ the "bread" and A the "cheese".

where the second equality comes from symmetry and the third one from idempotency. By the way, the above expression could be further manipulated to re-obtain equation (1.11):

$$\mathbf{y}'\mathbf{M}_\mathbf{X}\mathbf{y} = \mathbf{y}'(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{y} = \mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y}-\hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{y}).$$

Example 1.4

Readers are invited to check (by hand or using a computer program of their choice) that, with the matrices used in example 1.2, $\mathbf{P}_\mathbf{X}$ *equals*

$$\mathbf{P}_\mathbf{X} = \begin{bmatrix} 0.3 & 0.2 & 0.1 & 0.4 & 0 & 0 \\ 0.2 & 0.175 & 0.15 & 0.225 & 0.125 & 0.125 \\ 0.1 & 0.15 & 0.2 & 0.05 & 0.25 & 0.25 \\ 0.4 & 0.225 & 0.05 & 0.575 & -0.125 & -0.125 \\ 0 & 0.125 & 0.25 & -0.125 & 0.375 & 0.375 \\ 0 & 0.125 & 0.25 & -0.125 & 0.375 & 0.375 \end{bmatrix}$$

and that does in fact satisfy the idempotency property.

In the present context, the advantage of using projection matrices is that the main quantities that appear in the statistical problem of approximating y_i via $\mathbf{x}_i$ become easy to represent in a compact and intuitive way:

Magnitude	Symbol	Formula
OLS Coefficients	$\hat{\boldsymbol{\beta}}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
Fitted values	$\hat{\mathbf{y}}$	$\mathbf{P}_\mathbf{X}\mathbf{y}$
Residuals	$\mathbf{e}$	$\mathbf{M}_\mathbf{X}\mathbf{y}$
Sum of squared residuals	SSR	$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{M}_\mathbf{X}\mathbf{y}$

Take for example the special case $\mathbf{X} = \boldsymbol{\iota}$. As we now know, the optimal solution to the statistical problem is using the sample average, so $\hat{\boldsymbol{\beta}} = \bar{Y}$: the fitted values are $\mathbf{P}_\boldsymbol{\iota}\mathbf{y} = \boldsymbol{\iota}\cdot\bar{Y}$ and the residuals are simply the deviations from the mean: $\mathbf{e} = \mathbf{M}_\boldsymbol{\iota}\mathbf{y} = \mathbf{y}-\boldsymbol{\iota}\cdot\bar{Y}$. Finally, deviance can be written as $\mathbf{y}'\mathbf{M}_\boldsymbol{\iota}\mathbf{y}$.

1.4.2 Measures of fit

We are now ready to tackle a very important issue. How good is our model? We know that $\hat{\boldsymbol{\beta}}$ is the best we can choose if we want to approximate y_i via $\hat{y}_i = \mathbf{x}_i'\boldsymbol{\beta}$, but nobody guarantees that our best should be particularly good. A natural way to rephrase this question is: how much information are we losing in the projection? We know the information loss is minimal, but it could still be quite large.

In order to answer this question, let us start from the following two inequalities:

$$0 \leq \hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} \leq \mathbf{y}'\mathbf{y}; \tag{1.15}$$

the first one is rather obvious, considering that $\hat{\mathbf{y}}'\hat{\mathbf{y}}$ is a sum of squares, and therefore non-negative. The other one, instead, can be motivated via $\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{M}_\mathbf{X}\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{e}'\mathbf{e}$; since $\mathbf{e}'\mathbf{e}$ is also a sum of squares, $\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} \leq \mathbf{y}'\mathbf{y}$. If we divide everything by $\mathbf{y}'\mathbf{y}$, we get

$$0 \leq \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y}} = R_u^2 \leq 1. \tag{1.16}$$

This index bears the name R_u^2 ("uncentred R-squared"), and, as the above expression shows, it's bounded by construction between 0 and 1. It can be given a very intuitive geometric interpretation: evidently, in $\mathbb{R}^n$ the points $\mathbf{0}$, $\mathbf{y}$ and $\hat{\mathbf{y}}$ form a right triangle (see also Figure 1.3), in which you get a "good" leg, that is $\hat{\mathbf{y}}$, and a "bad" one, the segment linking $\hat{\mathbf{y}}$ and $\mathbf{y}$, which is congruent to $\mathbf{e}$: we'd like the bad leg to be as short as possible. After Pythagoras' theorem, the R_u^2 index gives us (the square of) the ratio between the good leg and the hypotenuse. Of course, we'd like this ratio to be as close to 1 as possible.

Example 1.5

With the matrices used in example 1.2, you get that $\mathbf{y}'\mathbf{y} = 24$ *and* $\mathbf{e}'\mathbf{e} = 4.325$*; therefore,*

$$R_u^2 = 1 - \frac{4.325}{24} \simeq 81.98\%$$

The R_u^2 index makes perfect sense geometrically, but hardly any from a statistical point of view: the quantity $\mathbf{y}'\mathbf{y}$ has a natural geometrical interpretation, but statistically it doesn't mean much, unless we give it the meaning

$$\mathbf{y}'\mathbf{y} = (\mathbf{y} - \mathbf{0})'(\mathbf{y} - \mathbf{0}),$$

that is, the SSR for a model in which $\hat{\mathbf{y}} = \mathbf{0}$. Such a model would be absolutely minimal, but rather silly as a model. Instead, we might want to use as a benchmark our initial proposal described in section 1.2, where $\mathbf{X} = \boldsymbol{\iota}$. In this case, the SSR is just the deviance of $\mathbf{y}$, that is the sum of squared deviations from the mean, which can be written as $\mathbf{y}'\mathbf{M}_\boldsymbol{\iota}\mathbf{y}$.

If $\boldsymbol{\iota} \in \text{Sp}(\mathbf{X})$ (typically, when the model contains a constant term, but not necessarily), then a decomposition similar to (1.15) is possible: since $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$, then obviously

$$\mathbf{y}'\mathbf{M}_\boldsymbol{\iota}\mathbf{y} = \hat{\mathbf{y}}'\mathbf{M}_\boldsymbol{\iota}\hat{\mathbf{y}} + \mathbf{e}'\mathbf{M}_\boldsymbol{\iota}\mathbf{e} = \hat{\mathbf{y}}'\mathbf{M}_\boldsymbol{\iota}\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e} \tag{1.17}$$

because if $\boldsymbol{\iota} \in \text{Sp}(\mathbf{X})$, then $\mathbf{M}_\boldsymbol{\iota}\mathbf{e} = \mathbf{e}$.[26] Therefore,

$$0 \leq \mathbf{e}'\mathbf{e} \leq \mathbf{y}'\mathbf{M}_\boldsymbol{\iota}\mathbf{y},$$

[26] Subsection 1.A.8 should help the readers who want this result proven.

where the second inequality comes from the fact that $\hat{\mathbf{y}}'\mathbf{M}_\iota\hat{\mathbf{y}}$ is a sum of squares and therefore non-negative.[27] The modified version of R^2 is known as **centred R-square**:

$$R^2 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}_\iota\mathbf{y}}. \tag{1.18}$$

The concept of R^2 that we normally use in econometrics is the centred one, and this is why the index defined at equation (1.16) has the "u" as a footer (from the word *uncentred*).

In a way, the definition of R^2 is implicitly based on a *comparison between different models*: one which uses all the information contained in $\mathbf{X}$ and another (smaller) one, which only uses ι, because $\mathbf{y}'\mathbf{M}_\iota\mathbf{y}$ is just the SSR of a model in which we regress $\mathbf{y}$ on ι. Therefore, equation (1.18) can be read as a way to compare the loss function for those two models.

In fact, this same idea can be pushed a little bit further: imagine that we wanted to compare model A and model B, in which B contains the same explanatory variables as A, plus some more. In practice:

$$\begin{array}{ll} \text{Model A} & \mathbf{y} \simeq \mathbf{X}\boldsymbol{\beta} \\ \text{Model B} & \mathbf{y} \simeq \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} = \mathbf{W}\boldsymbol{\theta} \end{array}$$

where $\mathbf{W} = [\mathbf{X} \quad \mathbf{Z}]$ and $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}$.

The matrix $\mathbf{Z}$ contains additional regressors to model A. It is important to realise that the information contained in $\mathbf{Z}$ could be perfectly relevant and legitimate, but also ridiculously useless. For example, a model for the academic performance of A. S. Tudent could well contain, as an explanatory variable, the number of pets A. S. Tudent's neighbours have, or the number of consonants in A. S. Tudent's mother's surname.

It's easy to prove that the SSR for model B is always smaller than that for A:

$$SSR_A = \mathbf{e}_a'\mathbf{e}_a \qquad SSR_B = \mathbf{e}_b'\mathbf{e}_b$$

where $\mathbf{e}_a = \mathbf{M}_\mathbf{X}\mathbf{y}$ and $\mathbf{e}_b = \mathbf{M}_\mathbf{W}\mathbf{y}$. Since $\mathbf{X} \in \text{Sp}(\mathbf{W})$, clearly[28] $\mathbf{P}_\mathbf{W}\mathbf{X} = \mathbf{X}$ and therefore

$$\mathbf{M}_\mathbf{W}\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{W},$$

which implies $\mathbf{M}_\mathbf{W}\mathbf{e}_a = \mathbf{e}_b$; as a consequence,

$$SSR_B = \mathbf{e}_b'\mathbf{e}_b = \mathbf{e}_a'\mathbf{M}_\mathbf{W}\mathbf{e}_a = \mathbf{e}_a'\mathbf{e}_a - \mathbf{e}_a'\mathbf{P}_\mathbf{W}\mathbf{e}_a \leq \mathbf{e}_a'\mathbf{e}_a = SSR_A.$$

More generally, if $\text{Sp}(\mathbf{W}) \supset \text{Sp}(\mathbf{X})$, then $\mathbf{y}'\mathbf{M}_\mathbf{W}\mathbf{y} \leq \mathbf{y}'\mathbf{M}_\mathbf{X}\mathbf{y}$ for any vector $\mathbf{y}$.

[27] Note that, in the rare but not impossible case $\iota \notin \text{Sp}(\mathbf{X})$, it is perfectly possible that $\mathbf{e}'\mathbf{e} < \mathbf{y}'\mathbf{M}_\iota\mathbf{y}$, so the centred version of the R^2 index may be negative.

[28] Some may say: "well, not so clearly". OK, here goes: $\mathbf{X} \in \text{Sp}(\mathbf{W})$ implies that there is a matrix $\mathbf{H}$ such that $\mathbf{X} = \mathbf{W}\mathbf{H}$. Hence, $\mathbf{P}_\mathbf{W}\mathbf{X} = \mathbf{P}_\mathbf{W}\mathbf{W}\mathbf{H} = \mathbf{W}\mathbf{H} = \mathbf{X}$.

The implication is: if we had to choose between A and B by using the SSR as a criterion, model B would always be the winner, *no matter how absurd the choice of the variables* $\mathbf{Z}$ *is*. The R^2 index isn't any better: proving that

$$SSR_B \leq SSR_A \Rightarrow R^2_B \geq R^2_A.$$

is a trivial exercise, so if you add any explanatory variable to an existing model, the R^2 index cannot become smaller.

A possible solution could be using a slight variation of the index, which goes by the name of **adjusted** R^2:

$$\overline{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}_\iota\mathbf{y}}\frac{n-1}{n-k}, \tag{1.19}$$

where n is the size of our dataset and k is the number of explanatory variables. It is easy to prove that if you add silly variables to a model, so that the SSR changes only slightly, the $n-k$ in the denominator should offset that effect. However, as we will see in section 3.3.2, the best way of choosing between models is by framing the decision in a proper inferential context.

One final thing on the R^2 index. Although it's perfectly legitimate to think that 0 is "bad" and 1 is "good", it would be unwise to automatically consider a number close to 0 (say, 10%) as "rather bad" or, symmetrically, a number close to 1 (say, 90%) as "pretty good": a model is an approximate description of the dependent variable y_i , insofar as the explanatory variables $\mathbf{x}_i$ contain relevant information. It may well be that the main determinants of y_i are unobservable, and therefore $\mathbf{x}_i$ only manages to capture a small portion of the overall dispersion of y_i. In these cases, the R^2 index will be very small, but it doesn't necessarily follow that our model is worthless: the relationship that it reveals between the dependent variable and the explanatory variables may be extremely valuable, even if the fraction of variance we explain is small. But again, this idea is more properly framed as a statistical inference issue, which is what chapter 3 is about.

1.4.3 Reparametrisations

Suppose that there are two researchers (Alice and Bob), who have the same dataset, which contains three variables: y_i, x_i and z_i. Alice performs OLS on the model

$$y_i \simeq \beta_1 x_i + \beta_2 z_i$$

Bob, instead, computes the new variables $s_i = x_i + z_i$ and $d_i = x_i - z_i$ and computes his coefficients using the transformed regressors as

$$y_i \simeq \gamma_1 s_i + \gamma_2 d_i.$$

How different will the two models be? Before delving into algebra, it is worth observing that Alice and Bob are using the same data, and it would be surprising

if they arrived at different conclusions. Moreover, Alice and Bob's choices are simply a matter of taste, and there's no "right" way to set up a model. One could compute s_i and d_i from x_i and z_i, or the other way around. In other words, the set of explanatory variables Alice and Bob are using are invertible transformations of one another, and therefore must contain the same information, expressed in a different way.

With this in mind, a relationship between the two sets of parameters is easy to find: start from Bob's model

$$\begin{aligned} y_i &\simeq \gamma_1 s_i + \gamma_2 d_i = \\ &= \gamma_1 (x_i + z_i) + \gamma_2 (x_i - z_i) = \\ &= (\gamma_1 + \gamma_2) x_i + (\gamma_1 - \gamma_2) z_i, \end{aligned}$$

so $\beta_1 = (\gamma_1 + \gamma_2)$ and $\beta_2 = (\gamma_1 - \gamma_2)$. Clearly, this entails that Bob's parameters can be recovered from Alice's as $\gamma_1 = \frac{\beta_1 + \beta_2}{2}$ and $\gamma_2 = \frac{\beta_1 - \beta_2}{2}$. It is perfectly legitimate to surmise that the two models are in fact equivalent, and should give the same fit.

More generally, it is possible to show that Alice's model can be written as $\mathbf{y} \simeq \mathbf{X}\boldsymbol{\beta}$ and Bob's model as $\mathbf{y} \simeq \mathbf{Z}\boldsymbol{\gamma}$, where $\mathbf{Z} = \mathbf{X}A$ and A is square and invertible. In the example above,

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

This simple fact has a very nice consequence on the respective projection matrices:

$$\begin{aligned} \mathbf{P}_{\mathbf{Z}} &= \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \\ &= \mathbf{X}A(A'\mathbf{X}'\mathbf{X}A)^{-1}A'\mathbf{X}' \\ &= \mathbf{X}A(A)^{-1}(\mathbf{X}'\mathbf{X})^{-1}(A')^{-1}A'\mathbf{X}' \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}_{\mathbf{X}}, \end{aligned}$$

that is, *the two projection matrices are the same*.[29] Therefore, $\mathrm{Sp}(\mathbf{X}) = \mathrm{Sp}(\mathbf{Z})$: Alice and Bob are projecting $\mathbf{y}$ onto the same space. It should be no surprise that they will get the same fitted values $\hat{\mathbf{y}}$ and the same residuals $\mathbf{e}$. As a further consequence, all the quantities that depend on the projection will be the same, such as the sum of squared residuals, the R^2 index and so on. As a matter of fact, Alice's and Bob's models are just the same model written in a different way, by a different representation choice, which uses different parameters. The relationship between the two sets of parameters is easy to show: since $\hat{\mathbf{y}}$ is the same for the two models, then it must also hold

$$\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\gamma}} = \mathbf{X}A\hat{\boldsymbol{\gamma}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

[29] If you find some of the passages above unclear, then Section 1.A.4 may be useful.

and therefore $\hat{\boldsymbol{\beta}} = A\hat{\boldsymbol{\gamma}}$ (and of course $\hat{\boldsymbol{\gamma}} = A^{-1}\hat{\boldsymbol{\beta}}$). The word we use in this context is **reparametrisation**: Bob's model is a reparametrisation of Alice's and vice versa. The difference between the two is just aesthetic, so to speak: in some cases, it could be more natural to interpret the coefficients of a model written in a certain way than another. This is a very common trick in applied economics, and an egregious example will be given in Section 5.5.

1.4.4 The Frisch-Waugh theorem

Projection matrices are also useful to illustrate a remarkable result, known as the **Frisch-Waugh theorem**:[30] given a model of the kind $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$, split $\mathbf{X}$ vertically into two sub-matrices $\mathbf{Z}$ and $\mathbf{W}$, and $\boldsymbol{\beta}$ accordingly

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{Z} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}$$

Applying equation (1.7) we get the following:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Z}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

It would seem that finding an analytical closed form for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as functions of $\mathbf{Z}$, $\mathbf{W}$ and $\mathbf{y}$ is quite difficult; fortunately, it isn't so: start from

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e} = \mathbf{Z}\hat{\boldsymbol{\beta}}_1 + \mathbf{W}\hat{\boldsymbol{\beta}}_2 + \mathbf{e}$$

and premultiply the equation above by $\mathbf{M}_\mathbf{W}$:

$$\mathbf{M}_\mathbf{W}\mathbf{y} = \mathbf{M}_\mathbf{W}\mathbf{Z}\hat{\boldsymbol{\beta}}_1 + \mathbf{e},$$

since $\mathbf{M}_\mathbf{W}\mathbf{W} = 0$ (by construction) and $\mathbf{M}_\mathbf{W}\mathbf{e} = \mathbf{e}$ (because $\mathbf{e} = \mathbf{M}_\mathbf{X}\mathbf{y}$, but $\mathrm{Sp}(\mathbf{W}) \subset \mathrm{Sp}(\mathbf{X})$, so $\mathbf{M}_\mathbf{W}\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{X}$).[31] Now premultiply by $\mathbf{Z}'$:

$$\mathbf{Z}'\mathbf{M}_\mathbf{W}\mathbf{y} = \mathbf{Z}'\mathbf{M}_\mathbf{W}\mathbf{Z}\hat{\boldsymbol{\beta}}_1$$

since $\mathbf{Z}'\mathbf{e} = 0$, because $\mathbf{Z}'\mathbf{M}_\mathbf{X} = 0$. As a consequence,

$$\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{Z}'\mathbf{M}_\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{M}_\mathbf{W}\mathbf{y} \tag{1.20}$$

Since $\mathbf{M}_\mathbf{W}$ is idempotent, an alternative way to write (1.20) could be

$$\hat{\boldsymbol{\beta}}_1 = \left[(\mathbf{Z}'\mathbf{M}_\mathbf{W})(\mathbf{M}_\mathbf{W}\mathbf{Z})\right]^{-1}(\mathbf{Z}'\mathbf{M}_\mathbf{W})(\mathbf{M}_\mathbf{W}\mathbf{y});$$

[30] In fact, many call this theorem the Frisch-Waugh-Lovell theorem, as it was Micheal Lovell who, in a paper appeared in 1963, generalised the original result that Frisch and Waugh had obtained 30 years earlier to its present form.

[31] If you're getting a bit confused, you may want to take a look at section 1.A.8.

therefore $\hat{\boldsymbol{\beta}}_1$ is the vector of the coefficients for a model in which the dependent variable is the vector of the residuals of $\mathbf{y}$ with respect to $\mathbf{W}$ and the regressor matrix is the matrix of residuals of $\mathbf{Z}$ with respect to $\mathbf{W}$. For symmetry reasons, you also obviously get a corresponding expression for $\hat{\boldsymbol{\beta}}_2$:

$$\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{W}'\mathbf{M}_{\mathbf{Z}}\mathbf{W}\right)^{-1}\mathbf{W}'\mathbf{M}_{\mathbf{Z}}\mathbf{y}$$

In practice, a perfectly valid algorithm for computing $\hat{\boldsymbol{\beta}}_1$ could be:

1. regress $\mathbf{y}$ on $\mathbf{W}$; take the residuals and call them $\tilde{\mathbf{y}}$;
2. regress each column of $\mathbf{Z}$ on $\mathbf{W}$; form a matrix with the residuals and call it $\tilde{\mathbf{Z}}$;
3. regress $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{Z}}$: the result is $\hat{\boldsymbol{\beta}}_1$.

This result is not just a mathematical curiosity, nor a computational gimmick: it comes in handy in a variety of situations for proving theoretical results. For example, we'll use this theorem more than once in chapters 3, 6 and 7.

An interpretation that the Frisch-Waugh theorem can be given is the following: the coefficients for a group of regressors measure the response of $\hat{\mathbf{y}}$ *having taken into account* the other ones or, as we say, "everything else being equal". The phrase normally used in the profession is "controlling for". For example: suppose that $\mathbf{y}$ contains data on the wages for n employees, that $\mathbf{Z}$ is their education level and $\mathbf{W}$ is a geographical dummy variable (North vs South). The vector $\tilde{\mathbf{y}} = \mathbf{M}_{\mathbf{W}}\mathbf{y}$ will contain the differences between the individual wages and the average wage *of the region where they live*, in the same way as $\tilde{\mathbf{Z}} = \mathbf{M}_{\mathbf{W}}\mathbf{Z}$ contains the data on education as deviation *from the regional mean*. Therefore, regressing $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{Z}}$ is a way to implicitly take into account that differences in wages between regions may depend on different educational levels. Consequently, by regressing $\mathbf{y}$ on both the "education" variable and the regional dummy variable, the coefficient for education will measure its effect on wages controlling for geographical effects.

1.5 An example

For this example, I got some data from the 2016 SHIW dataset;[32] our dataset contains 1917 individuals, who are full-time employees.[33] We are going to use four variables, briefly described in Table 1.1. Our dependent variable is going to be w, the natural logarithm of the hourly wage in Euro. The set of explanatory

[32] SHIW is the acronym for "Survey on Household Income and Wealth", provided by the Bank of Italy, which is a very rich and freely available dataset: see https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/.

[33] I can send you the details on the construction of the dataset from the raw data, if you're interested. Just send me an email.

variables was chosen in accordance with some vague and commonsense idea of the factors that can account for differences in wages. We would expect that people with higher education and/or longer work experience should command a higher wage, but we would also use the information on gender, because we are aware of an effect called "gender gap", that we might want to take into account.

Variable	Description	Mean	Median	S. D.	Min	Max
w	Log hourly wage	2.22	2.19	0.364	0.836	4.50
g	dummy, male = 1	0.601	1.00	0.490	0.00	1.00
e	education (years)	11.7	13.0	3.60	0.00	21.0
a	work experience (years)	27.4	29.0	10.9	0.00	58.0

Table 1.1: Wage example

The data that we need to compute $\hat{\boldsymbol{\beta}}$ are:

$$\begin{aligned}
\mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1917 & 1153 & 22493 & 52527 \\ 1153 & 1153 & 13299 & 32691 \\ 22493 & 13299 & 288731 & 594479 \\ 52527 & 32691 & 594479 & 1666703 \end{bmatrix} \\
\mathbf{X}'\mathbf{y} &= \begin{bmatrix} 4253.3716 \\ 2633.5507 \\ 51038.9769 \\ 116972.6710 \end{bmatrix} \\
\mathbf{y}'\mathbf{y} &= 9690.62
\end{aligned}$$

The reader is invited to check that the inverse of $\mathbf{X}'\mathbf{X}$ is (roughly)

$$(\mathbf{X}'\mathbf{X})^{-1} = 10^{-5} \cdot \begin{bmatrix} 1356.9719 & -120.7648 & -63.9068 & -17.6027 \\ -120.7648 & 220.4901 & 1.2056 & -0.9488 \\ -63.9068 & 1.2056 & 4.4094 & 0.4177 \\ -17.6027 & -0.9488 & 0.4177 & 0.4844 \end{bmatrix}$$

and therefore we have

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1.3289 \\ 0.1757 \\ 0.0526 \\ 0.0061 \end{bmatrix} \qquad \mathbf{e}'\mathbf{e} = 177.9738 \qquad R^2 = 29.76\%$$

but it is much more common to see results presented in a table like Table 1.2.

At this point, there are quite a few numbers in the table above that we don't know how to read yet, but we have time for this: chapter 3 is devoted entirely to this purpose. The important thing for now is that we have a reasonably efficient way to summarise the information on wages via the following model:

$$\hat{w}_i = 1.33 + 0.176 g_i + 0.053 e_i + 0.006 a_i$$

```
            coefficient   std. error    t-ratio    p-value
  ------------------------------------------------------------
  const      1.32891       0.0355309     37.40     2.86e-230 ***
  male       0.175656      0.0143224     12.26     2.42e-33  ***
  educ       0.0526218     0.00202539    25.98     1.02e-127 ***
  wexp       0.00608615    0.000671303    9.066    2.97e-19  ***

Mean dependent var   2.218765   S.D. dependent var   0.363661
Sum squared resid    177.9738   S.E. of regression   0.305015
R-squared            0.297629   Adjusted R-squared   0.296528
```

Table 1.2: Wage example — OLS output

where w_i is the log wage for individual i, g_i is their gender, and the rest follows.

In practice, if we had a guy who studied for 13 years and has worked for 20 years, we would guess that the log of his hourly wage would be

$$1.33 + 0.176 \cdot 1 + 0.052 \cdot 13 + 0.006 \cdot 20 \simeq 2.31$$

which is roughly €10 an hour (which sounds reasonable).

The quality of the approximation is not bad: the R^2 index is roughly 30%, which means that if we compare the loss functions for our model and the one we get if we has just used the average wage, we get

$$0.298 = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}_\iota\mathbf{y}} \Longrightarrow \mathbf{e}'\mathbf{e} = 0.702 \cdot \mathbf{y}'\mathbf{M}_\iota\mathbf{y};$$

if you consider the dazzling complexity of the factors that potentially dictate why two individuals get different wages, the fact that a simple linear rule involving only three variables manages to describe 30% of the heterogeneity between individual is surprisingly good.

Of course, nothing is stopping us from interpreting the sign and magnitude of our OLS coefficients: for example, the coefficient for education is about 5%, and therefore the best way to use the educational attainment variable for summarising the data we have on wages is by saying that each year of extra education gives you a guess which is about 5% higher.[34] Does this imply that you get positive returns to education in the Italian labour market? Strictly speaking, it doesn't. This number yields a fairly decent approximation to our dataset of 1917 people. To assume that the same regularity should hold for others is totally unwarranted. And the same goes for the gender gap: it would seem that being male shifts your fitted wage by 17.5%. But again, at the risk of being pedantic, all we can say is that among our 1917 data points, males get (on average) more

[34] One of the reasons why we economists love logarithms is that they auto-magically turn absolute changes into relative ones: $\beta_2 = \frac{\mathrm{d}w}{\mathrm{d}e} = \frac{\mathrm{d}\ln(W)}{\mathrm{d}e} = \frac{1}{W}\frac{\mathrm{d}W}{\mathrm{d}e} \simeq \frac{\Delta W/W}{\Delta e}$. In other words, the coefficient associated with the educational variable gives you a measure of the relative change in wage in response to a unit increase in education.

money than females with the same level of experience and education. Coincidence? We should be wary of generalisations, however tempting they may be to our sociologist self.

And yet, these thoughts are perfectly natural. The key ingredient to give scientific legitimacy to this sort of mental process is to frame it in the context of statistical inference, which is the object of the next chapter.

1.A Assorted results

This section contains several results on matrix algebra, in the simplest form possible. If you want an authoritative reference, my advice is to get one of Horn and Johnson (2012), Abadir and Magnus (2005) or Lütkepohl (1996), which are all excellent and use a notation and style that is close to what we use in econometrics.

1.A.1 Matrix differentiation rules

The familiar concept of a derivative of a function of a scalar can be generalised to functions of a vector

$$y = f(\mathbf{x}),$$

where you have a real number y for every possible vector $\mathbf{x}$. For example, if $y = x + w^z$, you can define the vector $\mathbf{x} = [x, w, z]'$. The generalisation of the concept of derivative is what we call the **gradient**, that is a vector collecting the partial derivatives with respect to the corresponding elements of $\mathbf{x}$. We adopt the convention by which the gradient is a *row* vector; hence, for the example above, the gradient is

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x} & \frac{\partial y}{\partial w} & \frac{\partial y}{\partial z} \end{bmatrix} = \begin{bmatrix} 1 & z w^{z-1} & \log(w) \cdot w^z \end{bmatrix}$$

The cases we'll need are very simple, because they generalise the simple univariate functions $y = ax$ and $y = ax^2$. Let's begin by

$$f(\mathbf{x}) = \mathbf{a}'\mathbf{x} = \sum_{i=1}^{n} a_i x_i;$$

evidently, the partial derivative of $f(\mathbf{x})$ with respect to x_i is just a_i; by stacking all the partial derivatives into a vector, the result is just the vector $\mathbf{a}$, and therefore

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} \mathbf{a}'\mathbf{x} = \mathbf{a}'$$

note that the familiar rule $\frac{\mathrm{d}}{\mathrm{d}x} ax = a$ is just a special case when a and x are scalars.

As for the quadratic form

$$f(\mathbf{x}) = \mathbf{x}' A \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j;$$

it can be proven easily (but it's rather boring) that

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\mathbf{x}'A\mathbf{x} = \mathbf{x}'(A + A')$$

and of course if A is symmetric (as in most cases), then $\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\mathbf{x}'A\mathbf{x} = 2 \cdot \mathbf{x}'A$. Again, note that the scalar case $\frac{\mathrm{d}}{\mathrm{d}x}ax^2 = 2ax$ is easy to spot as a special case.

One last thing: the convention by which differentiation expands "by row" turns out to be very useful because it makes the chain rule for the derivatives "just work" automatically. For example, suppose you have $\mathbf{y} = A\mathbf{x}$ and $\mathbf{z} = B\mathbf{y}$; of course, if you need the derivative of $\mathbf{z}$ with respect to $\mathbf{x}$ you may proceed by defining $C = B \cdot A$ and observing that

$$\mathbf{z} = B(A\mathbf{x}) = C\mathbf{x} \Longrightarrow \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = C$$

but you may also get the same result via the chain rule, as

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = B \cdot A = C.$$

1.A.2 Vector spaces

Here we will draw heavily on the fact that a vector with n elements can be thought of as a point in an n-dimensional space: a scalar is a point on the real line, a vector with two elements is a point on a plane, and so on. Actually, the notation $\mathbf{x} \in \mathbb{R}^n$ is a concise way of saying that $\mathbf{x}$ has n elements.

There are two basic operations we can perform on vectors: (i) multiplying a vector by a scalar and (ii) summing two vectors. In both cases, the result you get is another vector. Therefore, if you consider k vectors with n elements each, it makes sense to define an operation called a **linear combination** of them:

$$\mathbf{z} = \lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_k\mathbf{x}_k = \sum_{j=1}^{k} \lambda_j\mathbf{x}_j;$$

note that the above could have been written more compactly in matrix notation as $\mathbf{z} = \mathbf{X}\lambda$, where $\mathbf{X}$ is a matrix whose columns are the vectors $\mathbf{x}_j$ and λ is a k-element vector.

The result is, of course, an n-element vector, that is a point in $\mathbb{R}^n$. But the k vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are also a cloud of k points in $\mathbb{R}^n$; so we may ask ourselves if there is any kind of geometrical relationship between $\mathbf{z}$ and $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$.

Begin by considering the special case $k = 1$. Here $\mathbf{z}$ is just a multiple of $\mathbf{x}_1$; longer, if $|\lambda_1| > 1$, shorter otherwise; mirrored across the origin if $\lambda_1 < 0$, in the same quadrant otherwise. Easy, boring. Note that, if you consider the set of all the vectors $\mathbf{z}$ you can obtain by all possible choices for λ_1, you get a straight line going through the origin, and of course $\mathbf{x}_1$; this set of points is called the space **spanned**, or **generated** by $\mathbf{x}_1$; or, in symbols, $\mathrm{Sp}(\mathbf{x}_1)$. It's important to note that

this won't work if $\mathbf{x}_1 = \mathbf{0}$: in this case, $\text{Sp}(\mathbf{x}_1)$ is not a straight line, but rather a point (the origin).

If you have two vectors, instead, the standard case occurs when they are not aligned with respect to the origin. In this case, $\text{Sp}(\mathbf{x}_1, \mathbf{x}_2)$ is a plane and $\mathbf{z} = \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2$ is a point somewhere on that plane. Its exact location depends on λ_1 and λ_2, but note that

- by a suitable choice of λ_1 and λ_2, no point on the plane is unreachable;
- no matter how you choose λ_1 and λ_2, you can't end up outside the plane.

However, if $\mathbf{x}_2$ is a multiple of $\mathbf{x}_1$, then $\mathbf{x}_2 \in \text{Sp}(\mathbf{x}_1)$ and $\text{Sp}(\mathbf{x}_1, \mathbf{x}_2) = \text{Sp}(\mathbf{x}_1)$, that is a line, and not a plane. In this case, considering $\mathbf{x}_2$ won't make $\text{Sp}(\mathbf{x}_1)$ "grow" in dimension, since $\mathbf{x}_2$ is already contained in it, so to speak.

In order to fully generalise the point, we use the concept of **linear independence**: a set of k vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is said to be linearly independent if none of them can be expressed as a linear combination of the remaining ones.[35] The case I called "standard" a few lines above happens when $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent.

1.A.3 Rank of a matrix

If we take k vectors with n elements each and we arrange them side by side so as to form an $(n \times k)$ matrix (call it $\mathbf{X}$), the maximum number of linearly independent columns of $\mathbf{X}$ is the **rank** of $\mathbf{X}$ ($\text{rk}(\mathbf{X})$ in formulae). The rank function enjoys several nice properties:[36]

1. $0 \leq \text{rk}(\mathbf{X}) \leq k$ (by definition);
2. $\text{rk}(\mathbf{X}) = \text{rk}(\mathbf{X}')$;
3. $0 \leq \text{rk}(\mathbf{X}) \leq \min(k, n)$ (by putting together the previous two); but if $\text{rk}(\mathbf{X}) = \min(k, n)$, and the rank hits its maximal value, the matrix is said to have "full rank";
4. $\text{rk}(A \cdot B) \leq \min(\text{rk}(A), \text{rk}(B))$; but in the special case when $A' = B$, then equality holds, and $\text{rk}(B'B) = \text{rk}(BB') = \text{rk}(B)$.

We can use the rank function to measure the dimension of the space spanned by $\mathbf{X}$. For example, if $\text{rk}(\mathbf{X}) = 1$, then $\text{Sp}(\mathbf{X})$ is a line, if $\text{rk}(\mathbf{X}) = 2$, then $\text{Sp}(\mathbf{X})$ is a plane, and so on. This number may be smaller than the number of columns of $\mathbf{X}$.

[35] The usual definition is that $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly independent if no linear combination $\sum_{j=1}^{k} \lambda_j \mathbf{x}_j$ is zero unless all the λ_j are zero. The reader is invited to check that the two definitions are equivalent.

[36] I'm not proving them for the sake of brevity: if you're curious, have a look at https://en.wikipedia.org/wiki/Rank_(linear_algebra).

A result we will not use very much (only in chapter 6), but is quite useful to know in more advanced settings is that, if you have a matrix A with n rows, k columns and rank r, it is always possible to write it as

$$A = UV'$$

where U is $(n \times r)$, V is $(k \times r)$, and both have rank r. For example, the matrix

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

can be written as

$$A = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix}$$

where

$$U = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad V = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Note that such decomposition is not unique: there are infinitely many pairs of matrices that satisfy the decomposition above. The example above would have worked just as well with

$$P = \begin{bmatrix} -10 \\ 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad Q = \begin{bmatrix} -0.1 \\ 0 \end{bmatrix}$$

and the reader can easily verify that $A = UV' = PQ'$.

1.A.4 Rank and inversion

A square matrix A is said to be "invertible" if there is another matrix B such that $AB = BA = I$, where I is the identity matrix. if B exists, it's also notated as A^{-1} and called the inverse of A; otherwise, A is said to be singular.

The mathematically accepted way to say "A is non-singular" is by writing $|A| \neq 0$, where the symbol $|A|$ is used for the **determinant** of the matrix A, which is a scalar function such that $|A| = 0$ if and only if A is singular.[37]

The concept of matrix inversion becomes quite intuitive if you look at it geometrically. Take a vector $\mathbf{x}$ with n elements. If you pre-multiply it by a square matrix A you get another vector with n elements:

$$\mathbf{y} = A\mathbf{x};$$

[37] You almost never need to compute a determinant by hand, so I'll spare you its definition. If you're curious, there's always Wikipedia.

in practice, A defines a displacement that takes you from a point in $\mathbb{R}^n$ to a point in the same space. Is it possible to "undo" this movement? If A takes you from $\mathbf{x}$ to $\mathbf{y}$, is there a matrix B that performs the return trip? If such a matrix exists, then

$$\mathbf{x} = B\mathbf{y}.$$

The only way to guarantee that this happens for every pair of vectors $\mathbf{x}$ and $\mathbf{y}$ is to have

$$AB = BA = I.$$

Now, note that $\mathbf{y}$ is a linear combination of the columns of A; intuition suggests that all the n separate pieces of information originally contained in $\mathbf{x}$ can get preserved during the trip only if the rank of A is n. In fact, it can be proven formally that if A is an $(n \times n)$ matrix, then $\mathrm{rk}(A) = n$ is a necessary and sufficient condition for A^{-1} to exist: for square matrices, full rank is the same thing as invertibility.

The world of matrix algebra is populated with results that appear unintuitive when you're used to the algebra of scalars. A notable one is: any matrix A (even singular ones; even non-square ones) admits a matrix B such that $ABA = A$ and $BAB = B$; B is called the "Moore-Penrose" pseudo-inverse, or "generalised" inverse. For example, the matrix $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ is singular, and therefore has no inverse. However, it's got a pseudo-inverse, which is A itself. In fact, all projection matrices are their own pseudo-inverses.

Roger Penrose has been awarded the 2020 Nobel prize for physics. Not for the generalised inverse, but you get the idea of how brilliant the guy is.

Computing an inverse in practice is very boring: this is one of those tasks that computers are very good at, while humans are not. The only interesting cases for which I'm giving you instructions on how to invert a matrix by hand are:

1. when A is 2×2. In this case, it's easy to memorise the explicit formula for the inverse:
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix};$$
2. when A is block-diagonal, that is it can be written as
$$A = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_m \end{bmatrix};$$
if the inverse exists, it has the same structure:
$$A^{-1} = \begin{bmatrix} A_1^{-1} & 0 & \cdots & 0 \\ 0 & A_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_m^{-1} \end{bmatrix}.$$

There are many nice properties that invertible matrices enjoy. For example:

- the inverse, if it exists, is unique; that is, if $AB = I = AC$, then $B = C$;
- the inverse of a symmetric matrix is also symmetric;
- the transpose of the inverse is the inverse of the transpose $((A')^{-1} = (A^{-1})')$;
- if a matrix is positive definite (see section 1.A.7), then its inverse is positive definite too;
- if A is invertible, then the only solution to $A\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$; conversely, if A is singular, then there exists at least one non-zero vector such that $A\mathbf{x} = \mathbf{0}$.
- if A and B are invertible, then $(AB)^{-1} = B^{-1}A^{-1}$.

1.A.5 Step-by-step derivation of the sum of squares function

The function we have to differentiate with respect to β is

$$L(\beta) = \mathbf{e}(\beta)'\mathbf{e}(\beta);$$

the elegant way to do this is by using the chain rule:

$$\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial \mathbf{e}(\beta)'}{\partial \beta}\mathbf{e}(\beta) + \left[\mathbf{e}(\beta)'\frac{\partial \mathbf{e}(\beta)}{\partial \beta}\right]' = 2 \cdot \frac{\partial \mathbf{e}(\beta)'}{\partial \beta}\mathbf{e}(\beta);$$

the reason why we have to transpose the second element of the sum in the equation above is conformability: you can't sum a row vector and a column vector.

Therefore, since $\mathbf{e}(\beta)$ is defined as $\mathbf{e}(\beta) = \mathbf{y} - \mathbf{X}\beta$, we have

$$\frac{\partial \mathbf{e}(\beta)}{\partial \beta} = -\mathbf{X}$$

and the necessary condition for minimisation is $\frac{\partial L(\beta)}{\partial \beta} = -2\mathbf{X}'\mathbf{e} = \mathbf{0}$, which of course implies equation 1.9.

1.A.6 Numerical collinearity

Collinearity can sometimes be a problem as a consequence of finite precision of computer algebra.[38] For example, suppose you have the following matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 4+\epsilon \end{bmatrix}$$

[38] If you find this kind of things intriguing I cannot but recommend chapter 1 in Epperson (2013); actually, the whole book!

ϵ	$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})$
0.1	$\begin{bmatrix} 1 & 0 \\ 9.09495e-13 & 1 \end{bmatrix}$
0.01	$\begin{bmatrix} 1 & -1.16415e-10 \\ 0 & 1 \end{bmatrix}$
0.001	$\begin{bmatrix} 1 & 7.45058e-09 \\ 2.23517e-08 & 1 \end{bmatrix}$
0.0001	$\begin{bmatrix} 0.999999 & 0 \\ 9.53674e-07 & 1 \end{bmatrix}$
1e-05	$\begin{bmatrix} 0.999756 & 0 \\ 0 & 0.999878 \end{bmatrix}$
1e-06	$\begin{bmatrix} 0.992188 & 0 \\ 0 & 0.992188 \end{bmatrix}$
1e-07	$\begin{bmatrix} 0.5 & 0 \\ 0.5 & 1 \end{bmatrix}$
1e-08	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

Table 1.3: Numerical precision

For $\epsilon > 0$, the rank of $\mathbf{X}$ is, clearly, 2; nevertheless, if ϵ is a very small number, a computer program[39] goes berserk; technically, this situation is known as **quasi-collinearity**. To give you an example, I used gretl to compute $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})$ for decreasing values of ϵ; Table 1.3 contains the results. Ideally, the right-hand side column in the table should only contain identity matrices. Instead, results are quite disappointing for $\epsilon = 1e-05$ or smaller. Note that this is not a problem specific to gretl (which internally uses the very high quality LAPACK routines), but a consequence of finite precision of digital computers.

This particular example is easy to follow, because $\mathbf{X}$ is a small matrix. But if that matrix had contained hundreds or thousands of rows, things wouldn't have been so obvious.

1.A.7 Definiteness of square matrices

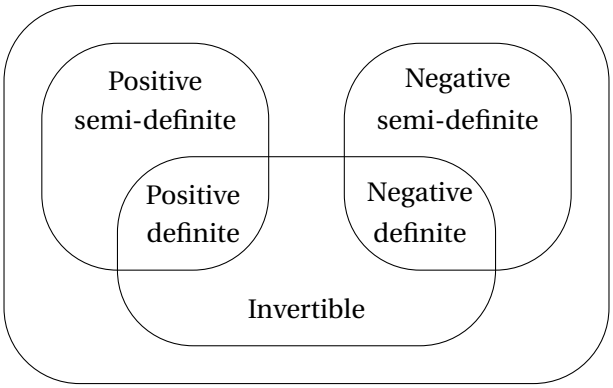
A square matrix B is *positive definite* (pd for short) if the quadratic form $\mathbf{x}'B\mathbf{x}$ returns a positive number for any choice of $\mathbf{x}$ and positive semi-definite (psd for short) if $\mathbf{x}'B\mathbf{x} \geq 0$.

If B is positive (semi-)definite, then $-B$ is negative (semi-)definite. Of course, it is entirely possible that $\mathbf{x}'B\mathbf{x}$ can take positive or negative values depending on $\mathbf{x}$, in which case B is said to be *indefinite*. If B is semi-definite and invertible, then it's also definite. Figure 1.4 may be helpful.[40] Ofter, the expressions "positive definite" and "positive semi-definite" are abbreviated as "pd" and "psd", re-

[39] I should say: a computer program not explicitly designed to operate with arbitrary precision. There are a few, but no statistical package belongs to this category, for very good reasons.

[40] In fact, figure 1.4 contains a slight inaccuracy. Finding it is left as an exercise to the reader.

Figure 1.4: Square matrices



spectively.

There are many interesting facts on psd matrices. A nice one is: if a matrix H exists such that $B = HH'$, then B is psd.[41] This, for example, gives you a quick way to prove that that I is pd and $\mathbf{P_X}$ is psd.

Some people use a special symbol for the cases when the difference between two matrices is pd or psd. The expression $A \succ B$ means that $A - B$ is pd, while $A \succeq B$ means that $A - B$ is psd.

1.A.8 A few more results on projection matrices

Consider an n-dimensional space and a matrix $\mathbf{X}$ with n rows, k columns and full rank. Of course, the columns of this matrix define a k-dimensional subspace that we call $\text{Sp}(\mathbf{X})$.

We would like to say something about the space spanned by matrices defined as $\mathbf{W} = \mathbf{X} \cdot A$. There are two cases of interest. The first one arises when A is square and invertible: in this case, $\text{Sp}(\mathbf{X}) = \text{Sp}(\mathbf{W})$, so $\mathbf{P_X} = \mathbf{P_W}$. The result is easy to prove: for any $\mathbf{y} \in \text{Sp}(\mathbf{X})$, there must be a vector $\boldsymbol{\beta}$ such that $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$. But then, by choosing $\boldsymbol{\gamma} = A^{-1}\boldsymbol{\beta}$, it's easy to see that $\mathbf{y}$ can also be written as $\mathbf{W}\boldsymbol{\gamma}$ and therefore $\mathbf{y} \in \text{Sp}(\mathbf{W})$; by a similar reasoning, it can also be proven that if $\mathbf{y} \in \text{Sp}(\mathbf{W})$ then $\mathbf{y}$ also belongs to $\text{Sp}(\mathbf{X})$, and therefore

$$\mathbf{y} \in \text{Sp}(\mathbf{X}) \iff \mathbf{y} \in \text{Sp}(\mathbf{W})$$

and the two sets are the same.

The equivalence of the two projection matrices can also be proven directly by using elementary results on matrix inversion (see section 1.A.4):

$$\mathbf{P_W} = \mathbf{X}A\left[A'\mathbf{X}'\mathbf{X}A\right]^{-1}A'\mathbf{X}' = \mathbf{X}AA^{-1}\left[\mathbf{X}'\mathbf{X}\right]^{-1}(A')^{-1}A'\mathbf{X}' = \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' = \mathbf{P_X}.$$

[41] Easy to prove, too. Try.

Let's now consider the case when A is a matrix with rank less than k (for example, a column vector). Evidently, any linear combination of the columns of $\mathbf{W}$ is also a linear combination of the columns of $\mathbf{X}$, and therefore each column of $\mathbf{W}$ is an element of $\text{Sp}(\mathbf{X})$. As a consequence, any vector that belongs to $\text{Sp}(\mathbf{W})$ also belongs to $\text{Sp}(\mathbf{X})$.

The converse is not true, however: some elements of $\text{Sp}(\mathbf{X})$ do not belong to $\text{Sp}(\mathbf{W})$ (allow me to skip the proof). In short, $\text{Sp}(\mathbf{W})$ is a subset of $\text{Sp}(\mathbf{X})$; in formulae, $\text{Sp}(\mathbf{W}) \subset \text{Sp}(\mathbf{X})$.

A typical example occurs when $\mathbf{W}$ contains some of the columns of $\mathbf{X}$, but not all. Let's say, without loss of generality, that $\mathbf{W}$ contains the leftmost $k-p$ columns of $\mathbf{X}$. In this case, the matrix A can be written as

$$A = \begin{bmatrix} \mathbf{I} \\ 0 \end{bmatrix}$$

where the identity matrix above has $k-p$ rows and columns, and the 0 zero matrix below has p and $k-p$ columns.

	$\mathbf{P_W}$	$\mathbf{M_W}$	$\mathbf{P_X}$	$\mathbf{M_X}$
$\mathbf{P_W}$	$\mathbf{P_W}$	0	$\mathbf{P_W}$	0
$\mathbf{M_W}$	0	$\mathbf{M_W}$	$\mathbf{P_X} - \mathbf{P_W}$	$\mathbf{M_X}$
$\mathbf{P_X}$	$\mathbf{P_W}$	$\mathbf{P_X} - \mathbf{P_W}$	$\mathbf{P_X}$	0
$\mathbf{M_X}$	0	$\mathbf{M_X}$	0	$\mathbf{M_X}$

Important: it is assumed that $\text{Sp}(\mathbf{W}) \subset \text{Sp}(\mathbf{X})$. All products commute.

Table 1.4: Projection matrices "multiplication table"

In this situation, the property $\mathbf{P_X W} = \mathbf{P_X X} A = \mathbf{X} A = \mathbf{W}$ implies some interesting consequences on the projection matrices for the spaces $\text{Sp}(\mathbf{W})$ and $\text{Sp}(\mathbf{X})$, that can be summarised in the "multiplication table" shown in Table 1.4. The reader is invited to prove them; it shouldn't take long.

Chapter 2

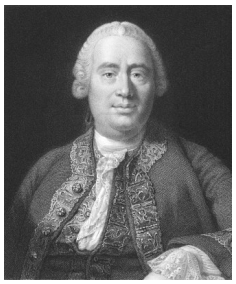
Some statistical inference

2.1 Why do we need statistical inference?

So far, we have followed a purely descriptive approach, trying to find the smartest possible method for compressing as much information as we can from the original data into a small, manageable container.

However, we are often tempted to read the evidence we have in a broader context. Strictly speaking, any statistic we compute on a body of data tells us something about those data, and nothing else. Thus, the OLS coefficients we compute are nothing but a clever way to squeeze the relevant information out of our dataset; however, we would often like to interpret the size and the magnitudes of the coefficients that we get out of our OLS calculations as something that tells us a more general story. In other words, we would like to perform what in philosophical language is known as **induction**.

In the 18th century, the Scottish philosopher David Hume famously argued against induction.



DAVID HUME

> When it is asked, *What is the nature of all our reasonings concerning matter of fact?* the proper answer seems to be, that they are founded on the relation of cause and effect. When again it is asked, *What is the foundation of all our reasonings and conclusions concerning that relation?* it may be replied in one word, Experience. But if we still carry on our sifting humour, and ask, *What is the foundation of all conclusions from experience?* this implies a new question, which may be of more difficult solution and explication.[1]

Inductive reasoning can be broadly formalised as follows:

1. Event X has always happened.

[1] D. Hume, *An Enquiry Concerning Human Understanding* (1748).

2. The future will be like the past.

3. Therefore, event X will happen in the future.

Even if you could establish statement 1 beyond any doubt, statement 2 is basically an act of faith. You may believe in it, but there is no rational argument one could convincingly use to support it. And yet, we routinely act on the premise of statement 2. Hume considered our natural tendency to rely on it as a *biological* feature of the human mind. And it's a *good thing*: if we didn't have this fundamental psychological trait, we'd be unable to learn anything at all;[2] the only problem is, it's logically unfounded.

Statistical inference is a way to make an inductive argument more rigorous by replacing statement number 2 with some assumptions that translate into formal statements our tendency to generalise, by introducing *uncertainty* into the picture. Uncertainty is the concept we use to handle situations in which our knowledge is partial. So for example we cannot predict which number will show up when we roll a die, although in principle it would be perfectly predictable, given initial conditions, using the laws of physics. We simply don't have the resources to perform such a monster computation, so we represent our imperfect knowledge through the language of probability, or, more correctly, via a probabilistic model; then, we assume that the same model will keep being valid in the future. Therefore, if we rolled a die 10 times and obtained something like

$$\mathbf{x} = [1, 5, 4, 6, 3, 3, 2, 6, 3, 4]$$

we would act on the assumption that if we keep rolling the same die we will observe something that, in our eyes, looks "just as random" as $\mathbf{x}$. To put it differently, our aim will not be to predict exactly which side the die will land on, but rather to make statements on how surprising or unsurprising certain outcomes will be.

Therefore, we use the idea of a **Data Generating Process**, or DGP. We assume that the DGP is the mechanism that Nature (or any divinity of your choice) has used to produce the data we observe, and will continue doing so for the data we have not observed yet. By describing the DGP via a mathematical structure (usually, but not necessarily, via a probability distribution), we try to come up with statistics $T(\mathbf{x})$ whose aim is not as much to describe the available data $\mathbf{x}$, but rather to describe the DGP that generated $\mathbf{x}$, and therefore, to provide us with some insight that goes beyond merely descriptive statistics.

Of course, in order to accomplish such an ambitious task, we need a set of tools to represent imperfect knowledge in a mathematical way. This is why we need probability theory.

[2] To be fair, this only applies to what Immanuel Kant called "synthetic" propositions. But maybe I'm boring you?

2.2 A crash course in probability

Disclaimer: in this section, I'll just go quickly though a few concepts that the reader should already be familiar with; as a consequence, it is embarrassingly simplistic and probably quite misleading in many ways. The reader who wants to go for the real thing might want to read (in increasing order of difficulty): Gallant (1997); Bierens (2011); Davidson (1994); Billingsley (1986). Having said this, let's go ahead.

2.2.1 Probability and random variables

The concept of probability has been the object of philosophical debate for centuries. The *meaning* of probability is still open for discussion,[3] but fortunately the *syntax* of probability is clear and undisputed since the great Soviet mathematician Andrej Nikolaevič Kolmogorov made probability a proper branch of measure theory.

ANDREJ NIKOLAEVIČ KOLMOGOROV

The meaning I give to the word probability here is largely operational: probability is a number between 0 and 1 that we attach to something called an **event**. Loosely speaking, an event is a statement that in our eyes could be conceivably true or false. Formally, an event is defined as a subset of an imaginary set, called the **state space**, and usually denoted by the letter Ω, whose elements ω are all the states of the world that our mind can conceive as possible. Probability is a function of subsets of Ω, which obeys a few properties the reader should already know, such as

$$P(\Omega) = 1, \quad P(\emptyset) = 0, \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

and so forth. Event A can be defined as the subset of Ω including all the states ω in which a statement A is true, and only those. $P(A)$ is the measure of A, where the technical word "measure" is a generalisation of our intuitive notion of "extension" (length, area, volume).[4] The familiar laws of probability are simple consequences of the way usual set operations (complement, union, intersection) work; let's waste no time on those.[5]

Random variables are a convenient way to map events to segments on the real line. That is, a **random variable** X is defined as a *measurable* function from

[3] The interested reader might want to have a look at Freedman and Stark (2016), section 2. You can download it from https://www.stat.berkeley.edu/~stark/Preprints/611.pdf.

[4] Warning: not all subsets can be associated with a corresponding probability: some are "non-measurable". Providing a simple example is difficult, this is deep measure theory: google "Vitali set" if you're curious.

[5] In most cases, intuition will suffice. For tricky cases, I should explain what a σ-algebra is, but I don't think that this is the right place for this, really.

Ω to $\mathbb{R}$; or, to put it differently, for any ω in Ω you get a corresponding real number $X(\omega)$. The requisite of measurability is necessary to avoid paradoxical cases, and simply amounts to requiring that, if we define A as the the subset of Ω such that

$$a < X(\omega) \leq b \Longleftrightarrow \omega \in A,$$

then A is a proper event. In practice, it must be possible to define $P(a < X \leq b)$ for any a and b. I will sometimes adopt the convention of using the acronym "rv" for random variables.

There are two objects that a random variable comes equipped with: the first is its **support**, which is the subset of $\mathbb{R}$ with all the values that X can take; in formulae, $X : \Omega \mapsto S \subseteq \mathbb{R}$, and the set S is sometimes indicated as $\mathrm{S}(X)$. For a six-sided die, $\mathrm{S}(X) = \{1,2,3,4,5,6\}$; if X is the time before my car breaks down, then $\mathrm{S}(X) = [0,\infty)$, and so on.

The other one is its **distribution function**, or **cumulative distribution function** (often abbreviated as **cdf**), defined as

$$F_X(a) = P(X \leq a),$$

which of course makes it easy to compute the probability of X being inside an interval as

$$P(a < X \leq b) = F_X(b) - F_X(a).$$

By their definition, cdfs enjoy three basic properties:

- $\lim_{a \to -\infty} F_X(a) = 0$;
- $\lim_{a \to \infty} F_X(a) = 1$;
- if $b > a$, then $F_X(b) \geq F_X(a)$; that is, $F_X(\cdot)$ is non-decreasing.

Apart from this, there's very little that can be said in general. However, in many cases it is assumed that $F_X(a)$ has a known functional form, which depends on a vector of parameters θ.

Two special cases are of interest:

1. The cdf is a function that goes up in steps; the support is a countable set, and the corresponding rv is said to be **discrete**; for every member of the support x it is possible to define $p(x) = P(X = x) > 0$; the function $p(x)$ is the so-called probability function.

2. The cdf is everywhere differentiable; the support is an interval on $\mathbb{R}$, (possibly, the whole real line), and the corresponding rv is said to be **continuous**; the derivative of $F_x(a)$ is known as the **density function** of X, or $f_X(a)$ and therefore, by definition,

$$P(a < X \leq b) = \int_a^b f_X(z)\mathrm{d}z;$$

in most cases, when the meaning is clear from the context, we just write the density function for X as $f(x)$.[6]

In the rest of the book, I will mostly use continuous random variables for examples; hopefully, generalisations to discrete rvs should be straightforward.

Of course, you can collect a bunch of random variables into a vector, so you have a **multivariate random variable**, or **random vector**. The multivariate extension of the concepts I sketched above is a little tricky from a technical viewpoint, but for our present needs intuition will again suffice. I will only mention that for a multivariate random variable $\mathbf{x}$ with k elements you have that

$$F_{\mathbf{x}}(\mathbf{a}) = P\left[(x_1 \leq a_1) \cap (x_2 \leq a_2) \cap \ldots \cap (x_k \leq a_k)\right]$$

If all the k elements of $\mathbf{x}$ are continuous random variables, then you can define the **joint density** as

$$f_{\mathbf{x}}(\mathbf{z}) = \frac{\partial^k F_{\mathbf{x}}(\mathbf{z})}{\partial z_1 \partial z_2 \cdots \partial z_k}.$$

The **marginal density** of the i-th element of $\mathbf{x}$ is just the density of x_i taken in isolation. For example, suppose you have a trivariate random vector $\mathbf{w} = [X, Y, Z]$: the marginal density for Y is

$$f_Y(a) = \int_{\mathrm{S}(X)} \int_{\mathrm{S}(Z)} f_{\mathbf{w}}(x, a, z)\, \mathrm{d}z\, \mathrm{d}x$$

2.2.2 Independence and conditioning

If $P(A)$ is the probabilistic evaluation we give of A, we may ask ourselves if we would change our mind when additional information becomes available. If we receive the news that event B has occurred, then we can safely exclude the event $\overline{B}$ from Ω.[7] In fact, after receiving the message "B is true", our state space Ω shrinks to B, because states in which B is false are no longer conceivable as possible.

The consequence for A is that the subset $A \cap \overline{B}$ is no longer possible; hence, we must update our probability measure so that $P(B)$ becomes 1, and consequently[8] $P(A)$ must be revised as

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

[6] I imagine the reader doesn't need reminding that the density at x is *not* the probability that $X = x$; for continuous random variables, the probability is only defined for intervals by the formula in the text, from which it follows that $P(X = x)$ is 0.

[7] When speaking about sets, I use the bar $\overline{\ }$ to indicate the complement.

[8] Technically, it's more complicated that this, because $P(B)$ may be 0, in which case the definition has to be adapted and becomes more technical. If you're interested, chapter 10 in Davidson (1994) is absolutely splendid.

You read the left-hand side of this definition as "the probability of A given B", which is of course what we call **conditional probability**. It should be clear that

$$P(A) = P(A|B) \iff P(A \cap B) = P(A) \cdot P(B)$$

Which means: "if you don't need to revise your evaluation of A after having received some message about B, then A and B have nothing to do with each other"; in this situation, A and B are said to be independent, and we write $A \perp\!\!\!\perp B$, so independence can be thought of as lack of mutual information. Note that independence is a symmetric concept: if A is independent of B, then B is independent of A, and vice versa.

Equation (2.1) has the following implication:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A),$$

so that

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

The expression above is interesting for many reasons. One is: in general, $P(A|B) \neq P(B|A)$, so, for example, the probability of dying from COVID if you're not vaccinated is not the same thing as the probability that someone who died from COVID was a no-vaxxer (think about it). Another reason is that this expression is the cornerstone of an approach to statistics known as **Bayesian**, after the English statistician and clergyman Reverend Thomas Bayes, who lived in the 18th century; I'm not going to use anything Bayesian in this book, but Bayesian methods are getting increasingly popular in many areas of econometrics.

The same concept can be applied to random variables. If

$$F_Y(z) = F_Y(z|a < X \leq b)$$

for any a and b, then evidently X carries no information about Y, and we say that the two random variables are independent: $Y \perp\!\!\!\perp X$. If this is not the case, it makes sense to consider the **conditional distribution** of Y on X, which describes our uncertainty about Y *once we have information about* X. So for example, if Y is the yearly expenditure on food by a household and X is the number of its components, it seems safe to say that $F(Y|X > 6)$ should be different from $F(Y|X < 3)$, because more people eat more food.

The case $a = b$ is important[9], because it gives us a tool for evaluating probabilities about Y in a situation when X is not uncertain at all, because in fact we observe its realisation $X = x$. In this case, we can define the *conditional* density as

$$f_{Y|X=x}(z) = \frac{f_{Y,X}(z,x)}{f_X(x)} \tag{2.2}$$

and when what we mean is clear from the context, we simply write $f(y|x)$.

Therefore, in many cases we will use the intuitive notion of X, the set of random variables we are conditioning Y on, as being "the relevant information

[9] Albeit special: a moment's reflection is enough to convince the reader that if X is continuous, the event $X = x$ has probability 0, and our naïve definition of conditioning breaks down. But again, treating the subject rigorously implies using measure theory, σ-algebras and other tools that I'm not willing to use in this book.

about Y that we have"; in certain contexts, this idea is expressed by the notion of an **information set**. However, a formalised description of this idea is, again, far beyond the scope of this book and I am contented to leave this to the reader's intuition.

2.2.3 Expectation

The expectation of a random variable is a tremendously important concept. A rigorous definition, valid in all cases, would require a technical tool called Lebesgue integral, that I'd rather avoid introducing. Luckily, in the two elementary special cases listed in section 2.2.1, its definition is quite simple:

$$\mathrm{E}[X] = \sum_{x \in \mathrm{S}(X)} x \cdot p(x) \qquad \text{for discrete rvs} \tag{2.3}$$

$$\mathrm{E}[X] = \int_{\mathrm{S}(X)} z \cdot f_X(z)\,\mathrm{d}z \qquad \text{for continuous rvs.} \tag{2.4}$$

The expectation of a function of X, $\mathrm{E}[h(X)]$, is defined simply as

$$\mathrm{E}[h(X)] = \int_{\mathrm{S}(X)} h(z) \cdot f_X(z)\,\mathrm{d}z$$

for continuous random variables and the parallel definition for the discrete case is obvious. The extension to multivariate rvs should also be straightforward: the expectation of a vector is the vector of expectations.

Some care must be taken, since $\mathrm{E}[X]$ may not exist, even in apparently harmless cases.

Example 2.1

If X is a uniform continuous random variable between 0 and 1, its density function is $f(x) = 1$ for $0 < x \leq 1$. Its expectation is easy to find as

$$\mathrm{E}[X] = \int_0^1 x \cdot 1\,\mathrm{d}x = \left[\frac{x^2}{2}\right]_0^1 = 1/2;$$

however, it's not difficult to prove that $\mathrm{E}[1/X]$ does not exist (the corresponding integral diverges):

$$\mathrm{E}[1/X] = \int_0^1 \frac{1}{x} \cdot 1\,\mathrm{d}x = \left[\log x\right]_0^1 = \infty.$$

However, it can be proven that if the support of X is finite, then $\mathrm{E}[X]$ is always bounded, and therefore exists. To be more specific: if $\mathrm{S}(X) = [a, b]$, where

a and b are finite, then $a < \mathrm{E}[X] < b$. The proof is easy and left to the reader as an exercise.

The expectation operator $\mathrm{E}[\cdot]$ is linear, and therefore we have the following simple rule for affine transforms (A and $\mathbf{b}$ must be non-stochastic):

$$\mathrm{E}[A\mathbf{x}+\mathbf{b}] = A\cdot\mathrm{E}[\mathbf{x}]+\mathbf{b} \tag{2.5}$$

For nonlinear transformation, things are not so easy. As a rule, $\mathrm{E}\left[g(X)\right] \neq g\left[\mathrm{E}[X]\right]$, and there's very little you can say in general.[10]

The expectation of the k-th power of X is called its k-th **moment**, so the first moment is $\mathrm{E}[X]$, the second moment is $\mathrm{E}\left[X^2\right]$ and so on. Of course, $\mathrm{E}[X^n]$ (with $n \geq 1$) may not exist, but if it does then $\mathrm{E}\left[X^{n-1}\right]$ is guaranteed to exist too.

The most egregious example of usefulness of moments is the definition of **variance**:[11] $\mathrm{V}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. The variance is always non-negative, and is the most widely used indicator of dispersion. Of course, in order to exist, the second moment of X must exist. Its multivariate generalisation is the **covariance matrix**, defined as

$$\mathrm{Cov}[\mathbf{x}] = \mathrm{E}\left[\mathbf{x}\mathbf{x}'\right] - \mathrm{E}[\mathbf{x}]\,\mathrm{E}[\mathbf{x}]'; \tag{2.6}$$

The properties of $\mathrm{Cov}[\mathbf{x}]$ should be well known, but let's briefly mention the most important ones: if $\Sigma = \mathrm{Cov}[\mathbf{x}]$, then

- Σ is symmetric;
- if $x_i \perp\!\!\!\perp x_j$, then $\Sigma_{ij} = 0$ (warning: the converse is not necessarily true);
- Σ is positive semi-definite.[12]

Definition 2.6 makes it quite easy to calculate the covariance matrix of an affine transform:[13]

$$\mathrm{Cov}[A\mathbf{x}+\mathbf{b}] = A\cdot\mathrm{Cov}[\mathbf{x}]\cdot A'. \tag{2.7}$$

Note that this result makes it quite easy to prove that if X and Y are independent rvs, then $\mathrm{V}[X+Y] = \mathrm{V}[X]+\mathrm{V}[Y]$ (hint: put X and Y into a vector and observe that its covariance matrix is diagonal).

2.2.4 Conditional expectation

The easiest way to see the conditional expectation of Y given X is by defining it as the expectation of Y with respect to $f(Y|X=x)$, that is

$$\mathrm{E}[Y|X=x] = \int_{\mathrm{S}(Y)} z\cdot f_{Y|X=x}(z)\,\mathrm{d}z.$$

[10] In fact, there is *something* more general we can say, when the transformation $g(X)$ is concave on the whole support of X: it's called *Jensen's lemma*. We will not use this result in this book, but the result is widely used in economics and econometrics; if you're interested, the idea is briefly explained in section 2.A.1.

[11] An alternative equivalent definition, perhaps more common, is $\mathrm{V}[X] = \mathrm{E}\left[(X-\mathrm{E}[X])^2\right]$.

[12] If you're wondering what "semi-definite" means, you may want to go back to section 1.A.7.

[13] The proof is an easy exercise, left to the reader.

If $f_{Y|X=x}(z)$ changes with x, the result of the integral (if it exists) should change with x too, so we may see $\mathrm{E}[y|x] = m(x)$ as a function of x. This function is sometimes called the **regression function** of Y on X.

Does $\mathrm{E}[y|x]$ have a closed functional form? Not necessarily, but if it does, it hopefully depends on a small number of parameters θ.

Example 2.2

Assume that you have a bivariate variable (Y,X) where Y is 1 if an individual catches COVID and 0 otherwise, and X is 1 if the same individual is vaccinated. Suppose that the joint probability is

	$X=0$	$X=1$
$Y=0$	*0.1*	*0.3*
$Y=1$	*0.3*	*0.3*

The probability of catching COVID among vaccinated people is $\frac{0.3}{0.3+0.3} = 50\%$, while for unvaccinated people it's $\frac{0.3}{0.1+0.3} = 75\%$. The same statement could have been stated in formulae as

$$\mathrm{E}[Y|X] = 0.75 - 0.25X,$$

which gives 0.5 if $X=1$ and 0.75 if $X=0$. The regression function of Y on X is linear ($\mathrm{E}[Y|X] = \theta_0 + \theta_1 X$), and it depends on the vector $\theta = [\theta_0, \theta_1]$.

Of course, if x is a random variable, $m(x)$ is too. Does it have an expectation? If so,

$$\mathrm{E}\left[\mathrm{E}\left[y|x\right]\right] = \mathrm{E}\left[y\right]. \tag{2.8}$$

This is called **law of iterated expectations**, and is in fact more general than it appears at first sight. For example, it applies to density functions too:

$$f(y) = \mathrm{E}\left[f(y|x)\right]$$

To continue with example 2.2, note that, since $\mathrm{E}[X] = 0.6$, $\mathrm{E}[Y] = \mathrm{E}[0.75 - 0.25 \cdot X] = 0.75 - 0.25 \cdot \mathrm{E}[X] = 0.7 - 0.25 \times 0.6 = 0.6$.

Example 2.3

As a more elaborate example, suppose that

$$\mathrm{E}[Y|X] = m(X) = 4X - 0.5X^2$$

and that $\mathrm{E}[X] = \mathrm{V}[X] = 1$. It follows that

$$\mathrm{E}[Y] = \mathrm{E}[m(X)] = 4 \cdot \mathrm{E}[X] - 0.5 \cdot \mathrm{E}\left[X^2\right] = 4 - 0.5 \cdot 2 = 3$$

where I used $\mathrm{V}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$.

It must be stressed that expressing the relationship between two random variables by means of the conditional expectation has no meaning on causal relationship. Example 2.2 above should not be taken to imply, by itself, that if you get vaccinated your chances of getting ill are lower, although the idea is very natural. More on this in Section 3.1.

2.3 Estimation

The best way to define the concept of an **estimator** is to assume that we observe some data $\mathbf{x}$, and that the DGP which generated $\mathbf{x}$ can be described by means of a vector of parameters θ. We assume to know nothing about θ, apart from the fact that it can be thought of as a vector with a certain number of elements (say, k), and that it belongs to a subset S of of $\mathbb{R}^k$ called the **parameter space**. An estimator is a statistic $\hat{\theta} = T(\mathbf{x})$ that should be "likely" to yield a value "close to" the parameters of interest θ.

To state the same idea more formally: since $\mathbf{x}$ is random and $\hat{\theta}$ is a function of $\mathbf{x}$, then $\hat{\theta}$ is a random variable too, and therefore it must have a support and a distribution function. Clearly, both will depend on those of $\mathbf{x}$, but ideally, we'd like to choose the function $T(\cdot)$ so that the support $\mathrm{S}(\hat{\theta})$ contains at least a neighbourhood of θ, and we'd like the probability of observing a realisation of $\hat{\theta}$ that is "near" θ, $P(\theta - \epsilon < \hat{\theta} < \theta + \epsilon) = P(|\hat{\theta} - \theta| < \epsilon)$, to be as close to 1 as possible.

The indispensable ingredient for evaluating those probabilities would be the distribution of $\hat{\theta} = T(\mathbf{x})$. However, it is almost always tremendously difficult to pin it down exactly, either because of the characteristics of $\mathbf{x}$, which could be a very complex random variable, or because the function $T(\cdot)$ could be very intricate. In fact, the cases when we're able to work out the exact distribution of $\hat{\theta}$ are exceptionally few. In very simple cases,[14] we may be able to compute $\mathrm{E}\left[\hat{\theta}\right]$ and perhaps even $\mathrm{V}\left[\hat{\theta}\right]$, which leads us to the well known concepts of unbiasedness and efficiency:

- the **bias** of $\hat{\theta}$ is the difference $\mathrm{E}\left[\hat{\theta}\right] - \theta$; therefore $\hat{\theta}$ is said to be **unbiased** if $\mathrm{E}\left[\hat{\theta}\right] = \theta$;
- $\hat{\theta}$ is more efficient than $\grave{\theta}$ if $\mathrm{V}\left[\grave{\theta}\right] - \mathrm{V}\left[\hat{\theta}\right] > 0$ (if both are unbiased).

The problem that makes these concepts not very useful is that, in many cases of interest, it's very hard, if not impossible, to compute the moments of $\hat{\theta}$ (in some cases, $\hat{\theta}$ may even possess no moments at all). So we need to use something else. Fortunately, asymptotic theory comes to the rescue.

2.3.1 Consistency

The estimator $\hat{\theta}$ is **consistent** if its probability limit is the parameter we want to estimate. To explain what this means, let us first define **convergence in proba-**

[14] Notably, when $\hat{\theta}$ is an affine function of $\mathbf{x}$.

bility:

$$X_n \xrightarrow{\text{p}} X \Longleftrightarrow \lim_{n\to\infty} P\left[|X_n - X| < \epsilon\right] = 1 \tag{2.9}$$

Also notated as $\text{plim}\,(X_n) = X$.

A description in words of the definition above is: given a sequence of random variables $X_1, X_2, \ldots, X_n$, we define a parallel sequence of events of the kind

$$|X_1 - X| < \epsilon, \quad |X_2 - X| < \epsilon, \quad \ldots, \quad |X_n - X| < \epsilon;$$

the sequence above can be read as a sequence of events, where $|X_i - X| < \epsilon$ means "X_i is more or less X".[15] Convergence in probability means that the sequence of probabilities for those events tends to 1; that is, the probability of X_n and X being "substantially different" becomes negligible if n is large.[16]

In general, the limit X could be a random variable, but we'll be mostly interested in the case when the limit is a constant: if $X_n \xrightarrow{\text{p}} a$, the chances of X_n being far from a become zero, and therefore the cdf of X_n tends to a step function which is 0 before a and 1 after it. Or, if X_n is continuous, the density function $f(X_n)$ collapses to a point.

This is exactly what happens, in many circumstances, when we compute the sample average in a data set. Imagine you have n observations: you can compute the average of the observations you have as they become available; that is,

$$\bar{X}_1 = X_1, \quad \bar{X}_2 = \frac{X_1 + X_2}{2}, \quad \bar{X}_3 = \frac{X_1 + X_2 + X_3}{3}, \cdots;$$

does the sequence $\bar{X}_n$ have a limit in probability? Or, in other words, if n is large enough, do we have good chances that $\bar{X}_n$ will be a number arbitrarily near to *something*? The question may sound abstract and technical, but in fact this is something that we implicitly do all the time, when we try something many times in the hope that our knowledge stabilises with repetition.

The conditions that must occur for this idea to make sense are studied by the so-called **Laws of Large Numbers**, or LLNs for short.[17] There are many different LLNs, that cover different cases. Basically, there are three dimensions to the problem that must be considered:

1. How heterogeneous are the X_i variables?
2. Are the X_i variables independent?
3. Can we assume the existence of at least some of the moments?

The simplest version of the LLN is due to the Soviet mathematician Aleksandr Khinchin, and sets very strong bounds on the first two conditions and

[15] Where $\epsilon > 0$ is the mathematically respectable way of saying "more or less".

[16] The curious reader might be interested in knowing that there are several other ways to define a similar concept. A particularly intriguing one is the so-called "almost sure" convergence.

[17] Technically, these are the *weak* LLNs. The strong version uses a different concept of limit.

relaxes the third one as much as possible: if $x_1, x_2, \ldots, x_n$ are independent and identically distributed (iid for short) and $\mathrm{E}[x_i] = m$, then $\bar{X} \xrightarrow{\mathrm{p}} m$. Other versions exist: for example, a different version of the LLN can be used if observations are not independent, but in that case more stringent assumptions are needed; allow me to skip these complications. For the curious, an example is provided in section 2.A.3.

Example 2.4

Let's toss a coin n times. The random variable representing the i-th toss is x_i, which is assumed to obey the following probability distribution (often referred to as a Bernoulli distribution*):*

$$x_i = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1-\pi \end{cases}$$

Note that the probability π is assumed to be the same for all x_i; that is, the coin we toss does not change its physical properties during the experiment. Moreover, it is safe to assume that what happens at the i-th toss has no consequences on all the other ones. In short, the x_i random variables are iid.

Does x_i have a mean? Yes: $\mathrm{E}[x_i] = 1 \cdot \pi + 0 \cdot (1-\pi) = \pi$. *Together with the iid property, this is enough for invoking the LLN and establishing that $\bar{X} = \hat{p} \xrightarrow{\mathrm{p}} \pi$. Therefore, we can take the empirical frequency $\hat{p}$ as a consistent estimator of the true probability π.*

The LLN becomes enormously powerful when coupled with another wonderful result, which is a special case of a powerful tool called **Slutsky's Theorem**, that I'm not exposing in full here. If $X_n \xrightarrow{\mathrm{p}} a$ and $g(\cdot)$ is continuous at a, then $g(X_n) \xrightarrow{\mathrm{p}} g(a)$ (note how much easier this property makes it to work with probability limits rather than expectations).

In the context of estimation, obviously we will want our estimators to be consistent:

$$\hat{\theta} \xrightarrow{\mathrm{p}} \theta \iff \lim_{n\to\infty} P\left[|\hat{\theta} - \theta| < \epsilon\right] = 1; \tag{2.10}$$

that is, we will want to use as estimators statistics that become increasingly unlikely to be grossly wrong. Fortunately, the combination of the LLN and Slutsky's Theorem provides a very nice way to devise estimators that are consistent by construction. If the average has a probability limit that is a continuous, invertible function of the parameter we want, we just apply a suitable transformation the the average and we're done: so for example if $\mathrm{E}[x_i] = 1/\theta$, then $\hat{\theta} = 1/\bar{X}$; if $\mathrm{E}[x_i] = e^{\theta}$, then $\hat{\theta} = \log(\bar{X})$; if $\mathrm{E}[x_i] = \theta^2$, then $\hat{\theta} = \sqrt{\bar{X}}$; and so on.

More generally, the extension to the case when θ is a vector is technically messier, but conceptually identical. This is known as the **method of moments**: it is by no means the only one used in inferential statistics, but it will suffice for our purposes. The core intuition that motivates it is relatively straightforward:

1. express the moments of the observables as continuous functions of the parameters of interest θ: $\mathbf{m} = m(\theta)$.
2. Estimate $\mathbf{m}$ via the corresponding sample moments $\hat{\mathbf{m}}$, using the LLN, so that $\hat{\mathbf{m}} \xrightarrow{\text{p}} \mathbf{m}$.
3. Estimate θ by inverting the correspondence between parameters and moments: $\hat{\theta} = m^{-1}(\hat{\mathbf{m}})$. This should guarantee consistency: $\hat{\theta} \xrightarrow{\text{p}} \theta$.

Example 2.5

Suppose you have a sample of iid random variables for which you know that

$$\begin{aligned} \mathrm{E}[X] &= \frac{p}{\alpha} \\ \mathrm{E}\left[X^2\right] &= \frac{p(p+1)}{\alpha^2}; \end{aligned}$$

and define the two statistics $m_1 = \bar{X} = n^{-1}\sum_i x_i$ and $m_2 = n^{-1}\sum_i x_i^2$. Clearly

$$\begin{aligned} m_1 &\xrightarrow{\text{p}} \frac{p}{\alpha} \\ m_2 &\xrightarrow{\text{p}} \frac{p(p+1)}{\alpha^2}. \end{aligned}$$

Now consider the statistic $\hat{p} = \frac{m_1^2}{m_2 - m_1^2}$. Since $\hat{p}$ is a continuous function of both m_1 and m_2,

$$\hat{p} = \frac{m_1^2}{m_2 - m_1^2} \xrightarrow{\text{p}} \frac{p^2/\alpha^2}{p(p+1)/\alpha^2 - p^2/\alpha^2} = \frac{p^2}{p^2 + p - p^2} = p,$$

So $\hat{p}$ is a consistent estimator of p.

But then, by the same token, by dividing $\hat{p}$ by m_1 you get that

$$\frac{\hat{p}}{m_1} = \frac{m_1}{m_2 - m_1^2} \xrightarrow{\text{p}} \frac{p}{p/\alpha} = \alpha,$$

so you get a second statistic, $\hat{\alpha} = \frac{m_1}{m_2 - m_1^2}$ which estimates α consistently.

From the discussion above, the meaning of an estimator being consistent should be rather clear: a consistent estimator is a statistic that becomes arbitrarily precise if your dataset is large enough, because its distribution tends to collapse to a single point if n goes to infinity. Interestingly, there are two ways in which an estimator may *not* be consistent: one case arises when $\hat{\theta}$ has a probability limit, but is different from the desired point; in other words, $\hat{\theta} \xrightarrow{\text{p}} c \neq \theta$. However, it may be the case that $\hat{\theta}$ does not have a probability limit at all. In that case, lack of consistency is a consequence of the distribution of our statistic not collapsing to a single point, but rather remaining spread across a set of values no matter how large n is.

2.3.2 Asymptotic normality

Consistency is important because we want our estimators to be reasonably precise for large samples, but this is almost never enough, as we may need to make more precise statements on the distribution of our estimators.

For example, imagine that we have two consistent estimators for the same quantity: that is, two different statistics $\hat{\theta}$ and $\tilde{\theta}$ that have the same probability limit θ. How do we choose which one to use? Consistency can't be used as a criterion, since they are both consistent: if we define

$$\begin{aligned} \widehat{P}_n &= P\left[|\hat{\theta}-\theta|<\epsilon\right] \\ \widetilde{P}_n &= P\left[|\tilde{\theta}-\theta|<\epsilon\right] \end{aligned}$$

clearly $\lim_{n\to\infty}\widehat{P}_n = \lim_{n\to\infty}\widetilde{P}_n = 1$, so a decision can't be made on these grounds. Nevertheless, if we could establish that, for n large enough, $\widehat{P}_n > \widetilde{P}_n$, so that our probability of being grossly wrong is lower if we use $\hat{\theta}$ instead of $\tilde{\theta}$, our preferred course of action would be obvious. Unfortunately, this is not an easy check: $\widehat{P}_n$ is defined as

$$\widehat{P}_n = \int_{\theta-\epsilon}^{\theta+\epsilon} \hat{f}(x)\mathrm{d}x,$$

where $\hat{f}(x)$ is the density function for $\hat{\theta}$ (clearly, a parallel definition holds for $\widetilde{P}_n$). In most cases, the analytical form of $\hat{f}(x)$ is very hard to establish, if not at all impossible. However, we could try to approximate the actual densities with something good enough to perform the required check. This is almost invariably achieved by resorting to a property called **asymptotic normality**, by which the unknown density $\hat{f}(x)$ can be approximated via a suitably chosen Gaussian density.[18]

At first sight, this sounds like a very ambitious task: how can we hope to make general statements on the distribution of an arbitrary function of arbitrarily distributed random variables? Besides, why the Gaussian density, rather than something else? What's so special about the bell-shaped curve?

And yet, there is a result that applies in a surprisingly large number of cases, and goes under the name of **Central Limit Theorem**, or CLT for short. Basically, the CLT says that, under appropriate conditions, when you observe a random variable X that can be conceivably thought of as the accumulation of a large number of random causes that are reasonably independent of each other, with none of them dominating the others in magnitude, there are very good chances that the distribution of X should be approximately normal.

The practical effect of this theorem is ubiquitous in nature; most natural phenomena follow (at least approximately) a Gaussian distribution; the width of leaves, the length of fish, the height of humans. The French mathematician Henri Poincaré is credited with the following remark:

[18] I assume that the reader is reasonably comfortable with the Gaussian distribution, but section 2.A.5 is there, just in case.

> Everyone is sure of this, Mr. Lippman told me one day, since the experimentalists believe that it is a mathematical theorem, and the mathematicians that it is an experimentally determined fact.[19]


HENRI POINCARÉ

In order to illustrate the concept, we have to begin by defining convergence in distribution:

$$X_n \xrightarrow{\mathrm{d}} X \Longleftrightarrow F_{X_n}(z) \to F_X(z) \tag{2.11}$$

When X_n converges in distribution to X, the difference between $P(a < X_n \leq b)$ and $P(a < X \leq b)$ becomes negligible for large n. So, for large n, we can approximate quite accurately the probability of events defined for X_n via the corresponding event defined for X.[20]

Note the fundamental difference between convergence in probability and in distribution: if $X_n \xrightarrow{\mathrm{p}} X$, then for large n each time we observe a realisation of X_n and X we can be fairly confident that the two numbers will be very close. If $X_n \xrightarrow{\mathrm{d}} X$ instead, there is no guarantee that $|X_n - X|$ will be small: the only thing we know is that they come from (nearly) the same probability distribution, and therefore all we can say is that $P(a < X_n \leq b)$ should be very close to $P(a < X \leq b)$. Convergence in distribution is useful because probabilities involving X are often much easier to compute than probabilities involving X_n.

Convergence in distribution is a much weaker concept than convergence in probability: for example, take a sequence $X_1, X_2, \ldots X_n$ of iid random variable with the same distribution F. Of course, by the definition we can say that $X_n \xrightarrow{\mathrm{d}} X$, where the distribution of X is, again, F, but there is very little we can say about the behaviour of the sequence itself.

On the other hand, if $X_n \xrightarrow{\mathrm{p}} X$, the fact that $\lim_{n\to\infty} P[|X_n - X| < \epsilon$ implies that, when n is large, $P(a < X_n < b) \simeq P(a < X < b)$ for every interval (a,b), and therefore $X_n \xrightarrow{\mathrm{d}} X$. This result is often spelt "convergence in probability implies convergence in distribution, but not vice versa", or $X_n \xrightarrow{\mathrm{p}} X \Rightarrow X_n \xrightarrow{\mathrm{d}} X$

Now imagine that the LLN holds and $\bar{X} \xrightarrow{\mathrm{p}} m$. Clearly, $\bar{X} - m \xrightarrow{\mathrm{p}} 0$. In many cases, it can be proven that multiplying that quantity by $\sqrt{n}$ gives you something that doesn't collapse to 0 but does not diverge to infinity either. The Central Limit Theorems analyse the conditions under which

$$\sqrt{n}(\bar{X} - m) \xrightarrow{\mathrm{d}} \mathcal{N}(0, v)\,, \tag{2.12}$$

[19] French original: *Tout le monde y croit cependant, me disait un jour M. Lippmann, car les expérimentateurs s'imaginent que c'est un théorème de mathématiques, et les mathématiciens que c'est un fait expérimental.*

[20] If I had wanted to interrupt the flow of the argument for the sake of accuracy, I should have said at this point that in many cases we should take into account the fact that the support of X_n may be discrete, and special care is needed to interpret what happens when $F_{X_n}(z)$ "takes a step". I thought that would have been rather pedantic, so this remark is confined to a footnote.

so that we can use a Gaussian density to approximate the distribution of the average. A multivariate version also exists, which is slightly more intricate from a technical point of view, but the intuition carries over straightforwardly.[21]

The approximation provided by the CLT can also be stated by using the symbol $\overset{\text{a}}{\sim}$, which means "approximately distributed as" (where the approximation gets better and better as n grows):

$$\sqrt{n}(\mathbf{w}-\mathbf{m}) \xrightarrow{\text{d}} \mathcal{N}(0,\Sigma) \Longrightarrow \mathbf{w} \overset{\text{a}}{\sim} \mathcal{N}\left(\mathbf{m}, \frac{1}{n}\Sigma\right)$$

If a certain quantity $\mathbf{w}$ converges in distribution to a Normal rv with covariance Σ, then we call Σ the **asymptotic variance** of $\mathbf{w}$, by which we mean that the distribution of $\mathbf{w}$ resembles more and more one of a normal rv whose variance is Σ so we can take Σ as an approximation of the variance of $\mathbf{w}$.[22] This is usually notated as $\mathrm{AV}[\mathbf{w}] = \Sigma$.

In the same way as the LLN, there are many versions of the CLT, designed to cover different cases. A simple version, close in spirit to Khinchin's LLN, was provided by Lindeberg and Lévy: if $x_1, x_2, \ldots, x_n$ are iid, $\mathrm{E}[x_i] = m$, and $\mathrm{V}[x_i] = v$, then equation (2.12) holds. In practice, the conditions are the same as in Khinchin's LLN, with the additional requirement that the variance of x_i must exist.

Example 2.6

Let's go back to example 2.4 (the coin-tossing experiment). Here not only the mean exists, but also the variance:

$$\mathrm{V}[x_i] = \mathrm{E}\left[x_i^2\right] - \mathrm{E}[x_i]^2 = \pi - \pi^2 = \pi(1-\pi)$$

Therefore, the Lindeberg-Lévy version of the CLT is readily applicable, and we have

$$\sqrt{n}(\hat{p}-\pi) \xrightarrow{\text{d}} \mathcal{N}(0, \pi(1-\pi)),$$

so the corresponding asymptotic approximation is

$$\hat{p} \overset{\text{a}}{\sim} \mathcal{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

In practice, if you toss a fair coin ($\pi = 0.5$) $n = 100$ times, the distribution of the relative frequency you get is very well approximated by a Gaussian random variable with mean 0.5 and variance 0.0025. Just so you appreciate how well the approximation works, consider that the event $0.35 < \hat{p} \leq 0.45$ has a true probability of 18.234%, while the approximation the CLT gives you is 18.219%. If

[21] At this point, the inquisitive reader may ask: why the square root of n? Why not n itself, or the cube root, or some other function of n? Section 2.A.4 offers an intuitive explanation of why it should be so.

[22] Note that, from a technical point of view, $\mathbf{w}$ may not have a variance for any n, although its limit distribution does. But let's not be pedantic.

you're interested in reproducing these numbers, Section 2.A.6 contains a small gretl script with all the necessary steps. Of course, you're strongly encouraged to translate it to any other software you like better.

The CLT, by itself, describes the convergence in distribution of averages. However, we need to see what happens to our estimators, that are usually functions of those averages. There are two tools that come in especially handy. The first one is sometimes called **Cramér's theorem**: if $X_n \xrightarrow{\text{p}} a$ (where a is a constant) and $Y_n \xrightarrow{\text{d}} Y$, then

$$X_n \cdot Y_n \xrightarrow{\text{d}} a \cdot Y. \tag{2.13}$$

The second result we will often use is the **delta method**: if your estimator $\hat{\theta}$ is defined as a differentiable transformation of a quantity which obeys a LLN and a CLT, there is a relatively simple rule to obtain the limit in distribution of $\hat{\theta}$;

$$\left\{ \begin{array}{c} \bar{X} \xrightarrow{\text{p}} m \\ \sqrt{n}\left(\bar{X}-m\right) \xrightarrow{\text{d}} \mathcal{N}(0,\Sigma) \end{array} \right\} \Longrightarrow \left\{ \begin{array}{c} \hat{\theta} = g(\bar{X}) \xrightarrow{\text{p}} \theta = g(m) \\ \sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{\text{d}} \mathcal{N}\left(0, J\Sigma J'\right) \end{array} \right\} \tag{2.14}$$

where n is the sample size and J is the Jacobian $\left.\frac{\partial g(x)}{\partial x}\right|_{x=m}$.

Example 2.7

Given a sample of iid random variables x_i for which $\mathrm{E}[x_i] = 1/a$ and $\mathrm{V}[x_i] = 1/a^2$, it is straightforward to construct a consistent estimator of the parameter a as

$$\hat{a} = \frac{1}{\bar{X}} \xrightarrow{\text{p}} \frac{1}{1/a} = a.$$

Its asymptotic distribution is easy to find: start from the CLT:

$$\sqrt{n}\left(\bar{X} - 1/a\right) \xrightarrow{\text{d}} \mathcal{N}\left(0, 1/a^2\right).$$

All we need is the Jacobian term, which is

$$J = \operatorname{plim} \frac{\mathrm{d}\hat{a}}{\mathrm{d}\bar{X}} = -\operatorname{plim} \frac{1}{\bar{X}^2} = -\frac{1}{1/a^2} = -a^2;$$

therefore, the asymptotic variance of $\hat{a}$ is given by

$$\mathrm{AV}[\hat{a}] = (-a^2)\frac{1}{a^2}(-a^2) = a^2,$$

and therefore

$$\sqrt{n}\left(\hat{a} - a\right) \xrightarrow{\text{d}} \mathcal{N}\left(0, a^2\right)$$

so we can use the approximation $\hat{a} \overset{\text{a}}{\sim} \mathcal{N}\left(a, \frac{a^2}{n}\right)$.

By using these tools, we construct estimators satisfying not only the consistency property, but also asymptotic normality. These estimators are sometimes termed **CAN** estimators (Consistent and Asymptotically Normal). Asymptotic normality is important for three reasons:

1. it can be used to compare two consistent estimators in terms of their relative asymptotic efficiency. Given two consistent estimators a and b for the same parameter m, we'll say that a is asymptotically more efficient than b if $\text{AV}[b] \geq \text{AV}[a]$ (that is, $\text{AV}[b] - \text{AV}[a]$ is a positive semi-definite matrix — see page 38).[23]

2. It provides a fairly general way to construct statistics for testing hypotheses, which is probably the most useful thing an applied scientist might want to do with data. The next section is just about this.

3. asymptotic normality makes it quite easy to construct **confidence intervals**: in order to illustrate the concept, suppose we have a scalar estimator $\hat{\theta}$, whose asymptotic distribution is

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{\text{d}} \mathcal{N}(0, \omega);$$

this means that, for a decently large value of n, we can approximate the distribution of $\hat{\theta}$ as $\hat{\theta} \overset{\text{a}}{\sim} N\left(\theta, \frac{\omega}{n}\right)$. This, in turn, implies that

$$P\left[\frac{|\hat{\theta} - \theta|}{\sqrt{\omega/n}} < 1.96\right] \simeq 95\%;$$

therefore, the chances that the interval

$$\hat{\theta} \pm 1.96 \times \sqrt{\omega/n}$$

contains the true value of θ are roughly 95%. This is what is called a 95% confidence interval. Of course, a 99% confidence interval would be somewhat larger. Generalising this to a vector of parameters would lead us to speaking of **confidence sets**; for example, when θ is a 2-parameter vector, the confidence set would be an ellipse.

Example 2.8

Consider the same setup as example 2.7, that is tossing a coin $n = 100$ times, and suppose we get "heads" 45 times. Therefore, our estimate of π would be $\hat{p} = 45/100 = 0.45$.

[23] This criterion is easy to understand when a and b are scalars: $\text{AV}[b] \geq \text{AV}[a]$. The vector case is more subtle: if our object of interest is estimating some scalar function of m (say, $g(m)$), then the two natural competing estimators would be $g(a)$ and $g(b)$, respectively, that are both consistent by Slutsky's theorem. However, by applying the delta method, it can be proven that $\text{AV}\left[g(b)\right] \geq \text{AV}\left[g(a)\right]$ in all cases.

Of course, this would imply that the asymptotic variance of our estimator can be itself estimated as

$$\hat{v} = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.45 \cdot 0.55}{100} = 0.002475$$

so, since its square root is $\sqrt{\hat{v}} = 0.04975$*, we may say that the interval*

$$A = [0.45 - 0.04975 \times 1.96, 0.045 + 0.04975 \times 1.96] = 0.45 \pm 0.0975 \simeq [0.35, 0.55]$$

has a very good chance (95%) of containing the true value of π*.*

This example should help the reader steer clear of a common misconception: it is often said "the parameter has a 95% chance of being between a and b" as if the parameter was random and the interval was fixed. It's the other way around. The value of the parameter is non-random (and unknown), whereas the bounds of the interval are random, since they are a function of the estimator. Therefore, a better choice of words would be "the interval between a and b has a 95% chance of containing the parameter". ⌟

2.4 Hypothesis Testing

The starting point is a tentative conjecture (called the **null hypothesis**, or H_0) that we make about the parameters of a DGP.[24] As I said at the beginning of Section 2.3, we take it for granted that the DGP parameters belong to a certain set S (the parameter space), but we may conjecture that in fact there is a certain subset of S (say, H), that also contains θ. In formulae:

$$H_0 : \theta \in H \subset S.$$

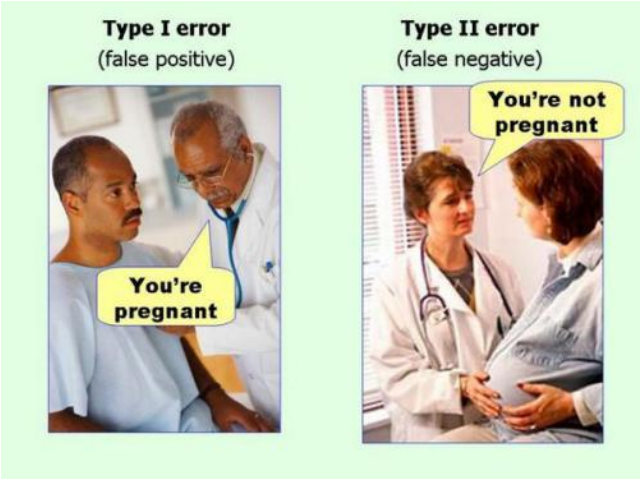
We would like to check whether our belief is consistent with the observed data. If what we see is at odds with our hypothesis, then reason dictates we should drop it in favour of something else.

The coin toss is a classic example. The one parameter in our DGP is π, that is a probability, so the parameter space is $S = [0,1]$. However, there is a subset of S that is of special significance, namely the point $H = \{0.5\}$, because if $\pi \in H$ (which trivially means $\pi = 0.5$), then the coin is fair.

We presume that the coin is fair, but it'd be nice if we could check. What we can do is flip it a number of times, and then decide, on the basis of the results, if our original conjecture is still tenable. After flipping the coin n times, we obtain a vector $\mathbf{x}$ of zeros and ones. What we want is a function $T(\mathbf{x})$ (called a **test statistic**) such that we can decide whether to reject H_0 or not.

[24] Disclaimer: this section is horribly simplistic. Any decent statistics textbook is far better than this. My aim here is just to lay down a few concepts that I will use in subsequent chapters, with no claim to rigour or completeness. My advice to the interested reader is to get hold of Casella and Berger (2002) or Gourieroux and Monfort (1995, volume 2). Personally, I adore Spanos' historical approach to the matter.

Figure 2.1: Types of Error



Note: in this case, H_0 is that the person is not *pregnant.*

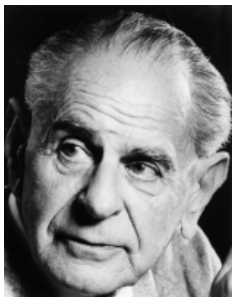
By fixing beforehand a subset R of the the support of $T(\mathbf{x})$ (called the "rejection region"), we can follow a simple rule: we reject H_0 if and only if $T(\mathbf{x}) \in R$. Since $T(\mathbf{x})$ is a random variable, the probability of rejecting H_0 will be between 0 and 1 regardless of the actual truth of H_0.[25] Therefore, there is a possibility that we'll end up rejecting H_0 while it's in fact true, but the opposite case, when we don't reject while in fact we should, is also possible. These two errors are known as **type I** and **type II** errors, respectively, and the following 4-way table appears in all statistics textbooks, but memorising the difference could be easier if you look at figure 2.1:

	Not Reject	Reject
H_0 true	OK	Type I error
H_0 false	Type II error	OK

This situation is not unlike the job of a judge in a criminal case. The judge starts from the premise that the defendant is not guilty, and then evidence is examined. By the "presumption of innocence" principle, the judge declares the defendant guilty only if the available evidence is overwhelmingly against H_0. Thus, type I error happens when an innocent person goes to jail; type II error is when a criminal gets acquitted.

This line of thought is very much in line with the idea philosophers call *falsificationism*, whose most notable exponent was the Austrian-British philosopher Karl Popper (see eg Popper (1968)). According to the falsificationist point of view, scientific progress happens via a series of rejections of previously made conjectures.

[25] Unless, of course, we do something silly such as deciding to always reject, or never. But in that case, what's the point of performing the experiment at all?



KARL POPPER

The hallmark of science (as opposed to other kinds of perfectly respectable mental endeavours, such as literary criticism, musical composition or political analysis) is that a conjecture can be *falsified*: we don't use evidence to *verify* a theory, but we can use it to prove that it's wrong.

Scientific theories always sound like: "when A, then B". Clearly, you can never verify a theory, because you can never rule out the possibility of observing A and not B; at most, you can say "the theory holds *so far*". But once you've observed one single case that disproves it, you need no more evidence to decide on the theory. When a theory is proved to be inconsistent with the available evidence, we move on to something better, and progress is made; but until a conjecture is not rejected, we adopt it as a tentative (and possibly wrong) explanation.

I am fully aware that the debate on philosophy of science has long established that falsificationism is untenable, as a description of scientific progress, on several accounts. What I'm saying is just that the statistical theory of hypothesis testing borrows quite a few ideas from the falsificationist approach. For a fuller account, check out Andersen and Hepburn (2016).

Therefore, strictly speaking, "not rejecting" doesn't mean "accepting". Rejection is always final; failure to reject is always provisional. That said, it is quite common (although incorrect) to use the word "accept" instead of "not reject", and I will do the same here. However, the reader should bear in mind is that "accepting" really means "accepting *for now*".[26]

The recipe we are going to use for constructing test statistics is simple: first, we will formulate our hypothesis of interest as $H_0 : g(\theta) = 0$, where θ are the DGP parameters and $g(\cdot)$ is a differentiable function. Then, given a CAN estimator $\hat{\theta}$, we evaluate the function at that point. Given consistency, we would expect $g(\hat{\theta})$ to be "small" if H_0 is true, since under H_0, $g(\hat{\theta}) \xrightarrow{\text{p}} 0$.

In order to build a rejection region, we need some criterion for deciding when $g(\hat{\theta})$ is large enough to force us to abandon H_0; we do so by exploiting asymptotic normality. By using the delta method, we can find an asymptotic approximation to the distribution of $g(\hat{\theta})$ as

$$g(\hat{\theta}) \overset{\text{a}}{\sim} \mathcal{N}\big(g(\theta), \Sigma\big);$$

under consistency, Σ should tend to a zero matrix; as for $g(\theta)$, that should be 0 if and only if H_0 is true. These two statements imply that the quadratic form $g(\hat{\theta})'\Sigma^{-1}g(\hat{\theta})$ should behave very differently in the two cases: if H_0 is false, it

[26] Some people use the word "retain" instead of "accept", which is certainly more correct, but unfortunately not very common.

should diverge to infinity, since plim$\left[g(\hat{\theta})\right] \neq 0$; if H_0 is true, instead, approximate normality implies that[27].

$$W = g(\hat{\theta})'\Sigma^{-1}g(\hat{\theta}) \overset{a}{\sim} \chi^2_p$$

where $p = \text{rk}(\Sigma)$. Hence, under H_0, the W statistic should take values typical of a χ^2 random variable.[28] Therefore, we should expect to see "small" values of W when H_0 is true and large values when it's false. The natural course of action is, therefore, to set the rejection region as $R = (c, \infty)$, where c, the **critical value**, is some number to be determined. Granted, there is always the possibility that $W > c$ even if H_0 is true. In that case, our decision to reject would imply a type I error. But since we can calculate the distribution function for W, we can set c to a prudentially large value. What is normally done is to set c such that the probability of a type I error (called the **size** of the test, and usually denoted by the Greek letter α) is a small number, typically 5%.

What people do in most cases is deciding which α they want to use and then set c accordingly, so that in many cases you see c expressed as a function if α (and written c_α), rather than the other way around.

But, I hear you say, what about type II error? Well, if W in fact diverges when H_0 is false, the the probability of rejection (also known as the **power** of the test) should approach 1, and we should be OK, at least when our dataset is reasonably large.[29] There are many interesting things that could and should be said about the power of tests, especially a truly marvellous result known as the *Neyman-Pearson lemma*, but I'm afraid this is not the place for this. See the literature cited at footnote 24.

Example 2.9

Let's continue example 2.6 here. So we have that the relative frequency is a CAN estimator for the probability of a coin showing "heads".

$$\sqrt{n}(\hat{p} - \pi) \xrightarrow{d} \mathcal{N}(0, \pi(1-\pi)).$$

Let's use this result for building a test for the "fair coin" hypothesis, $H_0 : \pi = 0.5$. We need a differentiable function $g(x)$ such that $g(x) = 0$ if and only if $x = 0.5$. One possible choice is

$$g(\pi) = 2\pi - 1$$

What we have to find is the asymptotic variance of $g(\hat{p})$, which is AV$\left[g(\hat{p})\right] = J \cdot \pi(1-\pi) \cdot J' = \omega$, *where* $J = \text{plim}\left(\frac{\partial g(x)}{\partial x}\right) = 2$, *so*

$$\sqrt{n}(g(\hat{p}) - g(\pi)) \xrightarrow{d} \mathcal{N}(0, \omega).$$

[27] To see why, see section 2.A.5

[28] The reason why I'm using the letter W to indicate the test is that, in a less cursory treatment of the matter, the test statistic constructed in this way could be classified as a "Wald-type" test.

[29] When the power of a test goes to 1 asymptotically, the test is said to be **consistent**. I know, it's confusing.

Under the null, $g(\pi) = 0$ and $\omega = 1$; therefore, the approximate distribution for $g(\hat{p})$ is

$$g(\hat{p}) \overset{a}{\sim} \mathcal{N}\left(0, n^{-1}\right)$$

and our test statistic is easy to build as

$$W = g(\hat{p}) \left[\frac{1}{n}\right]^{-1} g(\hat{p}) = n \cdot (2\hat{p} - 1)^2$$

A simple numerical example: suppose $n = 100$ and $\hat{p} = 46\%$. The value of W equals $W = 100 \cdot 0.08^2 = 0.64$. Is the number 0.64 incompatible with the presumption that the coin is fair? Not at all: if the coin is in fact fair W should come from a random variable that in 95% of the cases takes values below 3.84. Therefore, there is no reason to change our mind about H_0. The reader may want to check what happens if the sample size of our experiment is set to $n = 1000$. ⌟

The reader may be a bit perplexed about my vagueness about the nature of the $g(\theta)$ function that we use to build the test. In order for the above to work, besides the obvious requisite $g(\theta) = 0$ under the null, we only need this function to be continuous and differentiable; therefore, there is a considerable degree of arbitrariness in the way the function can be chosen. In the example above, for the hypothesis $H_0 : p = 0.5$ I employed $g(p) = 2p - 1$, but I could have chosen several alternatives, such as

$$g(p) = 1/2 - p \quad \text{or} \quad g(p) = \log(2p) \quad \text{or} \quad g(p) = 3 - 9^p.$$

Are they all equivalent? To cut a long story short, asymptotically, yes. In finite samples, no, but there is no rule to tell which choice is the "best". As a consequence, one normally goes for the one that implies the least computational effort. In example 2.9 I used $g(p) = 2p - 1$ simply because its Jacobian is $\frac{\partial g(p)}{\partial p} = 2$ and everything becomes nice and simple.

2.4.1 The p-value

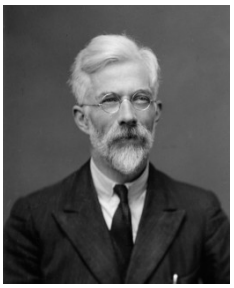
The way to make decisions on H_0 that I illustrated in the previous section is perfectly legitimate, but what people do in practice is slightly different, and involves a quantity known as the p-value.

The traditional method of checking the null hypothesis is based on the construction of a test statistic with a known distribution under H_0; once the size of the test (most often, 5%) is decided, the corresponding critical value c is found and then the realised statistic W is compared to c. If $W > c$, reject; otherwise, don't. This method makes perfect sense if finding the critical value for a given size α is complicated: you compile a table of critical values for a given α once and for all, and then every time you perform a test you just check your realised

value of W against the number in the table. All 20th century statistics textbooks contained an appendix with tables of critical values for a variety of distributions.

With the advent of cheap computing, it has become easy to compute c as a function of α, as well as performing the inverse calculation, since algorithms for computing the cdf of a χ^2 random variable are fast, precise and efficient. Thus, an equivalent route is often followed: after computing W, you calculate the probability that a χ^2 variable should take values greater than W.

That number is called the p-value for the test statistic W. Clearly, if $W > c$, the p-value must be smaller than α, so the decision rule can be more readily stated as "reject H_0 if the p-value is smaller than α".

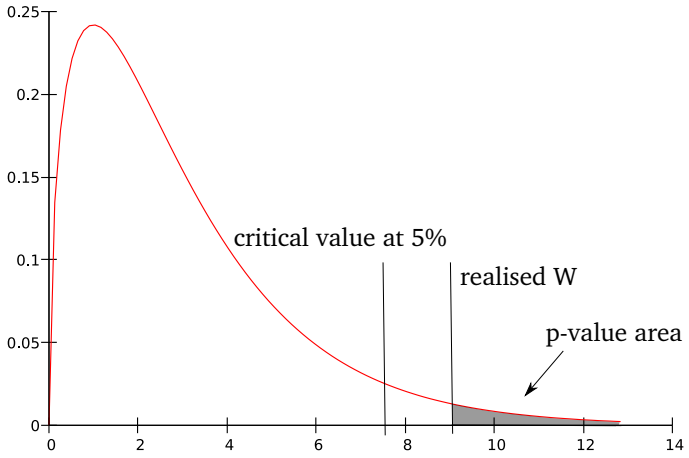


RONALD FISHER

In fact, according to its inventor, Sir Ronald Aylmer Fisher (arguably, the greatest statistician of all time), the p-value can be seen as a continuous (or monotonous) summary statistic of how well the data are compatible with the hypothesis;[30] in Fisher's own words, when we see a small p-value, "*[e]ither* an exceptionally rare chance has occurred *or* the theory [...] is not true".[31]

Figure 2.2 shows an example in which $W = 9$, and is compared against a χ^2_3 distribution. The corresponding 95% percentile is 7.815, so with $\alpha = 0.05$ the null should be rejected. Alternatively, we could compute the area to the right of the number 9 (shaded in the Figure), which is 2.92%; obviously, 2.92% < 5%, so we reject.

Figure 2.2: p-value example



To make results even quicker to read, most statistical packages adopt a graph-

[30] Thanks to Sven Schreiber for putting it so clearly and concisely.

[31] Fisher RA. *Statistical Methods and Scientific Inference.* Ed 2, 1959. On this subject, if you're into the history of statistics, you might like Biau et al. (2009).

ical convention, based on ornating the test statistic with a variable number of '$*$' characters, usually called "stars". Their meaning is as follows:

Stars	Meaning
(none)	p-value greater than 10%
*	p-value between 5% and 10%
**	p-value between 1% and 5%
***	p-value under 1%

Therefore, when you see 3 stars you "strongly" reject the null, but you don't reject H_o where no stars are printed. One star means "use your common sense".

In fact, I'd like add a few words on "using your common sense": relying on the p-value for making decisions is OK; after all, that's what it was invented for. However, you should avoid blindly following the rule "above 5% → yes, below 5% → no". You should always be aware of the many limitations of this kind of approach: for example,

- even if all the statistical assumption of your model are met, the χ^2 distribution is just an approximation to the actual density of the test. Therefore, the quantiles of the χ^2 density may be (slightly) misleading, especially so when your sample is not very large;
- even if the test was in fact exactly distributed as a χ^2 variable, type I and type II errors are always possible; actually, if you choose 5% as your significance level (like everybody does), you *will* make a mistake in rejecting H_0 one time out of twenty;
- and besides, why 5%? Why not 6%? Or 1%? In fact, someone once said:

  Q: *Why do so many colleges and grad schools teach p = 0.05?*

  A: *Because that's still what the scientific community and journal editors use.*

  Q: *Why do so many people still use p = 0.05?*

  A: *Because that's what they were taught in college or grad school.*

  (for more details, see Wasserstein and Lazar (2016)).

That said, I don't want you to think "OK, the p-value is rubbish": it isn't. Actually, it's the best tool we have for the purpose. But like any other tool (be it a screwdriver, a microwave oven or a nuclear reactor), in order to use it effectively, you must be aware of its shortcomings.

2.5 Identification

A common problem in econometrics is **identification** of a model. The issue is quite complex, and I cannot do justice to it in an introductory book such as this,

so I'll just sketch the main ideas with no pretence to rigour or completeness. Basically, a model is said to be identified with reference to a question of interest if the model's probabilistic structure is informative on that question.

A statistical model is, essentially, a probabilistic description of the data that we observe. When we perform inference on a dataset we assume that

- our available dataset is a realisation of some probabilistic mechanism (the DGP — see section 2.1);
- the salient features of the DGP can be described by a parameter vector θ;
- the data we observe are such that asymptotic theory is applicable (for example, n is large and the data are iid) and we can define statistics that we can use as estimators or tests;
- our question of interest can be phrased as a statement on the vector θ.

The importance of the first three items in the list should be clear to the reader from the past sections of this chapter. In this section, we will discuss the fourth one.

The vector θ contains parameters that describe the probability distribution of our data, but in principle the empirical problem we are ultimately interested in is expressed as a vector of **parameters of interest** ψ. That is, the parameters θ characterise the DGP, while ψ is a formalised description of the aspect of reality we are trying to analyse. For example, in a typical econometric model, ψ may contain quantities such as the elasticity of demand for a certain good to its own price, the causal effect of a policy on a target variable, the risk aversion parameter for the representative individual in a macroeconomic model, and so on.

What is the relationship between ψ and θ? If we take ψ as being a stylised description of reality, and θ as a stylised description of what we observe, then θ should be a known function of ψ, that we assume known:

$$\theta = M(\psi).$$

In some cases, the relationship is trivial; often, $M(\cdot)$ is just the identity function, $\theta = \psi$, but sometimes this is not the case.

Statistics gives us the tools to estimate θ; is this enough to estimate ψ? It depends on the function $M(\cdot)$; if the function is invertible, and we have a CAN estimator $\hat{\theta}$, a possible estimator for ψ is

$$\hat{\psi} = M^{-1}(\hat{\theta}).$$

If $M(\cdot)$ is continuous and differentiable, then its inverse will share these properties, so we can use Slutsky's theorem and the delta method and $\hat{\psi}$ is a CAN estimator too. In this case, we say that the model is **identified**.

In some cases, however, the function $M(\cdot)$ is not invertible, typically when different values of ψ give rise to the same θ.[32] In other terms, two alternative descriptions of the world give rise to the same observable consequences: if

$$M(\psi_1) = M(\psi_2)$$

for $\psi_1 \neq \psi_2$, we would observe data from the same DGP (described by θ) in both cases; this situation is known as **observational equivalence**, and ψ_1 and ψ_2 are said to be **observationally equivalent**. In these cases, being able to estimate θ, even in an arbitrarily precise way, doesn't tell us if the "true" description of the world is ψ_1 or ψ_2. This unfortunate case is known in econometrics as **under-identification**.

Example 2.10

Suppose you have an urn full of balls, some white and some red. Call w the number of white balls and r the number of red balls. We want to estimate both w and r.

Suppose also that the only experiment we can perform works as follows: we can extract one ball from the urn as many times as we want, but we must put it back after extraction (statisticians call this "sampling with replacement"). Define the random variable x_i as 1 if the ball is red. Clearly

$$x_i = \begin{cases} 1 & \text{with probability} \quad \pi = \frac{r}{w+r} \\ 0 & \text{with probability} \quad (1-\pi_i) = \frac{w}{w+r} \end{cases}$$

In this case, the probability distribution of our data is completely characterised by the parameter π; as we know, we have a perfectly good way to estimate π; since the data are iid, $\bar{X}$ is a CAN estimator of π and testing hypotheses on π is easy.

If, however, the parameters of interest are $\psi = [r, w]$, there is no way to estimate them separately, because the function $\theta = M(\psi)$ is not invertible, for the very simple reason that the relationship between the DGP parameter π and our parameters of interest w and r

$$\pi = \frac{r}{w+r}$$

is one equation in two unknowns. Therefore, in the absence of extra information we are able to estimate π (the proportion *of red balls) as precisely as wanted, but there is no way to estimate r (the* number *of red balls).*

Even if we knew the true *value of π, there would still be an infinite array of observationally equivalent descriptions of the urn. If, say, $\pi = 0.3$ the alternatives $\psi_1 = [3, 10]$, $\psi_2 = [15, 50]$, $\psi_3 = [3000, 10000]$, etc would all be observationally equivalent.*

[32] The technical way to say this would be "the $M(\cdot)$ function is not injective".

Identification of a model can be a very serious concern in some settings: if a model is under-identified, we may be able to estimate consistently the parameters that describe the data, but this wouldn't be helpful for the economic question we are ultimately after. In this book, we will not encounter any of these cases, except for the models I will describe in chapter 6, but you should be aware of the potential importance of the problem.

2.A Assorted results

2.A.1 Jensen's lemma

As we argued in Section 2.2.3, if $g(\cdot)$ is not a linear function, generally $\mathrm{E}\left[g(X)\right] \neq g\left[\mathrm{E}\left[X\right]\right]$. However, when $g(\cdot)$ is concave, we have a usable result that comes as an inequality:

$$\mathrm{E}\left[g(X)\right] \leq g\left(\mathrm{E}\left[X\right]\right)$$

For example, if $\mathrm{E}\left[X\right] = 1$, we can be sure that $\mathrm{E}\left[\log(X)\right]$ is negative, just because the logarithm is a concave function (provided, of course, that $\mathrm{E}\left[\log(X)\right]$ exists), since $\mathrm{E}\left[\log(X)\right] \leq \log(1) = 0$.

This remarkable result is easy to prove if $g(\cdot)$ is also differentiable, since in this case $g(\cdot)$ is said to be concave between a and b if

$$g(x) \leq g(x^*) + g'(x^*)(x - x^*) \tag{2.15}$$

for each $x^* \in (a,b)$. Now assume that the interval (a,b) is the support of the rv X, which possesses an expectation. Clearly, $a < \mu = \mathrm{E}\left[X\right] < b$; this implies that equation (2.15) holds when $x^* = \mu$, and therefore

$$\mathrm{E}\left[g(X)\right] \leq \mathrm{E}\left[g(\mu) + g'(\mu)(X - \mu)\right] = g(\mu) + g'(\mu) \cdot \mathrm{E}\left[(X - \mu)\right]$$

Since obviously $\mathrm{E}\left[X - \mu\right] = 0$, it follows that

$$\mathrm{E}\left[g(X)\right] \leq g(\mu) = g\left[\mathrm{E}\left[X\right]\right],$$

as required. Note that

- by linearity, $\mathrm{E}\left[-X\right] = -\mathrm{E}\left[X\right]$, so if the function is convex instead of concave, you can flip the inequality, because the negative of a concave function is convex: $\mathrm{E}\left[g(X)\right] \geq g\left(\mathrm{E}\left[X\right]\right)$;

- it is possible to prove Jensen's lemma in the more general case when $g(\cdot)$ is not everywhere differentiable in (a,b), but that's a bit more intricate (see for example Williams (1991), page 61).

2.A.2 Markov's and Chebyshev's inequalities

These two inequalities are nice because they provide a link between the moments of a random variable and its distribution. Apparently, one may think that pieces of information like $\mathrm{E}[X] = 3$ or $\mathrm{V}[X] = 1$ say nothing on the distribution of X, but in fact they do, up to a point.

Let's begin by **Markov's inequality**. It states that if W is a random variable with *positive* support and expectation $\mathrm{E}[W] = m$, then

$$P[W \geq a] \leq \frac{m}{a}. \tag{2.16}$$

for any a. The proof is surprisingly easy:[33]

$$\begin{aligned} m &= \int_0^\infty w f(w)\mathrm{d}w = \int_0^a w f(w)\mathrm{d}w + \int_a^\infty w f(w)\mathrm{d}w \geq \int_a^\infty w f(w)\mathrm{d}w \geq \\ &\geq \int_a^\infty a f(w)\mathrm{d}w = a\int_a^\infty f(w)\mathrm{d}w = a \cdot P[W \geq a]. \end{aligned}$$

So for example if you knew that the expectation of a non-negative random variable W was 4, you could safely say that $P[W > 8] \leq 1/2$ without knowing anything on the distribution of X. Cool.

Now take a random variable X, with arbitrary support and $\mathrm{E}[X] = m$. Provided the second moment exists, define $S = (X - m)^2$, so $\mathrm{E}[S]$ is the variance of X. Clearly, S cannot be negative (it's a square), so Markov's inequality (2.16) applies directly and

$$P[S \geq a] \leq \frac{\mathrm{V}[X]}{a}.$$

The left-hand side of this inequality can be rewritten as $P[|X - m| \geq \sqrt{a}]$, so

$$P[|X - m| \geq \sqrt{a}] \leq \frac{\mathrm{V}[X]}{a},$$

and obviously

$$P\left[|X - m| \leq \sqrt{a}\right] \geq 1 - \frac{\mathrm{V}[X]}{a}; \tag{2.17}$$

this special case of Markov's inequality is known as **Chebyshev's inequality**. To give you an idea of how remarkable this result is, imagine that all we knew about X is that

$$\mathrm{E}[X] = 10 \qquad \text{and} \qquad \mathrm{V}[X] = 1.$$

Applying the formula above with $a = 4$ yields

$$P\left[|X - 10| \leq 2\right] = P\left[8 < X < 12\right] \geq 1 - \frac{1}{4} = 3/4, \tag{2.18}$$

so, at least 75% of the distribution of X is between 8 and 12. Considering how little we know about the distribution of X (is it discrete? is it symmetric? does it have one maximum? two? none at all?), I consider this a rather impressive result.

[33] Here I'm using integrals as if W was necessarily continuous, but the theorem in fact holds for any kind of random variable.

2.A.3 More on consistency

We stated in section 2.3.1 that if $x_1, x_2, \ldots, x_n$ are independent and identically distributed (iid for short) and $\mathrm{E}[x_i] = m$, then $\bar{X} \xrightarrow{\mathrm{p}} m$. However, this is only a sufficient condition, and is by no means necessary.

In this subsection, I will provide an example of an alternative scenario under which $\bar{X} \xrightarrow{\mathrm{p}} m$ even though the x_i variables may not be iid: let's just say that $\mathrm{E}[x_i] = m$, although nothing is said about heterogeneity or independence. In this example, however, it is crucial that they all possess a variance $v_i = \mathrm{V}[x_i]$.

Suppose we have a vector $\mathbf{x}$ of size n containing our observations, that are not necessarily independent nor identical. However, we do require that they possess second moments and use Σ to indicate the covariance matrix of $\mathbf{x}$:

$$\mathrm{V}[\mathbf{x}] = \Sigma = \begin{bmatrix} \mathrm{V}[x_1] & \mathrm{Cov}[x_1, x_2] & \ldots & \mathrm{Cov}[x_1, x_n] \\ \mathrm{Cov}[x_1, x_2] & \mathrm{V}[x_2] & \ldots & \mathrm{Cov}[x_2, x_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_1, x_n] & \mathrm{Cov}[x_2, x_x] & \ldots & \mathrm{V}[x_n] \end{bmatrix}$$

First, let's have a look at the moments of $\bar{X}$; its first moment is trivial to find, since

$$\mathrm{E}\left[\bar{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum x_i\right] = \frac{1}{n}\sum \mathrm{E}[x_i] = \frac{nm}{n} = m.$$

Now note that the average $\bar{X}$ can be written as $\bar{X} = \frac{1}{n}\boldsymbol{\iota}'\mathbf{x}$, and therefore its variance can be easily calculated by the rule (2.7). Therefore,

$$\mathrm{V}\left[\bar{X}\right] = \frac{1}{n^2} \cdot \boldsymbol{\iota}'\Sigma\boldsymbol{\iota}$$

What can we say about $\boldsymbol{\iota}'\Sigma\boldsymbol{\iota}$? First, given the properties of $\boldsymbol{\iota}$, this is simply the sum of all the elements of Σ; second, since Σ is positive semi-definite by construction, this cannot be a negative number, but it may be a large positive one. Especially so, considering that the size of Σ grows with n.

We must now examine what happens to $\boldsymbol{\iota}'\Sigma\boldsymbol{\iota}$ as $n \to \infty$. When the x_i rvs are iid, this is easy, since in this special case Σ is just a multiple of the identity matrix; hence, in the iid case, $\Sigma = v \cdot \mathbf{I}$ and $\boldsymbol{\iota}'\Sigma\boldsymbol{\iota} = n \cdot v$. In a more general case, the non-diagonal elements may be non-zero (which could happen for dependent observations), or the elements on the diagonal may be heterogeneous (which could happen in the non-identical case). However, it may still be that, despite these complications, $\boldsymbol{\iota}'\Sigma\boldsymbol{\iota}$ behaves asymptotically as a linear function of n. To be more precise, it may happen that

$$\lim_{n\to\infty} \frac{\boldsymbol{\iota}'\Sigma\boldsymbol{\iota}}{n} = K,$$

where K is some constant. For example, in the iid case, K would just be equal to v. In all these cases, you have that, for large n,

$$\mathrm{V}\left[\bar{X}\right] \simeq \frac{K}{n}.$$

The most immediate consequence of the equation above is that $\mathbb{V}\left[\bar{X}\right]$ tends to 0 for large n, and therefore the desired result is a simple consequence of Chebyshev's inequality (2.18) applied to $\bar{X}$, where ε is any positive real:

$$\lim_{n\to\infty} P[|\bar{X} - m| < \varepsilon] \geq 1 - \frac{K}{n\varepsilon^2} \to 1 \iff \bar{X} \xrightarrow{\text{p}} m.$$

Note that this case is nearly useless in more elaborate (and realistic) cases than the average, because being able to compute the moments of our quantities of interest is extremely rare, but still gives you a nice idea of the kind of conditions can be used to prove consistency.

2.A.4 Why $\sqrt{n}$?

Here I'll give you an intuitive account of the reason why, in the standard cases, the Central Limit Theorem works by using $\sqrt{n}$ as the normalising transformation instead of some other power of n.

Let's use the same scenario described in Section 2.A.3, that is a vector of observations $\mathbf{x}$ of size n, with common mean $\mathbb{E}[x_i] = m$ and covariance matrix

$$\mathbb{V}[\mathbf{x}] = \Sigma = \begin{bmatrix} \mathbb{V}[x_1] & \mathbb{C}\text{ov}[x_1, x_2] & \dots & \mathbb{C}\text{ov}[x_1, x_n] \\ \mathbb{C}\text{ov}[x_1, x_2] & \mathbb{V}[x_2] & \dots & \mathbb{C}\text{ov}[x_2, x_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}[x_1, x_n] & \mathbb{C}\text{ov}[x_2, x_x] & \dots & \mathbb{V}[x_n] \end{bmatrix}.$$

As I proved in section 2.A.3, we can approximate $\mathbb{V}\left[\bar{X}\right]$ as

$$\mathbb{V}\left[\bar{X}\right] \simeq \frac{K}{n},$$

where K is some positive real number. Thus,

$$\mathbb{V}\left[n^{\alpha}\left(\bar{X} - m\right)\right] \simeq K n^{2\alpha - 1}.$$

Therefore, the only way to multiply $\left(\bar{X} - m\right)$ by a power of n and have that the variance of the result is a constant is to choose that $\alpha = 1/2$, which of course gives you $\sqrt{n}$.

When observations are not iid, there may be cases when $\boldsymbol{\iota}'\Sigma\boldsymbol{\iota}$ grows at a rate that is different from n. In these cases, the normalising factor needed to achieve convergence in distribution is actually different from the square root. This typically happens when the x_i rvs come from a time-series sample, and the degree of dependence between nearby observations can be substantial. The beginning of chapter 5 contains a brief discussion of "persistence" in time series.

2.A.5 The normal and χ^2 distributions

A continuous random variable X is a **standard normal** random variable when its support is $\mathbb{R}$, and its density function is

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

as depicted in figure 2.3.

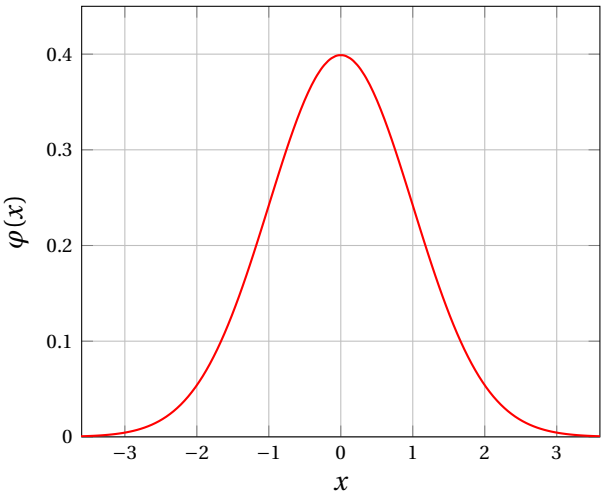


Figure 2.3: Standard normal density function

As is well known, $\varphi(x)$ has no closed-form indefinite integral: that is, it can be proven that the function $\Phi(x)$, whose derivative is $\varphi(x)$, does exist, but cannot be written as a combination of "simple" functions (the proof is very technical). Nevertheless, it's quite easy to approximate numerically, so every statistical program (heck, even spreadsheets) will give you excellent approximations via clever numerical methods. If you're into this kind of stuff, Marsaglia (2004) is highly recommended.

As the reader certainly knows, this object was discovered[34] by C. F. Gauss (the guy on page 13), so it's also also known as a **Gaussian** random variable. By playing with integrals a little[35], it can be proven that $\mathrm{E}[X] = 0$ and $\mathrm{V}[X] = 1$. One of the many nice properties of Gaussian rvs is that an affine transformation of a normal rv is also normal. Therefore, by the rules for expected values (see section 2.2.3), if X is a standard normal rv, then $Y = m + s \cdot X$ is a normal rv with mean m and variance s^2. Its density function is

$$f(y) = \frac{1}{\sqrt{2\pi s^2}} \exp\left\{-\frac{(y-m)^2}{2s^2}\right\}$$

[34] Or invented? Interesting point.

[35] If you want to have some fun with the moments of the standard normal distribution, you'll find the result $\frac{\mathrm{d}\varphi(x)}{\mathrm{d}x} = -x\varphi(x)$ very useful, because it implies that $\int x\varphi(x)\mathrm{d}x = -\varphi(x)$.

A compact way to say this is $Y \sim \mathcal{N}(m, s^2)$.

In fact, one can define a **multivariate normal** random variable as a random vector $\mathbf{x}$ with density

$$f(\mathbf{x}) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left\{(\mathbf{x}-\mathbf{m})'\Sigma^{-1}(\mathbf{x}-\mathbf{m})\right\},$$

or, in short, $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \Sigma)$, where n is the dimension of $\mathbf{x}$, $\mathbf{m}$ is its expectation and Σ its covariance matrix. The multivariate version of this random variable also enjoys the linearity property, so if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \Sigma)$, then

$$\mathbf{y} = A\mathbf{x} + \mathbf{b} \sim \mathcal{N}\left(A\mathbf{m} + \mathbf{b}, A\Sigma A'\right). \tag{2.19}$$

It is easy to overlook how amazing this result is: the fact that $\mathrm{E}[A\mathbf{x}+\mathbf{b}] = A\mathrm{E}[\mathbf{x}] + \mathbf{b}$ is true for any distribution and does not depend on Gaussianity; and the same holds for the parallel property of the variance. The special thing about the Gaussian distribution is that a linear transformation of a Gaussian rv is itself Gaussian. And this is a *very* special property, that is only shared by a few distributions (for example: if you take a linear combination of two Bernoulli rvs, the result is not Bernoulli-distributed).

The Gaussian distribution has a very convenient feature: contrary to what happens in general, if X and Y have a joint normal distribution (that is, the vector $\mathbf{x} = [Y, X]$ is a bivariate normal rv), absence of correlation implies independence (again, this can be proven quite easily: nice exercise left to the reader). Together with the linearity property, this also implies another very important result: if $\mathbf{y}$ and $\mathbf{x}$ are jointly Gaussian, then the conditional density $f(\mathbf{y}|\mathbf{x})$ is Gaussian as well. In formulae:

$$f(\mathbf{y}|\mathbf{x}) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left\{(\mathbf{y}-\mathbf{m})'\Sigma^{-1}(\mathbf{x}-\mathbf{m})\right\},$$

where

$$\begin{aligned}
\mathbf{m} &= \mathrm{E}\left[\mathbf{y}|\mathbf{x}\right] = \mathrm{E}\left[\mathbf{y}\right] + B'(\mathbf{x} - \mathrm{E}[\mathbf{x}]) \\
B &= \Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{x},\mathbf{y}} \\
\Sigma &= \mathrm{V}\left[\mathbf{y}|\mathbf{x}\right] = \Sigma_{\mathbf{y}} - \Sigma_{\mathbf{x},\mathbf{y}}'\Sigma_{\mathbf{x}}^{-1}\Sigma_{\mathbf{x},\mathbf{y}}
\end{aligned}$$

where $\Sigma_{\mathbf{y}}$ is the covariance matrix of $\mathbf{y}$, $\Sigma_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$ and $\Sigma_{\mathbf{x},\mathbf{y}}$ is the matrix of covariances between $\mathbf{x}$ and $\mathbf{y}$.

Example 2.11

For example, suppose that the joint distribution of y and $\mathbf{x} = [x_1, x_2]$ is normal, with

$$\mathrm{E}\begin{bmatrix} y \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\mathrm{V}\begin{bmatrix} y \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

then you have

$$\mathrm{E}\left[y\right]=1 \quad \mathrm{E}\left[\mathbf{x}\right]=[2,3]' \quad \Sigma_{\mathbf{y}}=3 \quad \Sigma_{\mathbf{x}}=\begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix} \quad \Sigma_{\mathbf{x},\mathbf{y}}=[0,1]'$$

and therefore

$$\begin{aligned} B &= \begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix}^{-1}\begin{bmatrix}0\\ 1\end{bmatrix}=\begin{bmatrix}-1\\ 1\end{bmatrix} \\ \Sigma &= 3-\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix}^{-1}\begin{bmatrix}0\\ 1\end{bmatrix}=3-1=2, \end{aligned}$$

since $\begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix}^{-1}=\begin{bmatrix}2 & -1\\ -1 & 1\end{bmatrix}$. *Thus, the conditional expectation of* y *given* $\mathbf{x}$ *equals*

$$\mathrm{E}\left[\mathbf{y}|\mathbf{x}\right]=1+[-1,1]'\begin{bmatrix}x_1-2\\ x_3-3\end{bmatrix}=-x_1+x_2$$

and in conclusion

$$y|\mathbf{x}\sim N\left[x_2-x_1,2\right].$$

Note that:

- *the conditional mean is a linear function of the* $\mathbf{x}$*; this needn't happen in general: it's a miraculous property of Gaussian random variables;*
- *the conditional variance is* not *a function of the* $\mathbf{x}$ *variables (it's a constant); again, this doesn't happen in general, but with Gaussian random variables, it does;*
- *if you apply the Law of Iterated Expectations (eq. (2.8)) you get*

$$\mathrm{E}\left[\mathrm{E}\left[y|\mathbf{x}\right]\right]=-\mathrm{E}\left[x_1\right]+\mathrm{E}\left[x_2\right]=-2+3=1=\mathrm{E}\left[y\right];$$

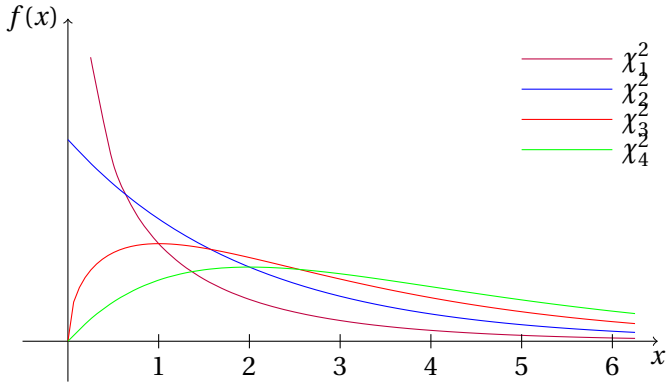
which is, in fact, unsurprising, but it's nice and reassuring. ⌟

If one instead needs to investigate the distribution of quadratic forms of Gaussian rvs, then another distribution arises, namely the **chi-square** distribution (χ^2 in symbols). The general result is that, if $\mathbf{x}\sim\mathcal{N}(0,\Sigma)$, then $\mathbf{x}'\Sigma^{-1}\mathbf{x}\sim\chi^2_n$, where n is the number of elements of $\mathbf{x}$, commonly known as the "degrees of freedom" parameter.

The support of the χ^2 density is over the non-negative reals; its shape depends on n (the degrees of freedom) in the following way:

$$f(x)=\frac{0.5^{n/2}}{\Gamma(n/2)}x^{n/2-1}e^{-x/2}.$$

Figure 2.4: Density function of χ^2_p, for $p = 1 \ldots 4$



The most common cases, where n ranges from 1 to 4, are shown in Figure 2.4.[36] Like the normal density, there is no way to write down the distribution function of χ^2 random variables, but numerical approximations work very well, so critical values are easy to compute via appropriate software. The 95% critical values for the cases $n = 1 \ldots 4$ are

degrees of freedom	1	2	3	4
critical value at 95%	3.84	5.99	7.81	9.49

For example, a χ^2_1 random variable takes values from 0 to 3.84 with probability 95%. Memorising them may turn out to be handy from time to time.

2.A.6 Gretl script to reproduce example 2.6

Input:

```
set verbose off
clear

# characteristics of the event

scalar p = 0.5
scalar n = 100
scalar lo = 36
scalar hi = 45
```

[36] In case you're wondering what $\Gamma(n/2)$ is, just google for "gamma function"; it's a wonderful object, you won't be disappointed. Suffice it to say that, if x is a positive integer, then $\Gamma(x) = (x+1)!$, but the gamma function is defined for all real numbers, except non-positive integers.

```
# true probability via the binomial distribution

matrix bin = pdf(B, p, n, seq(lo, hi)') # Binomial probabilities
scalar true = sumc(bin)

# approximation via the Central Limit Theorem

scalar m = p*n                  # mean
scalar s = sqrt(p*(1-p)*n)      # standard error
scalar z0 = (lo - 0.5 - m)/s    # subtract 0.5 to compensate for continuity
scalar z1 = (hi + 0.5 - m)/s    # add 0.5 to compensate for continuity

# "cnorm" = Normal distribution function

scalar appr = cnorm(z1) - cnorm(z0)

# printout

printf "probability of \"heads\" = %g\n", p
printf "number of tosses = %g\n", n
printf "probability of heads between %d and %d:\n", lo, hi
printf "true = %g, approximate via CLT = %g\n", true, appr
```

Output:

```
probability of "heads" = 0.5
number of tosses = 100
probability of heads between 36 and 45:
true = 0.182342, approximate via CLT = 0.182194
```

Chapter 3

Using OLS as an inferential tool

3.1 The regression function

In this chapter, we well revisit the OLS statistic and give it an inferential interpretation. As we will see, under many circumstances OLS is a consistent and asymptotically normal estimator. The first question that springs to mind is: what is OLS an estimator of, exactly?

Generally speaking, statistical inference can be a very useful tool when an observable variable y can be thought of as being influenced by a vector of observables $\mathbf{x}$, but only via a complicated causal chain, that possibly includes several other unobservable factors. Thus, we represent y as a random variable, so as to acknowledge our uncertainty about it; however, we have reasons to believe that y may not be independent from $\mathbf{x}$, so further insight can be gained by studying the conditional distribution $f(y|\mathbf{x})$.

Figure 3.1: Conditional distribution of the stature of children given their parents'

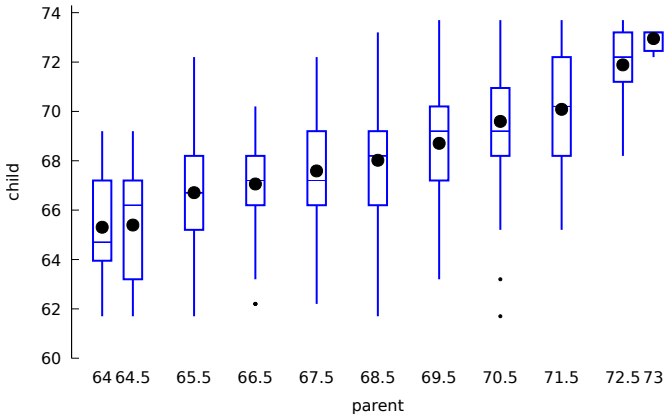
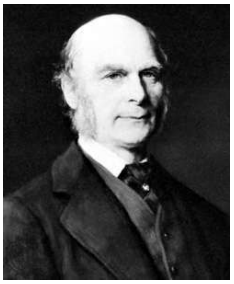


Figure 3.1 is my rendition of a celebrated dataset, that was studied by Galton (1886). Galton assembled data on the body height of 928 individuals (y), and

matched them against the average height of their parents (**x**). Data are in inches.



FRANCIS GALTON

It is natural to think that somebody's stature is the result of a multiplicity of causes, but surely the hereditary component cannot be negligible. Therefore, the interest in $f(y|\mathbf{x})$. For each observed value in of **x** in the sample, Figure 3.1 shows the corresponding boxplot.

Maybe not all readers are familiar with boxplots, so allow me to explain how to read the "candlesticks" in the figure: each vertical object consists of a central "box", from which two "whiskers" depart, upwards and downwards. The central box encloses the middle 50% of the data, i.e. it is bounded by the first and third quartiles. The "whiskers" extend from each end of the box for a range equal at most to 1.5 times the interquartile range. Observations outside that range are considered outliers[1] and represented via dots. A line is drawn across the box at the median. Additionally, a black dot indicates the average.

The most notable feature of Figure 3.1 is that the boxes seem to go up together with **x**; that is, the distribution of y shifts towards higher values as **x** grows. However, even considering the subset of observations defined as the children whose parents were of a certain height, some dispersion remains. For example, if we focus on $\mathbf{x} = 65.5$, you see from the third candlestick from the left that the minimum height is about 62 and the maximum is about 72, while the mean is between 66 and 68 (in fact, the precise figure is 67.059).

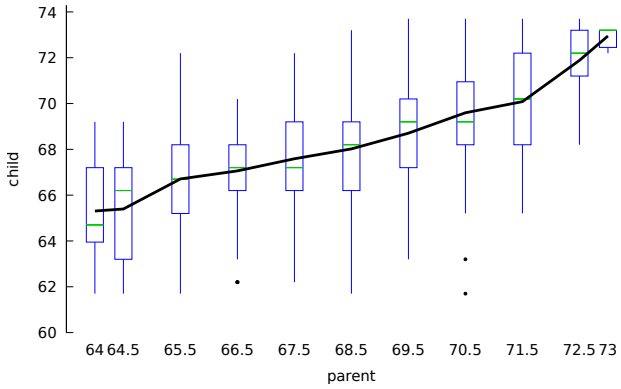
Historical curiosity: if you use OLS to go through those points such as to minimise the SSR, you will find that the fitted line is $\hat{c}_i = 23.9 + 0.646 p_i$, where c_i stands for "child" and p_i for "parent". The fact that the slope of the fitted line is less than 1 prompted Galton to observe that the tendency for taller parents was to have children who were taller than the average, but not as much as themselves (and of course the same, in reverse, happened to shorter parents). Galton described this state of things as "Regression towards Mediocrity", and the term stuck.

This method of inquiry is certainly interesting, but also very demanding: if **x** had been a vector of characteristics, rather than a simple scalar, the analysis would have been too complex to undertake. However, we may focus on a more limited problem, that is hopefully amenable to a neat and tidy solution.

Instead of studying $f(y|\mathbf{x})$, we could focus on the conditional expectation $\mathrm{E}[y|\mathbf{x}]$ (assuming it exists): this object contains the information on how the centre of the distribution of y varies across different values of **x**, and in most cases is just what we want. If you join the dots in figure 3.1, you get an upward-sloping line like in Figure 3.2, that suits very well our belief that taller parents should have, as a rule, taller children.

[1] An "outlier" is a data point that is far away from the rest.

Figure 3.2: Regression function of the stature of children given their parents'



The first step for making this intuition operational is to define the random variable $\varepsilon \equiv y - \mathrm{E}\left[y|\mathbf{x}\right]$, so that y can be written (by definition) as $\mathrm{E}\left[y|\mathbf{x}\right] + \varepsilon$. For historical reasons, the random variable ε is called the **disturbance** (see also section 3.A.2). A very important property of the random variable ε is that it's orthogonal to $\mathbf{x}$ by construction:[2]

$$\mathrm{E}\left[\mathbf{x} \cdot \varepsilon\right] = \mathbf{0} \tag{3.1}$$

The proof is simple: call $\mathrm{E}\left[y|\mathbf{x}\right] = m(\mathbf{x})$. Since $\varepsilon = y - m(\mathbf{x})$, clearly

$$\mathrm{E}\left[\varepsilon|\mathbf{x}\right] = \mathrm{E}\left[y|\mathbf{x}\right] - \mathrm{E}\left[m(\mathbf{x})|\mathbf{x}\right] = m(\mathbf{x}) - m(\mathbf{x}) = 0;$$

therefore, by the law of iterated expectations (see Section 2.2.4),

$$\mathrm{E}\left[\mathbf{x} \cdot \varepsilon\right] = \mathrm{E}\left[\mathbf{x} \cdot \mathrm{E}\left[\varepsilon|\mathbf{x}\right]\right] = \mathrm{E}\left[\mathbf{x} \cdot 0\right] = \mathbf{0}.$$

Finally, assume that $m(\mathbf{x})$ is a simple function, whose shape is governed by a few parameters.[3] The choice that is nearly universally made is that of a linear[4] function: $\mathrm{E}\left[y|\mathbf{x}\right] = \mathbf{x}'\boldsymbol{\beta}$. If we observe multiple realisations of y and $\mathbf{x}$, then we can write

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{3.2}$$

or, in matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.3}$$

Note the difference between equation (3.3) and the parallel OLS decomposition $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$, where everything on the right-hand side of the equation is an

[2] Warning: as shown in the text, $\mathrm{E}\left[\varepsilon|\mathbf{x}\right] = 0 \Longrightarrow \mathrm{E}\left[\mathbf{x} \cdot \varepsilon\right] = \mathbf{0}$, but the converse is not necessarily true.

[3] In fact, there are techniques for estimating the regression function directly, without resorting to assumptions on its functional form. These techniques are grouped under the term **nonparametric regression**. Their usage in econometrics is rather limited, however, chiefly because of their greater computational complexity and of the difficulty of computing marginal effects.

[4] As for what we mean exactly by "linear", see 1.3.2.

observable statistic. Instead, the only observable item in the right-hand side of (3.3) is $\mathbf{X}$: $\boldsymbol{\beta}$ is an unobservable vector of parameters and $\boldsymbol{\varepsilon}$ is an unobservable vector of random variables. Still, the similarity is striking; it should be no surprise that, under appropriate conditions, $\hat{\boldsymbol{\beta}}$ is a CAN estimator of $\boldsymbol{\beta}$, which we prove in the next section.

3.2 Main statistical properties of OLS

A handy consequence of equation (3.3) is that the OLS statistic can be written as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \tag{3.4}$$

As any estimator, $\hat{\boldsymbol{\beta}}$ has a distribution, and its finite-sample properties can be studied, in some cases. For example, its unbiasedness is very easy to prove: if $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$, then

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right] = \boldsymbol{\beta} + \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}\right] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{\beta}.$$

And therefore, by the law of iterated expectations,

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{E}\left[\mathrm{E}\left[\hat{\boldsymbol{\beta}}|\mathbf{X}\right]\right] = \mathrm{E}\left[\boldsymbol{\beta}\right] = \boldsymbol{\beta}.$$

However, nobody cares about unbiasedness nowadays. Moreover, in order to say something on the distribution of $\hat{\boldsymbol{\beta}}$ we'd need assumptions on the distribution of $\boldsymbol{\varepsilon}$, which is something we'd rather avoid doing. Therefore, we'll use asymptotic results. Of course, we will assume that the data are such that limit theorems apply (iid being but an example).

3.2.1 Consistency

In order to prove consistency, start from equation (3.4) and rewrite matrix products as sums:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left[\sum_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1}\sum_i \mathbf{x}_i\varepsilon_i = \boldsymbol{\beta} + \left[\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1}\frac{1}{n}\sum_i \mathbf{x}_i\varepsilon_i. \tag{3.5}$$

Let's analyse the two terms on the right-hand side separately: in order to do so, it will be convenient to define the vector

$$\mathbf{z}_i = \mathbf{x}_i\varepsilon_i; \tag{3.6}$$

given equation (3.1), $\mathrm{E}[\mathbf{z}] = \mathbf{0}$, so a straightforward application of the LLN gives

$$\frac{1}{n}\sum_i \mathbf{x}_i\varepsilon_i \xrightarrow{\mathrm{p}} \mathbf{0}. \tag{3.7}$$

As for the limit of the first term, assume that $n^{-1}\mathbf{X}'\mathbf{X}$ has one, and call it Q:[5]

$$\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i' \xrightarrow{\text{p}} Q; \tag{3.8}$$

if Q is invertible, then we can exploit the fact that inversion is a continuous transformation as follows

$$\left[\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1} \xrightarrow{\text{p}} Q^{-1},$$

so, after putting the two pieces together,

$$\hat{\boldsymbol{\beta}} \xrightarrow{\text{p}} \boldsymbol{\beta} + Q^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}.$$

The OLS statistic, therefore, is a consistent estimator of the parameters of the conditional mean, or to be more technical, of the derivative of $\mathrm{E}\left[y|\mathbf{x}\right]$ with respect to $\mathbf{x}$, which is constant by the linearity hypothesis.

One may think that the whole argument would break down if the assumption of linearity were violated. This is not completely true: even in many cases when $\mathrm{E}\left[y|\mathbf{x}\right]$ is non linear, it may be proven that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of the parameters of an object called *Optimal Linear Predictor*, which includes the linearity as a special case. But this is far too advanced for a book like this.

It's important here to ensure that $\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'$ converges to an invertible matrix; there are two main reasons while this requirement may fail to hold:

1. it may not converge to any limit; this would be the case if, for example, the vector $\mathbf{x}$ possessed no second moments;[6]

2. it may converge to a singular matrix; this, for example, would happen in cases such as $x_t = \phi^t$, where $|\phi| < 1$.

However, in ordinary circumstances, such problems should not arise.

3.2.2 Asymptotic normality

From equation (3.5),

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left[\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1} \frac{1}{\sqrt{n}}\sum_i \mathbf{x}_i\varepsilon_i$$

[5] Ordinarily, Q will be equal to $E(\mathbf{x}_i\mathbf{x}_i')$. It may be interesting to know that the proeprties of OLS can be worked out in more exotic cases, where Q is more complicated or may even not exist. These, cases, however, are far too complicated to be analysed here.

[6] In fact, there are cases when this situation may be handled by using a scaling factor other than n^{-1}; but let's ignore such acrobatics.

we already know from the previous subsection that $\left[\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1} \xrightarrow{\text{p}} Q^{-1}$, but what happens to the second term as n grows to infinity? Define $\mathbf{z}_i$ as in equation (3.6). Therefore, by the CLT,

$$\frac{1}{\sqrt{n}}\sum_i \mathbf{x}_i\varepsilon_i = \sqrt{n}\bar{\mathbf{Z}} \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0},\Omega),$$

where $\Omega \equiv \mathrm{V}[\mathbf{z}_i] = \mathrm{E}\left[\mathbf{z}_i\mathbf{z}_i'\right]$. Therefore, we can use Cramér's theorem (see (2.13)) as follows: since

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &= \left[\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1}\frac{1}{\sqrt{n}}\sum_i \mathbf{x}_i\varepsilon_i \\
\left[\frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1} &\xrightarrow{\text{p}} Q^{-1} \\
\frac{1}{\sqrt{n}}\sum_i \mathbf{x}_i\varepsilon_i &\xrightarrow{\text{d}} \mathcal{N}(\mathbf{0},\Omega)
\end{aligned}$$

the quantity $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ converges to a normal rv multiplied by the nonstochastic matrix Q^{-1}; therefore, the linearity property for Gaussian rvs applies (see eq. (2.19)), and as a consequence

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{\text{d}} \mathcal{N}\left(\mathbf{0},Q^{-1}\Omega Q^{-1}\right). \tag{3.9}$$

In order to say something on Ω, we can use the law of iterated expectations (see section 2.2.4) to re-write it as follows:

$$\Omega = \mathrm{E}\left[\mathbf{z}_i\mathbf{z}_i'\right] = \mathrm{E}\left[\varepsilon_i^2\cdot\mathbf{x}_i\mathbf{x}_i'\right] = \mathrm{E}\left[\mathrm{E}\left[\varepsilon_i^2|\mathbf{x}_i\right]\cdot\mathbf{x}_i\mathbf{x}_i'\right].$$

The quantity $\mathrm{E}\left[\varepsilon_i^2|\mathbf{x}_i\right]$ is a bit of a problem here. We proved quite easily that $\mathrm{E}[\varepsilon|\mathbf{x}] = 0$ as a natural consequence of the way ε is defined (see section 3.1, page 79) . However, we know nothing about its conditional second moment (the conditional variance of y, if you like). For all we know, it may even not exist; or if it does, it could be an arbitrary (and possibly quite weird) function of $\mathbf{x}$. The only thing we can be sure of is that the function $h(\mathbf{x}) = \mathrm{E}\left[\varepsilon^2|\mathbf{x}\right]$ (sometimes called the **skedastic** function) must be positive, since it's the expectation of a square and of course the support of ε^2 is the positive real line, or possibly a subset.

In some cases, one could be tempted to set up a model in which we assume a functional form for the skedastic function in the same way as we do for the regression function. This, however, is very seldom done: the computational complexity is greater than OLS and there is little interest in the parameters of the conditional variance.

That said, there are certain situations where the main object of interest is the conditional variance instead of the conditional mean, like for example in certain models used in finance. A fuller discussion of this topic would lead to a concept called **heteroskedasticity**, which is the object of section 4.2.

In the present chapter, we will assume that $\mathrm{E}\left[\varepsilon^2|\mathbf{x}\right]$ is a positive constant, traditionally labelled σ^2:

$$\mathrm{E}\left[\varepsilon^2|\mathbf{x}\right] = \sigma^2; \tag{3.10}$$

the most important implication of this assumption is that the conditional variance $\mathrm{V}\left[y_i|\mathbf{x}_i\right]$ is constant for all observations $i = 1\ldots n$; this idea is known as **homoskedasticity**.[7] This assumption can be visualised in terms of Figure 3.1 as the idea that all the boxplots look roughly the same, and all you get by moving along the horizontal axis is that they may go up and down as an effect of $\mathrm{E}\left[y|\mathbf{x}\right]$ not being constant, but never change their shape and size. How realistic this assumption is in practice remains to be seen, and a sizeable part of chapter 4 will be devoted to this issue, but for the moment let's just pretend this is not a problem.

Therefore, under the homoskedasticity assumption,

$$\Omega = \mathrm{E}\left[\mathbf{z}_i\mathbf{z}_i'\right] = \mathrm{E}\left[\sigma^2 \cdot \mathbf{x}_i\mathbf{x}_i'\right] = \sigma^2 \cdot \mathrm{E}\left[\mathbf{x}_i\mathbf{x}_i'\right] = \sigma^2 Q$$

and equation (3.9) simplifies to

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mathbf{0}, \sigma^2 Q^{-1}\right). \tag{3.11}$$

This result is also important because it provides the basis for justifying the usage of OLS as an estimator of $\boldsymbol{\beta}$ on the grounds of its efficiency. Traditionally, this is proven via the so-called **Gauss-Markov** theorem, which, however, relies quite heavily on small-sample results that I don't like very much.[8] In fact, there is a much more satisfactory proof that OLS is *asymptotically semiparametrically* efficient, but it's considerably technical, so it's way out of scope here.[9] Suffice it to say that, under homoskedasticity, OLS is hard to beat in terms of efficiency.

We can estimate consistently Q via $n^{-1}\mathbf{X}'\mathbf{X}$ and σ^2 via[10]

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n} \xrightarrow{\mathrm{p}} \sigma^2 \tag{3.12}$$

so the approximate distribution for OLS that we use is

$$\hat{\boldsymbol{\beta}} \overset{\mathrm{a}}{\sim} N\left[\boldsymbol{\beta}, \hat{V}\right]$$

where $\hat{V} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.

A word of warning: the expression above $\hat{V}$ is not the one used by the majority of econometrics textbooks, and by none of the major econometric software

[7] From the Greek prefix "homo" (same); "skedastic", in an econometric context, means "that has to do with variance".

[8] If you really really care, a proof is given in section 3.A.3, but I don't care about it very much myself.

[9] See Hansen (2019), sections 7.20 and 7.21 if you're interested.

[10] Proof of this is unnecessary, but if you insist, go to subsection 3.A.1.

packages. Instead, in the more popular variant an alternative estimator of σ^2 is used:

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k},$$

where k is the number of elements of $\boldsymbol{\beta}$; the number $n-k$ is commonly known as **degrees of freedom**. It's easy to prove that s^2 is also consistent, but (as can be proven via some neat algebra trick that I'm sparing you here) it's also unbiased: $\mathrm{E}\left[s^2\right] = \sigma^2$.

The difference between the two variants is negligible if the sample size n is reasonably large, so you can use either; or to put it otherwise, if using s^2 instead of $\hat{\sigma}^2$ makes a substantial difference, then n is probably so small that in my opinion you shouldn't be using statistical inference in the first place. And besides, I see no convincing reason why unbiasedness should be considered a virtue in this context. The usage of $\hat{\sigma}^2$, instead, makes many things easier and is consistent with all the rest of procedures that econometricians use beyond OLS, in which asymptotic results are uniformly used. Having said this, let's move on.

3.2.3 In short

To summarise: a set of conditions that are necessary for OLS to be interpreted as a CAN estimator of something meaningful are:

1. we have a sample of n observations on y_i (our dependent variable) and $\mathbf{x}_i$ (our explanatory variables) that satisfies some basic requirements so that asymptotic theory is applicable as a reasonable approximation of the behaviour of sample statistics. For example, the observation are iid and all moments exist.

2. The conditional expectation of y on $\mathbf{x}$ exists and is linear: $\mathrm{E}\left[y|\mathbf{x}\right] = \mathbf{x}'\boldsymbol{\beta}$.

3. The matrix $n^{-1}\mathbf{X}'\mathbf{X}$ converges in probability to an invertible matrix Q.

4. The conditional variance $\mathrm{V}\left[y|\mathbf{x}\right]$ is a constant, that we call σ^2.

If the above is true, then $\hat{\boldsymbol{\beta}}$ can be regarded as a CAN estimator of the parameters of the conditional mean, that can be used, in turn, to compute marginal effects or, as we shall see in the next section, to perform hypothesis tests. Note that the above hypotheses are *sufficient*, but some of them may be relaxed to some degree, and we will do so in the next chapters.

The reader may also find it interesting that an alternative set of assumptions customarily known as the **classical assumptions** was traditionally made when teaching econometrics in the twentieth century. In my opinion, using the classical assumptions for justifying the usage of OLS as an estimator is a relic of the past, but if you're into the history of econometric thought, I wrote a brief description in section 3.A.2.

3.3 Specification testing

3.3.1 Tests on a single coefficients

Sometimes, we would like to test hypotheses on elements of $\boldsymbol{\beta}$, so that we can decide optimally on which explanatory variables we must include in our regression function. This is often called "specification testing".

Let's begin by testing a very simple hypothesis: $H_0 : \beta_i = 0$. The practical implication of H_0 in a model like

$$\mathrm{E}\left[y|\mathbf{x}\right] = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k = \sum_{j=1}^{k} x_j\beta_j$$

is that the impact of x_i on $\mathrm{E}\left[y|\mathbf{x}\right]$ is 0, and therefore that the i-th explanatory variable is irrelevant, since it has no effect on the regression function.

Note that under H_0 there are two equally valid representations of the regression function; one that includes x_i, the other one that doesn't. For example, for $i = 2$ and $k = 3$,

$$\text{Model A} \qquad y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \varepsilon_i \tag{3.13}$$

$$\text{Model B} \qquad y_i = x_{1i}\beta_1 \qquad\quad + x_{3i}\beta_3 + \varepsilon_i \tag{3.14}$$

Clearly, if H_0 was true, model B would be preferable, chiefly on the grounds of parsimony;[11] however, if H_0 was false, only model A would be a valid representation of the regression function.

As I explained in section 2.4, in order to test H_0 we need to find a differentiable function of $\boldsymbol{\beta}$ such that $g(\boldsymbol{\beta}) = 0$ if and only if H_0 is true. In this case, this is very easy: define $\mathbf{u}_i$ as the "extraction vector", that is a vector of zeros, except for the i-th element, which is 1.[12] The extraction vector takes its name by the fact that the inner product of $\mathbf{u}_j$ by any vector $\mathbf{a}$ returns the j-element of $\mathbf{a}$. More generally, the product $A \cdot \mathbf{u}_j$ yields the j-th column of A, while $\mathbf{u}_i' A$ yields its i-th row.[13] Evidently, $\mathbf{u}_i' A \mathbf{u}_j = A_{ij}$, the element of A on row i and column j.

By defining $g(\boldsymbol{\beta}) = \mathbf{u}_i'\boldsymbol{\beta}$, the hypothesis $H_0 : \beta_i = 0$ can be written as $H_0 : \mathbf{u}_i'\boldsymbol{\beta} = 0$, and the Jacobian term is simply $\mathbf{u}_i'$. Hence

$$\sqrt{n}\left[\mathbf{u}_i'\hat{\boldsymbol{\beta}} - \mathbf{u}_i'\boldsymbol{\beta}\right] \xrightarrow{\mathrm{d}} N\left[0, \mathbf{u}_i' V \mathbf{u}_i\right]$$

so our test statistic is

$$W = \hat{\boldsymbol{\beta}}'\mathbf{u}_i \left(\mathbf{u}_i'\hat{V}\mathbf{u}_i\right)^{-1} \mathbf{u}_i'\hat{\boldsymbol{\beta}} = \frac{\hat{\beta}_i^2}{v_{ii}} = t_i^2 \tag{3.15}$$

[11] In fact, we will argue in section 3.5 that OLS on model B would produce a more efficient estimator of the remaining coefficients.

[12] Some call the vector $\mathbf{u}_i$ a **basis vector**. Others simply say "the i-th column of the identity matrix".

[13] As always, the reader should verify the claim, instead of trusting me blindly.

where $t_i = \frac{\hat{\beta}_i}{se_i}$, v_{ii} is the i-th element on the diagonal of $\hat{V}$ and $se_i = \sqrt{v_{ii}}$. The quantity se_i is usually referred to as the **standard error** of $\hat{\beta}_i$. Of course, the null hypothesis would be rejected if $W > 3.84$ (the 5% critical value for a χ^2_1 distribution). Of course, this implies that we'd reject when $|t| > \sqrt{3.84} = 1.96$.

In fact, it's rather easy to prove that we could use a slight generalisation of the above for constructing a test for $H_0 : \beta_i = a$, where a is any real number you want, and that such a test takes the form

$$t_{\beta_i = a} = \frac{\hat{\beta}_i - a}{se_i} \tag{3.16}$$

Clearly, we can use the t statistic to decide whether a certain explanatory variable is irrelevant or not, and therefore choose between model A and model B. In the next subsection, I will show how this idea can be nicely generalised so as to frame the decision on the best specification via hypothesis testing.

Note, also, that the t statistic can also be used "in reverse" to construct confidence intervals in the same way as discussed at the end of Section 2.3.2: instead of asking ourselves what the decision on H_0 would be for a given a, we may look for the values of a that would lead us to a given decision. A hypothesis of the kind $H_0 : \beta_j = a$ is not rejected whenever

$$-1.96 < \frac{\hat{\beta}_i - a}{se_i} < 1.96;$$

therefore, the range of values for a that would lead to accepting H_0 is

$$\hat{\beta}_j - 1.96 \cdot se_i < a < \hat{\beta}_j + 1.96 \cdot se_i.$$

In other words, the interval $\hat{\beta}_j \pm 1.96 \cdot se_i$ contains all the values of a that we may consider not contradictory to the observed data.

3.3.2 More general tests

The general idea I will pursue here can be stated as follows: we assume we have a true representation of the conditional mean, that I will call the **unrestricted** model; in the previous subsection, that was called "model A". Additionally, we conjecture that the parameters β may obey some **restrictions** (also known as **constraints**); by incorporating those into our unrestricted model, we would obtain a more compact representation of the regression function. This, however, would be valid only if our conjecture was true. This model is known as the **restricted** model, and was labelled "model B" in the previous subsection. We make our decision on the model to adopt by testing if our conjecture via a standard hypothesis test.

In order to exemplify this idea, I will use the following unrestricted model:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \varepsilon_i; \tag{3.17}$$

in the previous subsection, we saw the the restricted model corresponding to the constraint $\beta_2 = 0$ is

$$y_i = x_{1i}\beta_1 + x_{3i}\beta_3 + \varepsilon_i.$$

Suppose now that instead of $\beta_2 = 0$, our conjecture was $\beta_1 = 1$; by inserting this equality into (3.17) we obtain

$$y_i = x_{1i} + x_{2i}\beta_2 + x_{3i}\beta_3 + \varepsilon_i$$

and therefore the restricted version of (3.17) would become

$$y_i - x_{1i} = x_{2i}\beta_2 + x_{3i}\beta_3 + \varepsilon_i,$$

so that, in fact, we would be studying the regression function of the observable variable $\tilde{y}_i = y_i - x_{1i}$ on x_{2i} and x_{3i}. Note that in this case we would have to redefine the dependent variable of our model.

One more example: suppose we combine (3.17) with the restriction $\beta_2 + \beta_3 = 0$: in this case, the constrained model turns out to be

$$y_i = x_{1i}\beta_1 + (x_{2i} - x_{3i})\beta_2 + \varepsilon_i.$$

Of course you can combine more than one restriction into a system:

$$\begin{cases} \beta_1 = 1 \\ \beta_2 + \beta_3 = 0, \end{cases}$$

and if you applied these to (3.17), the constrained model would turn into

$$y_i - x_{1i} = (x_{2i} - x_{3i})\beta_2 + \varepsilon_i.$$

The best way to represent constraints of the kind we just analysed is via the matrix equation

$$R\boldsymbol{\beta} = \mathbf{d},$$

where the matrix R and the vector $\mathbf{d}$ are chosen so as to express the constraints we want to test. The examples above on model (3.17) are translated into the $R\boldsymbol{\beta} = \mathbf{d}$ form in the following table:

Constraint	R	$\mathbf{d}$	Restricted model
$\beta_3 = 0$	$\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$	0	$\mathrm{E}\left[y_i \vert \mathbf{x}_i\right] = x_{1i}\beta_1 + x_{2i}\beta_2$
$\beta_1 = 1$	$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$	1	$\mathrm{E}\left[y_i - x_{1i} \vert \mathbf{x}_i\right] = x_{2i}\beta_2 + x_{3i}\beta_3$
$\beta_2 + \beta_3 = 0$	$\begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$	0	$\mathrm{E}\left[y_i \vert \mathbf{x}_i\right] = x_{1i}\beta_1 + (x_{2i} - x_{3i})\beta_2$
$\begin{cases} \beta_1 = 1 \\ \beta_2 = \beta_3 \end{cases}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\mathrm{E}\left[y_i - x_{1i} \vert \mathbf{x}_i\right] = (x_{2i} - x_{3i})\beta_2$

Since the Jacobian of $R\boldsymbol{\beta} - \mathbf{d}$ with respect to $\boldsymbol{\beta}$ is just the matrix R, we can adapt the apparatus of Section 3.3.1 to the present case and decide on the appropriateness of the restriction by computing the statistic

$$W = (R\hat{\boldsymbol{\beta}} - \mathbf{d})' \left[R\hat{V}R'\right]^{-1} (R\hat{\boldsymbol{\beta}} - \mathbf{d}) \tag{3.18}$$

and matching it to the χ^2 distribution with p degrees of freedom, p being the number of constraints (the number of rows of the R matrix, if you prefer).

It should be noted, at this point, that checking for the compatibility of a conjecture such as $R\boldsymbol{\beta} = \mathbf{d}$ may be a good idea for several reasons that go beyond the simple task of choosing the most parsimonious representation for the regression function. The hypothesis itself could be of interest: for example, the coefficient β_j could measure the response of the incidence of autism to the percentage of vaccinated children. From a policy perspective, it would be extremely important if $H_0 : \beta_j = 0$ were rejected (I'd be very surprised if it were).

Additionally, econometric models are often written in terms of parameters that can be given a direct interpretation in terms of economic theory. As an example, take a Cobb-Douglas production function: $Q = AK^{\alpha_1}L^{\alpha_2}$. The reader is doubtlessly familiar enough with microeconomics to need no reminder that scale economies are constant if and only if $\alpha + \alpha_2 = 1$. The production function, in logs, reads

$$q = a + \alpha_1 k + \alpha_2 l,$$

where $k = \log(K)$ and $l = \log(L)$. If we could perform an experiment in which we vary k and l to our liking and observe the resulting q, it would be very natural to estimate the parameter vector

$$\boldsymbol{\beta} = \begin{pmatrix} a \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$$

by means of an OLS regression, and the hypothesis "constant returns to scale" would simply amount to $R\boldsymbol{\beta} = \mathbf{d}$, with

$$R = [0 \quad 1 \quad 1] \qquad \mathbf{d} = 1.$$

Nevertheless, if we knew for certain (by some supernatural revelation) that our production function displays in fact constant returns to scale, we would like our estimate of $\boldsymbol{\beta}$ to incorporate the information $\alpha_1 + \alpha_2 = 1$, and there is no reason why $\hat{\boldsymbol{\beta}}$ should.

In section 3.5, we will develop an alternative estimator, known as the **Restricted Least Squares** estimator (or RLS for short), which integrates sample data with one or more a priori constraints on the parameter vector. As we will see, this will have the additional advantage of providing us with to calculate the W test statistic, by comparing the SSRs for the two versions of the model.

3.4 Example: reading the output of a software package

Now it's time for a hands-on example:[14] Table 3.1 contains a regression on a dataset containing data about 2610 home sales in Stockton, CA from Oct 1, 1996 to Nov 30, 1998;[15] the dependent variable is the natural logarithm of their sale price. For this example, the software package I used is gretl, but the output is more or less the same with every other program.

The model we're going to estimate can be written as follows:

$$p_i = \beta_0 + \beta_1 s_i + \beta_2 b_i + \beta_3 a_i + \beta_4 x_i + \varepsilon_i \tag{3.19}$$

where p_i is the log price of the i-th house and the explanatory variables are:

Legend		
`lsize`	s_i	log of living area, hundreds of square feet
`baths`	b_i	number of baths
`age`	a_i	age of home at time of sale, years
`pool`	x_i	= 1 if home has pool, 0 otherwise

Models of this type, where the dependent variable is the price of a good and the explanatory variables are its features, are commonly called **hedonic models**. In this case (like in most hedonic models), the dependent variable is in logarithm; therefore, the effect of all coefficients must be interpreted as the impact on that variable on the *relative* change in the house price (see footnote 34 in Chapter 1).

As you can see, the output is divided into two tables; the most interesting is the top one, which contains $\hat{\boldsymbol{\beta}}$ and some more statistics. I'll describe the contents of the bottom table in section 3.4.2.

3.4.1 The top table: the coefficients

The top table contains one row for each regressor. In the five columns you have:

1. the regressor name
2. the corresponding element of $\hat{\boldsymbol{\beta}}$, that is $\hat{\beta}_i$;
3. the corresponding standard error, that is $se_i = \sqrt{s^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}_{ii}}$;
4. the ratio of those two numbers, that is the t-ratio (see eq. 3.15)

[14] After reading this section, the reader might want to go back to section 1.5 and read it again from a different perspective.

[15] Data are taken from Hill et al. (2018); if you use gretl, you can find the data in `gdt` format at `http://www.learneconometrics.com/gretl/poe5/data/stockton5.gdt`. This is part of the rich offering you find on Lee Adkins' excellent website `http://www.learneconometrics.com/gretl/`, which also contains Lee's book. Highly recommended.

```
Dependent variable: lprice

            coefficient   std. error    t-ratio    p-value
  ------------------------------------------------------------
  const       8.85359     0.0483007     183.3      0.0000   ***
  lsize       1.03696     0.0232121      44.67     0.0000   ***
  baths      -0.00515142  0.0130688      -0.3942   0.6935
  age        -0.00238675  0.000270997    -8.807    2.29e-18 ***
  pool        0.106793    0.0226757       4.710    2.61e-06 ***

Mean dependent var   11.60193   S.D. dependent var   0.438325
Sum squared resid    157.8844   S.E. of regression   0.246187
R-squared            0.685027   Adjusted R-squared   0.684544
F(4, 2605)           1416.389   P-value(F)           0.000000
Log-likelihood      -42.58860   Akaike criterion     95.17721
Schwarz criterion    124.5127   Hannan-Quinn         105.8041
```

Table 3.1: Example: house prices in the US

5. the corresponding p-value, possibly with the little stars on the right (see section 2.4.1).

Note that gretl, like all econometric packages I know, gives you the "finite-sample" version of the standard errors, that is those computed by using s^2 as a variance estimator instead of $\hat{\sigma}^2$, which is what I personally prefer, but the difference would be minimal.

For the interpretation of each row, let's begin by the `lsize` variable:[16] the coefficient is positive, so that in our dataset bigger houses sell for higher prices, which of course stands to reason. However, the magnitude of the coefficient is also interesting: 1.037 is quite close to one. Since the house size is also expressed in logs, we could say that the *relative* response of the house price to a *relative* variation in the house size is 1.037. For example, if we compared two houses where house A is bigger than house B by 10% (and all other characteristics were the same), we would expect the value of house A to be 10.37% higher than that of house B.

As the reader knows, this is what in economics we call an elasticity: for a continuous function you have that

$$\eta = \frac{\mathrm{d}y}{\mathrm{d}x} \cdot \frac{x}{y} = \frac{\mathrm{d}\log y}{\mathrm{d}\log x}$$

because $\frac{\mathrm{d}\log y}{\mathrm{d}y} = \frac{1}{y}$ and therefore $\mathrm{d}\log y = \frac{\mathrm{d}y}{y}$. So, any time you see something like $\log(y) = a + b\log(x)$, you can safely interpret b as the elasticity of y to x.

From an economic point of view, therefore, we would say that the elasticity of the house price to its size is about 1. What is more interesting, the standard

[16] "Why not the constant?" you may ask. *Nobody* cares about the constant.

error for that coefficient is about 0.023, which gives a t-ratio of 44.67, and the corresponding p-value is such an infinitesimal number that the software just gives you 0.[17] If we conjectured that there was no effect of size on price, that hypothesis would be strongly rejected on the grounds of empirical evidence. In the jargon of applied economists, we would say that size is *significant* (in this case, very significant).

If, instead, we wanted to test the more meaningful hypotheses $H_0 : \beta_1 = 1$, it would be quite easy to compute the appropriate t statistic as per equation (3.16):

$$t = \frac{\hat{\beta}_1 - 1}{se_1} = \frac{1.03696 - 1}{0.0232121} = 1.592$$

and the corresponding p-value would be about 11.1%, so we wouldn't reject H_0.

On the other hand, we get a slightly negative effect for the number of baths (−0.00515142). At first sight, this does not make much sense, since you would expect that the more baths you have in the house, the more valuable your property is. How come we observe a negative effect?

There are two answers to this question: first, the p-value associated to the coefficient is 0.6935, which is way over the customary 0.05 level. In other words, an applied economist would say that the `baths` variable is *not significant*. This does not mean that we can conclusively deduce that there is no effect. It means that, if there is one, it's too weak for us to detect (and, for all we know, it might as well be positive instead, albeit quite limited). Moreover, this is the effect of the number of baths *other things being equal*, as we know from the Frisch-Waugh theorem (see section 1.4.4). In this light, the result is perhaps less surprising: why should the number of baths matter, given the size of the house? Actually, a small house filled with baths wouldn't seem such a great idea, at least to me.

On the contrary, the two variables `age` and `pool` are highly significant. The coefficient for age, for example, is about -0.002: each year of age decreases the house value by about 0.2%, which makes sense. The coefficient for the dummy variable `pool` is about 0.107, so it would seem that having a pool increases the house value by a little over 10%, which, again, makes sense.

3.4.2 The bottom table: other statistics

Let's begin with the easy bits; the first line of the bottom table contains descriptive statistics for the dependent variable: mean (about 11.6) and standard deviation (about 0.43). The next line contains the sum of squared residuals $\mathbf{e}'\mathbf{e}$ (157.88) and the square root of $s^2 = \mathbf{e}'\mathbf{e}/(n-k)$, which is in this case about 0.246. Since our dependent variable is in logs, this means that the "typical" size of the approximation errors for our model is roughly 25%. The line below contains the R^2 index and its adjusted variant (see eq. 1.19). Both versions are around 68.5%, so that our model, all in all, does a fair job at describing price differentials between houses, especially given how little information on each properties

[17] In case you're curious: I can't compute the number exactly, but it's smaller than 10^{-310}.

we have. However, since our estimate of σ^2 is quite sizeable, we shouldn't expect our model to give us a detailed description of individual house prices.

The line below contains a test[18] commonly called "overall specification test": it is a joint test for all coefficient being zero apart from the constant. The null hypothesis is, basically, that none of your regressors make any sense and your model is utter rubbish. Luckily, in our case the p-value is infinitesimal, so we reject.

On the next line, you get the **log-likelihood** for the model:

$$L = -\frac{n}{2}\left[1 + \ln(2\pi) + \ln(\hat{\sigma}^2)\right].$$

This number is of very little use by itself;[19] in this book, it's only important because it provides the essential ingredient for calculating the so-called **Information Criteria** (IC), that are widely used tools for comparing *non-nested* models.

We say that a model *nests* another one when the latter is a special case of the former. For example, the two models (3.13) and (3.14) are nested, because model (3.14) is just model (3.13) in the special case $\beta_2 = 0$. If model B is nested in model A, choosing between A and B is rather easy: all you need is a proper test statistic; I will provide a detailed exposition in Section 3.5. However, we may have to choose between the two alternatives in which nesting is impossible:

$$\begin{aligned} y_i &\simeq \mathbf{x}_i'\boldsymbol{\beta} \\ y_i &\simeq \mathbf{z}_i'\boldsymbol{\gamma} \end{aligned}$$

Information criteria start from the value of the log-likelihood (multiplied by -2) and add a *penalty function*, which is increasing in the number of parameters. The gretl package, that I'm using here, reports three criteria: the Akaike IC (AIC), the Schwartz IC (BIC, where B is for "Bayesian") and the one by Hannan and Quinn (HQC), which differ by the choice of penalty function:

$$\text{AIC} = -2L + 2k \tag{3.20}$$

$$\text{BIC} = -2L + k\log n \tag{3.21}$$

$$\text{HQC} = -2L + 2k\log\log n \tag{3.22}$$

The rule is to pick the model that minimises information criteria. It may be interesting to know that, for the case of linear model that we are examining in the present context, the quantity $-2L$ equals

$$-2L = n\left[K - \log(\hat{\sigma}^2)\right],$$

[18] This test, technically, is of the F variety — see section 3.5.1 for its definition.

[19] If the data are iid and $f(y|\mathbf{x})$ is Gaussian, then $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2. I chose not to include this topic into this book, but the interested reader will find a nice and compact exposition in Verbeek (2017), chapter 6. Other excellent treatments abound, but the curious reader may want to check out chapters 14–15 of Ruud (2000). If you want to go for the real thing, grab Gourieroux and Monfort (1995), volume 1, chapter 7.

where K is a not particularly interesting constant. Therefore, minimising the information criteria amounts to choosing a model that fits the data "well" without using "too many" parameters.

At times, it may happen that this algorithm gives conflicting results depending on which IC you choose. There is a huge literature on this, but my advice in these cases is "don't trust the AIC much".[20] An alternative to information criteria that has become very popular in recent years (especially because the machine learning people are crazy about it) is the so-called **cross-validation criterion**: you'll find more about it in Section 3.A.4.[21]

3.5 Restricted Least Squares and hypothesis testing

The **Restricted Least Squares** statistic (or **RLS** for short) is an estimator of the parameter vector that, like OLS, uses the available data in the most effective way but at the same time, unlike OLS, satisfies by construction a set of p restrictions of the type $R\boldsymbol{\beta} = \mathbf{d}$. In other words, we are looking for a vector $\tilde{\boldsymbol{\beta}}$ that minimises the SSR under the condition that a certain set of linear constraints are satisfied:

$$\tilde{\boldsymbol{\beta}} = \underset{R\boldsymbol{\beta}=d}{\text{Argmin}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|; \tag{3.23}$$

compare (3.23) with equation (1.14): OLS is defined as the unconstrained SSR minimiser (we can choose $\hat{\boldsymbol{\beta}}$ among all k-element vectors); RLS, instead, can only be chosen among those vectors $\boldsymbol{\beta}$ that satisfy $R\boldsymbol{\beta} = \mathbf{d}$. Figure 3.3 exemplifies the situation for $k = 2$.

Define the restricted residuals as $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$; we will be interested in comparing them with the OLS residuals, so in this section we will denote them as $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{M}_\mathbf{X}\mathbf{y}$ to make the distinction typographically evident.

A couple of remarks can already be made even without knowing what the solution to the problem in (3.23) is. First, since $\hat{\boldsymbol{\beta}}$ is an unrestricted minimiser, $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ cannot be larger than the constrained minimum $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$. However, the inequality $\hat{\mathbf{e}}'\hat{\mathbf{e}} \leq \tilde{\mathbf{e}}'\tilde{\mathbf{e}}$ can be made more explicit by noting that

$$\mathbf{M}_\mathbf{X}\tilde{\mathbf{e}} = \mathbf{M}_\mathbf{X}\left[\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right] = \mathbf{M}_\mathbf{X}\mathbf{y} = \hat{\mathbf{e}}$$

and therefore

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = \tilde{\mathbf{e}}'\mathbf{M}_\mathbf{X}\tilde{\mathbf{e}} = \tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \tilde{\mathbf{e}}'\mathbf{P}_\mathbf{X}\tilde{\mathbf{e}}$$

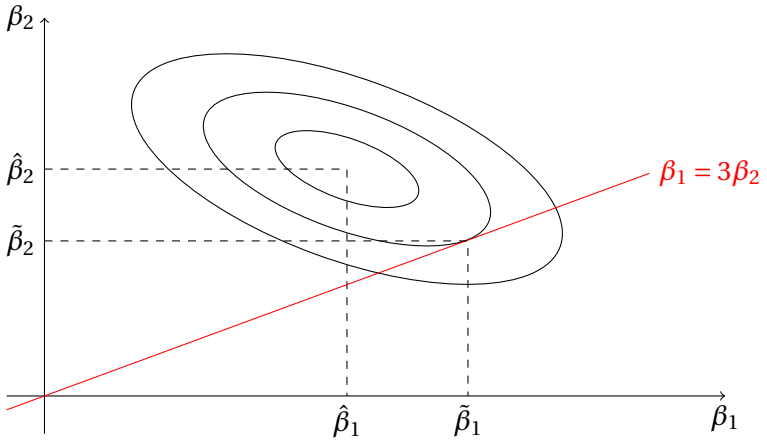
so that

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}} = \tilde{\mathbf{e}}'\mathbf{P}_\mathbf{X}\tilde{\mathbf{e}} \tag{3.24}$$

[20] If you want some more detail, see section 15.4 in Davidson and MacKinnon (2004) or section 3.2.2 in Verbeek (2017). However, the literature on statistical methods for selecting the "best" model (whatever that may mean) is truly massive; see for example the "model selection" entry in (Durlauf and Blume, 2008).

[21] In fact, the cross validation criterion can be shown to be roughly equivalent to the AIC.

Figure 3.3: Example: two-parameter vector



The ellipses are the contour lines of the function $\mathbf{e}'\mathbf{e}$*. The constraint is* $\beta_1 = 3\beta_2$*. The number of parameters* k *is 2 and the number of constraint* p *is 1. The unconstrained minimum is* $\hat{\beta}_1, \hat{\beta}_2$*; the constrained minimum is* $\tilde{\beta}_1, \tilde{\beta}_2$*.*

where the right-hand side of the equation is non-negative, since $\tilde{\mathbf{e}}'\mathbf{P}_{\mathbf{X}}\tilde{\mathbf{e}}$ can be written as $(\mathbf{P}_{\mathbf{X}}\tilde{\mathbf{e}})'(\mathbf{P}_{\mathbf{X}}\tilde{\mathbf{e}})$, which is a sum of squares.[22]

In order to solve (3.23) for $\tilde{\boldsymbol{\beta}}$, we need to solve a constrained optimisation problem, which is not complicated once you know how to set up a Lagrangean. The details, however, are not important here and I'll give you the solution straight away:

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}R'\left[R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - \mathbf{d}); \tag{3.25}$$

derivation of this result is provided in the separate subsection 3.A.5, so you can skip it if you want.

The statistical properties of $\tilde{\boldsymbol{\beta}}$ are proven in section 3.A.6, but the most important point to make here are: if $R\hat{\boldsymbol{\beta}} = \mathbf{d}$, then $\tilde{\boldsymbol{\beta}}$ is consistent just like the OLS estimator $\hat{\boldsymbol{\beta}}$, but has the additional advantage of being more efficient. If, on the contrary, $R\hat{\boldsymbol{\beta}} \neq \mathbf{d}$, then $\hat{\boldsymbol{\beta}}$ is inconsistent. The practical consequence of this fact is that, if we were certain that the equation $R\hat{\boldsymbol{\beta}} = \mathbf{d}$ holds, we would be much better off by using an estimator that incorporates this information; but if our conjecture is wrong, our inference would be invalid.

It's also worth noting that nobody uses expression (3.25) as a computational device. The simplest way to compute $\tilde{\boldsymbol{\beta}}$ is to run OLS on the restricted model and then "undo" the restrictions: for example, if you take model (3.17), reproduced

[22] In fact, the same claim follows more elegantly by the fact that $\mathbf{P}_{\mathbf{X}}$ is, by construction, positive semi-definite.

here for your convenience

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \varepsilon_i$$

and want to impose the set of restrictions $\beta_1 = 1$ and $\beta_2 = \beta_3$, what you would do is estimating the constrained version

$$y_i - x_{1i} = (x_{2i} + x_{3i})\beta_2 + \varepsilon_i, \tag{3.26}$$

that can be "unravelled" as

$$y_i = x_{1i} \cdot 1 + x_{2i}\beta_2 + x_{3i}\beta_2 + \varepsilon_i$$

and then forming $\hat{\boldsymbol{\beta}}$ as $[1, \tilde{\beta}_2, \tilde{\beta}_2]$, where $\tilde{\beta}_2$ is the OLS estimate of equation (3.26).

Nevertheless, equation (3.25) is useful for proving an important result. Let's define the vector λ as

$$\lambda = \left[R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - \mathbf{d});$$

by premultiplying (3.25) by $\mathbf{X}$ we get:

$$\mathbf{X}\tilde{\boldsymbol{\beta}} = \tilde{\mathbf{y}} = \hat{\mathbf{y}} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'\lambda$$

which in turn implies

$$\tilde{\mathbf{e}} = \hat{\mathbf{e}} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}R'\lambda$$

By using $\hat{\mathbf{e}} = \mathbf{M}_{\mathbf{X}}\mathbf{y} \Longrightarrow \hat{\mathbf{e}}'\mathbf{X} = 0$, we can use the above equation to get

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \hat{\mathbf{e}}'\hat{\mathbf{e}} + \lambda' R(\mathbf{X}'\mathbf{X})^{-1}R'\lambda$$

but by the definition of λ,

$$\lambda' R(\mathbf{X}'\mathbf{X})^{-1}R'\lambda = (R\hat{\boldsymbol{\beta}} - \mathbf{d})'\left[R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - \mathbf{d})$$

so finally

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}} = (R\hat{\boldsymbol{\beta}} - \mathbf{d})'\left[R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - \mathbf{d}). \tag{3.27}$$

Note that the right-hand side of equation (3.27) is very similar to (3.18). In fact, if our estimator for σ^2 is $\hat{\sigma}^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}/n$, we can combine equations (3.18), (3.24) and (3.27) to write the W statistic as:

$$W = \frac{(R\hat{\boldsymbol{\beta}} - \mathbf{d})'\left[R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - \mathbf{d})}{\hat{\sigma}^2} = n\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}}{\hat{\mathbf{e}}'\hat{\mathbf{e}}}. \tag{3.28}$$

Therefore, we can compute the same number in two different ways: one implies a rather boring sequence of matrix operations, using only ingredients that are available after the estimation of the unrestricted model. The other one, instead, requires estimating both models, but at that point 3 scalars (the SSRs and the number of observations) are sufficient for computing W.

3.5.1 Two alternative test statistics

There are two other statistics that can be used to perform a test on H_0 instead of the W test. One is the so-called F test, which is the traditional statistic taught in all elementary econometric courses, treated with reverence in all introductory econometrics books and the one that all software packages report. It can be written as

$$F = \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}}{s^2}\frac{1}{p} = \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}}{\hat{\mathbf{e}}'\hat{\mathbf{e}}} \cdot \frac{n-k}{p}; \tag{3.29}$$

it can be easily seen that there are two differences between F and W: F uses s^2, the unbiased estimator of σ^2 that I showed you at the end of section 3.2.2, instead of $\hat{\sigma}^2$; moreover, you also have the number of restrictions p in the denominator. Of course, it's easy to compute them from one another:

$$W = p \cdot F\frac{n}{n-k} \quad \Longleftrightarrow \quad F = \frac{n-k}{n}(W/p)$$

so in the standard case, when n is much larger than k, you have that $W \simeq p \cdot F$. Since their p-values are always practically the same, there is no statistical ground for preferring either. The reason why the econometricians of yore were attached to the F test was because its distribution is known even for small samples if ε_i is normal, so you don't need asymptotics. In my opinion, however, small samples are something you should steer clear of anyway, and postulating normality of ε_i is, as a rule, just wishful thinking. So, my advice is: use W.

The other statistic we can use is more interesting: if H_0 is true, the restricted model is just as correct as the unrestricted one. Therefore, one could conceivably estimate σ^2 by using $\tilde{\mathbf{e}}$ instead of $\hat{\mathbf{e}}$:

$$\tilde{\sigma}^2 = \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{n} \tag{3.30}$$

It can be proven that intuition is right, and if H_0 is true $\tilde{\sigma}^2$ is indeed consistent for σ^2. If we use $\tilde{\sigma}^2$ instead of $\hat{\sigma}^2$ in equation (3.28), we obtain the so-called **LM statistic**:[23]

$$LM = n\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}}{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}} = n\frac{\tilde{\mathbf{e}}'\mathbf{P}_{\mathbf{X}}\tilde{\mathbf{e}}}{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}, \tag{3.31}$$

where equality comes from (3.24). Since $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}$ cannot be less than $\hat{\mathbf{e}}'\hat{\mathbf{e}}$, in finite samples LM will always be smaller than W. However, under H_0 they tend to the same probability limit, and therefore under the null LM will also by asymptotically distributed as χ^2_p.[24]

The nice feature of the LM statistic is that it can be computed via a neat trick, known as **auxiliary regression**:

[23] The reason for the name is that this test statistic can be shown to be a "Lagrange Multiplier" test if normality of ε_i was assumed. Its validity, however, does not depend on this assumption. A fuller discussion of this point would imply showing that OLS is a maximum likelihood estimator under normality, which is something I'm not willing to do. See also footnote 19 in Section 3.4.2.

[24] The reader may want to verify that alternative formulations of the W and LM statistics are possible using $\hat{\sigma}^2$ and $\tilde{\sigma}^2$, or the R^2 indices from the two models.

1. run OLS on the constrained model and compute the residuals $\tilde{\mathbf{e}}$;
2. run OLS on a model where the dependent variable is $\tilde{\mathbf{e}}$ and the regressors are the same as in the unconstrained model;
3. take R^2 from this regression and multiply it by n. What you get is the *LM* statistic.

The last step is motivated by the fact that you can write R^2 as $\frac{\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y}}{\mathbf{y}'\mathbf{y}}$, so in the present case $\frac{\tilde{\mathbf{e}}'\mathbf{P}_\mathbf{X}\tilde{\mathbf{e}}}{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}$ is R^2 from the auxiliary regression.

Example 3.1

As an example, let's go back to the house pricing model we used as an example in Section 3.4. In Section 3.4.1 we already discussed two hypotheses of interest, namely:

- *The price elasticity is 1, and*
- *the number of baths has no effect.*

Testing for these two hypotheses separately is easy, via t-tests, which is just what we did a few pages back. As for the joint hypothesis, the easiest thing to do is setting up the resticted model as follows: combine equation (3.19) with $\beta_1 = 1$ and $\beta_2 = 0$. The restricted model becomes

$$p_i - s_i = \beta_0 + \beta_3 a_i + \beta_4 x_i + \varepsilon_i. \tag{3.32}$$

Note the redefinition of the dependent variable: if p_i is the log of the house price and s_i is its size in square feet, then $p_i - s_i$ is the log of is price per square foot, or unit price if you prefer. In fact, the hypothesis $\beta_1 = 1$ implies that if you have two houses (A and B) that are identical on all counts, except that A is twice as big as B, then the price of A should be twice that for B. Therefore, this hypothesis says implicitly that you can take into account appropriately the size of the property simply by focusing on its price per square foot, which is what we do in model (3.32). Estimating (3.32) via OLS gives

```
Dependent variable: lup

            coefficient   std. error    t-ratio    p-value
  ------------------------------------------------------------
  const       8.94603     0.00833147   1074       0.0000   ***
  age        -0.00247396  0.000232353   -10.65    6.07e-26 ***
  pool        0.115810    0.0221914       5.219   1.94e-07 ***

Mean dependent var   8.881042   S.D. dependent var   0.253030
Sum squared resid    158.1204   S.E. of regression   0.246277
R-squared            0.053395   Adjusted R-squared   0.052669
```

Superficially, one may think that our restricted model is much worse than the unrestricted one, as the R^2 drops from 68.5% to 5.3%. However, this is not a fair comparison, because in the restricted model the dependent variable is redefined and the denominator of the two R^2 indices is not the same. The SSRs are, instead, perfectly comparable, and the change you have between the full model and the unrestricted one is 157.88 → 158.12, which looks far less impressive, so we are drawn to think that the restricted model is not much worse in terms of fit. We could take this as an indication that our maintained hypothesis is not particularly at odds with the data.

This argument can be made rigorous by computing the W statistic:

$$W = 2610 \cdot \frac{158.1204 - 157.8844}{157.8844} = 3.9014$$

you get a statistic that is smaller than the critical value at 5% of $\chi^2_2 = 5.99$, so we accept both hypotheses again (the p-value is about 0.124). This time, however, the test was performed on the joint *hypothesis. It may well happen (examples are not hard to construct) that you may accept two hypotheses separately but reject them jointly (the converse should never happen, though).*

The LM test, instead, can be computed via an auxiliary regression as follows: take the residuals form model (3.32) and regress them against the explanatory variables of the unrestricted model (3.19). In this case, you get

```
            coefficient    std. error    t-ratio   p-value
  ------------------------------------------------------------
  const      -0.0924377     0.0483007     -1.914    0.0558  *
  lsize       0.0369564     0.0232121      1.592    0.1115
  baths      -0.00515142    0.0130688     -0.3942   0.6935
  age         8.72115e-05   0.000270997    0.3218   0.7476
  pool       -0.00901706    0.0226757     -0.3977   0.6909

SSR = 157.884, R-squared = 0.001493
```

The auxiliary regression per se *is not particularly interesting: its parameters don't have any meaningful interpretation.*[25]*. For us, it's just a computational device we use to compute the LM test statistic: take $R^2 = 0.001493$ and multiply it by the number of observations (2610); you get*

$$LM = 0.001493 \times 2610 = 3.89558,$$

which is practically identical to W, hence the conclusion is the same. ⌟

3.6 Exogeneity and causal effects

This section is short, but very important: The 2021 Nobel Prize for Economics was awarded to David Card, Joshua Angrist and Guido Imbens, precisely for

[25] For curiosity: the SSR is the same as in the restricted model. Can you prove why analytically? It's not difficult.

their work on causal effects, which has been enormously influential, especially in labour economics. The issue here is not about the statistical properties of $\hat{\boldsymbol{\beta}}$, but rather on its interpretation as an estimator of $\boldsymbol{\beta}$, so it fits in well at this point of the book, although the point we pursue here will be discussed in much greater detail in Chapter 6.

What does $\boldsymbol{\beta}$ measure? If $\mathrm{E}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, then $\boldsymbol{\beta}$ is simply defined as

$$\boldsymbol{\beta} = \frac{\partial \mathrm{E}[y|\mathbf{x}]}{\partial \mathbf{x}};$$

therefore, in a wage equation, the coefficient for education tells us simply by how much, on average, wage varies across different education levels.

It would be tempting to say "by how much your wage would change if you got one extra year of education", but unfortunately this statement would be unwarranted.[26] There are many reasons why the conditional mean may not be a good indicator of causality: for example, people may just stop attending school, or university, the moment they are able to earn a decent wage. If that were the case, the regression function of wage with respect to education would be flat, if not negative (because the smartest people would spend a shorter time in education). But this wouldn't mean that education has a negative effect on wages. In fact, quite the contrary: people who get more education would do so to compensate for their lesser ability. Of course, this example is a bit of a stretch, but should give you a hint as to why inferring causality from a regression coefficient may be a *very* bad idea.

More in general, there are situations when the causal relationship between y and $\mathbf{x}$ works in such a way that the conditional mean does not capture causality, but only the *outcome* of the process, which can be quite different, as in the example I just made.

In these cases, the traditional phrase that we use in the economics community is "$\mathbf{x}$ is **endogenous**" (as opposed to **exogenous**). If regressors are endogenous, then the regression parameters have nothing to do with causal effects; put another way, the parameters of interest $\boldsymbol{\beta}$ are not those that describe the conditional mean, and therefore, if you define ε_i as $y_i - \mathbf{x}_i'\boldsymbol{\beta}$, the first consequence is that the property $\mathrm{E}[\varepsilon_i|\mathbf{x}_i] = 0$ doesn't hold anymore, and so $\mathrm{E}[\varepsilon_i\mathbf{x}_i] = \mathrm{Cov}[\varepsilon_i, \mathbf{x}_i] \neq \mathbf{0}$. This is why in many cursory treatment of the subject, endogeneity is described as an "illness" that happens when the regressors are correlated with the disturbance term. Of course, that is an oversimplification: a more rigorous statement would be that in some cases the causal effects can be different from the conditional mean; if you define the disturbances as deviations from the causal effects, non-zero correlation between regressors and disturbances follows by construction.

[26] The ongoing debate in contemporary econometrics on the issue of differentiating between correlation and causation is truly massive. For a quick account, read chapter 3 in the latest best-seller in econometrics, that is Angrist and Pischke (2008), or simply google for "Exogeneity".

Note that this problem is not a shortcoming of OLS *per se*: the job of OLS is to estimate consistently the parameters of the conditional expectation $\frac{\partial \mathrm{E}[y|\mathbf{x}]}{\partial \mathbf{x}}$. If the nature of the problem is such that our parameters of interest $\boldsymbol{\beta}$ are a different object and we insist on equating them with what OLS returns (thereby giving OLS a misleading interpretation), it's a hermeneutical problem, not a statistical one.

The preferred tool in the econometric tradition for estimating causal effects is an estimator called **Instrumental Variable** estimator (or IV for short), but you'll have to wait until chapter 6 for it.

3.7 Prediction

Once a model is estimated and we have CAN estimators of $\boldsymbol{\beta}$ and σ^2, we may want to answer to the following question: if we were given a new datapoint for which the vector of covariates is known and equal to $\check{\mathbf{x}}$, what could we say about the value of the dependent variable $\check{y}$ for that new observation?

In order to give a sensible answer, let's begin by noting a few obvious facts: of course, y is a random variable, so we cannot predict it exactly. If we knew the true DGP parameters $\boldsymbol{\beta}$ and σ^2 we could say, however, that

$$\mathrm{E}\left[\check{y}|\check{\mathbf{x}}\right] = \check{\mathbf{x}}'\boldsymbol{\beta} \qquad \mathrm{V}\left[\check{y}|\check{\mathbf{x}}\right] = \sigma^2.$$

With this in hand, we could even build a confidence interval.[27] We have two ways we can follow here:

1. make no assumption on the distribution of ε. In this case, we need to use tools such as Chebyshev's inequality (see Section 2.A.2); this would be commendable, but is very rarely done.

2. Make some claim on the distribution of ε: everyone's favourite is the normal distribution, which leads to

$$P(|\check{y} - m| < 1.96\sigma) \simeq 95\% \tag{3.33}$$

   where $m = \check{\mathbf{x}}'\boldsymbol{\beta}$.

Unfortunately, we don't observe $\boldsymbol{\beta}$; instead, we observe $\hat{\boldsymbol{\beta}}$, so, assuming that the best way to make a point prediction of a random variable is to take its expectation,[28] the best we can do to predict $\mathbf{y}$ is computing

$$\hat{y} = \check{\mathbf{x}}'\hat{\boldsymbol{\beta}}.$$

[27] If you need to refresh the notion of confidence interval, go back to the end of section 2.3.2.

[28] This may seem obvious, but actually isn't: this choice is optimal if the loss function we employ for evaluating prediction is quadratic (see section 1.2). If the loss function was linear, for example, we'd have to use the median. But let's just stick to what everybody does.

Note, however, that $\hat{\boldsymbol{\beta}}$ is a random variable, with its own variance, so the confidence interval around $\hat{y}$ has to take this into account. Formally, let us define the prediction error as

$$e^* \quad = \check{y} - \hat{y} = (\check{\mathbf{x}}'\boldsymbol{\beta} + \check{\varepsilon}) - \check{\mathbf{x}}'\hat{\boldsymbol{\beta}} = \quad \check{\varepsilon} + \check{\mathbf{x}}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}});$$

the expression above reveals that our prediction can be wrong for two reasons: (a) because $\check{\varepsilon}$ is inherently unpredictable: our model does not contain all the feaures that describe the dependent variable and its variance is a measure of how bad our model is and (b) our sample is not infinite, and therefore we don't observe the DGP parameter $\boldsymbol{\beta}$, but only its estimate $\hat{\boldsymbol{\beta}}$.

If $\check{\varepsilon}$ can be assumed independent of $\hat{\boldsymbol{\beta}}$ (as is normally safe to do), then the variance of the difference is the sum of the variances:

$$\mathrm{V}\left[e^*\right] = \mathrm{V}\left[(\check{y} - \hat{y})|\check{\mathbf{x}}\right] = \mathrm{V}\left[\check{\varepsilon}\right] + \mathrm{V}\left[\check{\mathbf{x}}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right] = \sigma^2 + \check{\mathbf{x}}' V \check{\mathbf{x}}.$$

Of course, when computing this quantity with real data, we replace variances with their estimates, so we use $\hat{\sigma}^2$ (or s^2) in place of σ^2, and $\hat{V}$ for V.

Example 3.2

Suppose we use the model shown in section 3.4.1 to predict the price for a house built 5 years ago, with 1500 square feet of living area, 2 baths and no pool. In this case,

$$\check{\mathbf{x}}' = [1 \quad 2.708 \quad 2 \quad 5 \quad 0];$$

(the number 2.708 is just $\log(1500/100)$*; since*

$$\hat{\boldsymbol{\beta}}' = [8.8536 \quad 1.037 \quad -0.00515 \quad -0.00239 \quad 0.107]$$

simple multiplication yields $\hat{y} = 11.6395$*. Therefore the predicted price is about US$ 113491 (that is,* $\exp(11.6395)$*).*[29]

As for the variance,[30] *we need* $\hat{V}$*, that is* $s^2(\mathbf{X}'\mathbf{X})^{-1}$*:*

$$\hat{V} = 0.0001 \times \begin{bmatrix} 23.330 & -9.914 & 2.288 & -0.025 & 1.724 \\ -9.914 & 5.388 & -2.245 & -0.010 & -0.705 \\ 2.288 & -2.245 & 1.708 & 0.016 & -0.021 \\ -0.025 & -0.010 & 0.016 & 0.001 & -0.001 \\ 1.724 & -0.705 & -0.021 & -0.001 & 5.142 \end{bmatrix}$$

It turns out that

$$\mathrm{V}\left[\hat{y}|\mathbf{x}\right] = \hat{\sigma}^2 + \check{\mathbf{x}}'\hat{V}\check{\mathbf{x}} = 0.0606082 + 6.28845 \cdot 10^{-05} = 0.0606711;$$

[29] Actually, the expectation of the exponential is not the exponential of the expectation, since the exponential function is everywhere convex (see Section 2.A.1), but details are not important here.

[30] Of course, we could have used the asymptotic version $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ and very little would have changed.

since $\sqrt{0.0606711} \simeq 0.2463$*, by assuming normality we can even calculate a 95% confidence interval around our prediction as*

$$\hat{y} \pm 1.96\sqrt{\mathrm{V}\left[\hat{y}\right]} = 11.6395 \pm 1.96 \times 0.2463$$

so we could expect that, with a probability of 0.95, the log price of our hypothetical house would be between 11.157 and 12.122, and therefore the price itself between $ 70000 and $ 180000 (roughly). You may feel unimpressed by such a wide range, and I wouldn't disagree. But on the other hand, consider that this is a very basic model, which only takes into account very few features of the property, so it would be foolish to expect it to be razor-sharp when it comes to prediction.

One last thing: you may have noticed, in the example above, that the variance of the predictor depends almost entirely on the "model uncertainty" component $\hat{\sigma}^2$ and very little on the "parameter uncertainty" component $\check{\mathbf{x}}'\hat{V}\check{\mathbf{x}}$. This is not surprising, in the light of the fact that, as $n \to \infty$, the latter component should vanish, since $\hat{\boldsymbol{\beta}}$ is consistent. Therefore, in many settings (notably, in time-series models, that we'll deal with in Chapter 5), the uncertainty on the prediction is tacitly assumed to come only from σ^2.

3.8 The so-called "omitted-variable bias"

In many econometrics textbooks, you can find an argument that goes more or less like this: assume that the true model is

$$y_i = x_i\beta_1 + z_i\beta_2 + \varepsilon_i; \tag{3.34}$$

if you try to estimate β_1 via a regression of y_i on x_i alone, you're going to end up with a bad estimate. The proof is rather easy:[31]

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i\left(x_i\beta_1 + z_i\beta_2 + \varepsilon_i\right)}{\sum_{i=1}^n x_i^2} = \\ &= \beta_1 + \beta_2\frac{\sum_{i=1}^n x_i z_i}{\sum_{i=1}^n x_i^2} + \frac{\sum_{i=1}^n x_i\varepsilon_i}{\sum_{i=1}^n x_i^2} \end{aligned} \tag{3.35}$$

If you take probability limits, you'll find that, even if $\mathrm{E}\left[\varepsilon_i|x_i\right] = 0$,

$$\hat{\beta}_1 \xrightarrow{\mathrm{p}} \beta_1 + \beta_2\frac{\mathrm{E}\left[x_i z_i\right]}{\mathrm{E}\left[x_i^2\right]}$$

and the estimator would be inconsistent unless $\beta_2 = 0$ and/or $\mathrm{E}\left[x_i z_i\right] = 0$. The solemn maxim the student receives at this point is "if you omit a relevant regressor (z_i in our case, that is relevant if $\beta_2 \neq 0$) then your estimates will be incorrect,

[31] I could extend this example to matrices, but it would be totally unnecessary to grasp my point.

unless the omitted variable miraculously happens to be uncorrelated with x_i". The phenomenon is usually called **omitted variable bias**.

I've always considered this remark quite useless, if not outright misleading, and seeing econometricians, who are far better than myself, repeating it over and over to generations of students is a constant source of wonder to me. I'll try to illustrate why, and convince you of my point.

The parameter β_1 in (3.34) is defined as the partial effect of x_i on the conditional mean of y_i *on both* x_i *and* z_i; that is, the effect of x on y *given* z. It would be silly to think that this quantity could be estimated consistently without using any information on z_i.[32] The statistic $\hat{\beta}_1$, as defined in (3.35) (which does ignore z_i), is nevertheless a consistent estimator of a different quantity, namely the partial effect of x_i on the conditional mean of y_i *on* x_i *alone*, that is

$$\mathrm{E}\left[y_i|x_i\right] = \beta_1 x_i + \beta_2 \mathrm{E}\left[z_i|x_i\right].$$

The objection that some put forward, at this point, is: "OK; but assume that equation (3.34) is my object of interest, and z_i is unobservable, or unavailable. Surely, you must be aware that the estimate you get by using x_i only is bogus." Granted. But then, I may reply, do you *ever* get a real-life case when you observe *all* the variables you would like to have in your conditioning set? I don't think so; take for example the model presented in section 3.4: in order to set up a truly complete model, you would have to have data on the state of the building, on the quality of life in the neighbourhood, on the pleasantness of the view, and so on. You should always keep in mind that the parameters of your model only make sense in the context of the observable explanatory variable that you use for conditioning.[33]

This doesn't mean that you should not worry about omitted variable bias at all. The message to remember is: the quantity we would like to measure (ideally) is "the effect of x on y *all else being equal*"; but what we measure by OLS is the effect of x on y *conditional on* z. Clearly, in order to interpret our estimate the way we would like to, z should be as close to "all else" as possible, and if you omit relevant factors from your analysis (by choice, or impossibility) you have to be extra careful in interpreting your results.

Example 3.3

I downloaded some data from the World Development Indicators[34] *website. The variables I'm using for this example are*

[32] I should add that if we had an observable variable w_i, which we knew for certain to be uncorrelated with z_i, you could estimate β_1 consistently via a technique called *instrumental variable* estimation, which is the object of Chapter 6.

[33] In fact, there is an interesting link the bias you get from variable omission and the one you get from endogeneity (see section 3.6). Maybe I'll write it down at some point.

[34] The World Development Indicator (or WDI for short) is a wonderful database, maintained by the World Bank, that collects a wide variety of variables for over 200 countries over a large time span. It is one of the most widely used resources in development economics and is publicly available at http://wdi.worldbank.org or through DBnomics https://db.nomics.world/.

Variable name	***Description***
`NY.GDP.PCAP.PP.KD`	*GDP per capita based on purchasing power parity (PPP).*
`SH.MED.BEDS.ZS`	*Hospital beds (per 1,000 people)*
`NV.AGR.TOTL.ZS`	*Agriculture, value added (% of GDP)*

For each country, I computed the logarithm of the average (between 2014 and 2018) of the available data, which left me with data for 69 countries. The three resulting variables are called `l_gdp`*,* `l_hbeds` *and* `l_agri`*. Now consider Table 3.2, which reports two OLS regressions. In the first one, we regress the number of hospital beds on the share of GDP from agriculture. As you can see, the parameter is negative and significant. However, when we add GDP to the equation, the coefficient of* `l_agri` *becomes insignificant (and besides, its sign changes). On the contrary, you find that GDP matters a lot.*

Dependent variable: l_hbeds

	(1)	(2)
const	1.090**	−4.876**
	(0.1168)	(1.231)
l_agri	−0.2916**	0.08467
	(0.05794)	(0.09217)
l_gdp		0.5655**
		(0.1163)
$\bar{R}^2$	0.2635	0.4496

(standard error in parentheses)

Table 3.2: Two regressions on WDI data

The correct interpretation for this result is: there is a significant link between medical quality (as measured by the number of hospital beds per 1000 inhabitants) and the share of GDP from agriculture. In other words, if you travel to a country where everybody works in the fields, you'd better not get ill. However, this fact is simply a by-product of differences between countries in terms of economic development.

Once you consider the conditional expectation of `l_hbeds` *on a wider information set, which includes GDP per capita,*[35] *the effect disappears. That is,* for a given level of economic development[36] *there is no visible link between hospital beds and agriculture. To put it more explicitly: if you compare two countries*

[35] In the applied economic jargon: "once you control for GDP".

[36] OK, GDP per capita is not a *perfect* measure of economic development, nor of happiness, nor of well-being. I know. I know about the Human Development Index. I know about that Latouche guy. I know about all these things. Just give me a break, will you?

where the agricultural sectors have a different size (say, Singapore and Burundi), you're likely to find differences in their health system quality. However, if you compare two countries with the same per capita GDP *(say, Croatia vs Greece, or Vietnam vs Bolivia) you shouldn't expect to find any association between agriculture and hospital beds.*

Does this mean that model (1) is "wrong"? No: it simply means that the two coefficients in the two models measure two different things: a "gross" effect in equation (1) and a "net" effect in equation (2).[37] *Does this mean that model (2) is preferable? Yes: model (2) gives you a richer picture (see how much larger $\bar{R}^2$ is) because it's based on a larger information set.*

3.A Assorted results

3.A.1 Consistency of $\hat{\sigma}^2$

From (3.3), $\mathbf{e} = \mathbf{M}_\mathbf{X}\mathbf{y} = \mathbf{M}_\mathbf{X}\varepsilon$. Therefore, the sum of squared residuals can be written as

$$\mathbf{e}'\mathbf{e} = \varepsilon'\mathbf{M}_\mathbf{X}\varepsilon = \varepsilon'\varepsilon - \varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon;$$

now, given the definition of $\hat{\sigma}^2$,

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n},$$

divide everything by n and take probability limits; the first bit is easy:

$$\frac{\varepsilon'\varepsilon}{n} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \xrightarrow{\text{p}} \text{E}\left[\varepsilon_i^2\right] = \sigma^2.$$

On the other hand, equations (3.7) and (3.8) say that

$$\frac{\varepsilon'\mathbf{X}}{n} \xrightarrow{\text{p}} \mathbf{0} \qquad \frac{\mathbf{X}'\mathbf{X}}{n} \xrightarrow{\text{p}} Q$$

and therefore

$$\hat{\sigma}^2 = \frac{\varepsilon'\mathbf{M}_\mathbf{X}\varepsilon}{n} = \frac{\varepsilon'\varepsilon}{n} - \frac{\varepsilon'\mathbf{X}}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'\varepsilon}{n} \xrightarrow{\text{p}} \sigma^2 - \mathbf{0}'Q^{-1}\cdot\mathbf{0} = \sigma^2.$$

3.A.2 The classical assumptions

The classical assumptions were used in the infancy of econometrics to justify OLS as an inferential method. They reflect a point of view that was quite natural in those days, that is the idea that statistical methods could be borrowed from

[37] The discerning reader will doubtlessly spot the parallel with the discussion we had on the Frisch-Waugh theorem in section 1.4.4.

other sciences and be employed on economic data with little or no modification. The starting point is the linear model in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

which of course implies $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$. The classical assumptions are:

1. $\mathbf{X}$ is a $n \times k$ non-stochastic matrix, with $n > k$ and $\text{rk}(\mathbf{X}) = k$;
2. $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

In this context, $\mathbf{x}_i'\boldsymbol{\beta}$ is interpreted as a "law of nature", which describes what happens to y_i under certain conditions, described by $\mathbf{x}_i$; this idea is borrowed directly from experimental science, where y_i is the outcome of the i-th experiment and $\mathbf{x}_i$ are the conditions under which the i-th experiment took place.

The disturbance term ε_i is just "random noise", coming from experimental errors, bad measurement or some other factor that is impossible to control fully; the idea here is that if ε_i was 0, by observing y_i we would observe the "law of Nature" $\mathbf{x}_i'\boldsymbol{\beta}$ in its "uncontaminated" form.

In a controlled experiment, these hypotheses are perfectly natural: the $\mathbf{x}_i$ are obviously non-random (because are decided by the experimenter); to surmise that ε_i is Gaussian is also quite natural, since it is the outcome of a multitude of a large number of small imperfections and some faith in the Central Limit Theorem is not totally unjustified.

In the adaptation to economic data, the "fixed-$\mathbf{X}$" assumption was recognised as untenable, so a second version of the classical assumptions allows for the possibility that $\mathbf{X}$ may be random, In that case, assumption 2 is replaced by

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

which in turn implies $\text{E}[\varepsilon_i|\mathbf{X}] = \text{E}[\varepsilon_i|\mathbf{x}_i] = 0$, and as a consequence $\text{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$.

Finally, note that in the classical world $\boldsymbol{\varepsilon}$ is assumed to be Gaussian, while no assumption of that kind was made in Section 3.3. Normality is necessary to derive the distribution of hypothesis tests such as the t test or the F test when the sample size is small; needless to say, this is neither necessary nor desirable in modern econometrics, where datasets are almost always rather large and the normality assumption is, at best, questionable. This is why in contemporary econometrics (and, as a result, in this book) we mainly rely on asymptotic inference, where Gaussianity is nearly useless.

3.A.3 The Gauss-Markov theorem

The Gauss-Markov theorem states that, under homoskedasticity, OLS is the most efficient estimator among all those that are (a) unbiased and (b) linear. Unbiasedness means that $\text{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$; linearity means that $\hat{\boldsymbol{\beta}}$ can be written as $\hat{\boldsymbol{\beta}} = \mathbf{L}'\mathbf{y}$.

The OLS statistic enjoys both properties ($\mathbf{L}'$ being equal to $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for OLS), but other statistics may too. This property is often condensed in the phrase "OLS is BLUE", where BLUE stands for "Best Linear Unbiased Estimator".

The proof is simple if we concentrate on the case when $\mathbf{X}$ is a matrix of fixed constants and does not contain random variables, because in this case we can shuffle $\mathbf{X}$ in and out of the expectation operator $\mathrm{E}[\cdot]$ any way we want. Considering the case when $\mathbf{X}$ contains random variables makes the proof more involved.

Here goes: take a linear estimator $\tilde{\boldsymbol{\beta}}$ defined as $\tilde{\boldsymbol{\beta}} = \mathbf{L}'\mathbf{y}$. In order for it to be unbiased, the following must hold

$$\mathrm{E}\left[\tilde{\boldsymbol{\beta}}\right] = \mathrm{E}\left[\mathbf{L}'\left(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right)\right] = \mathbf{L}'\mathbf{X}\boldsymbol{\beta} + \mathbf{L}'\mathrm{E}\left[\boldsymbol{\varepsilon}\right] = \boldsymbol{\beta};$$

it is safe to assume that $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, so the unbiasedness requirement amounts to $\mathbf{L}'\mathbf{X} = I$. Note that, in the standard case, there are infinitely many matrices that satisfy this requirement, since $n > k$ and $\mathbf{X}$ is a "tall" matrix. In the OLS case, $\mathbf{L}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and the requirement is met. Therefore, under unbiasedness, $\tilde{\boldsymbol{\beta}}$ can be written as $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{L}'\boldsymbol{\varepsilon}$.

Now consider the variance of $\tilde{\boldsymbol{\beta}}$:

$$\mathrm{V}\left[\tilde{\boldsymbol{\beta}}\right] = \mathrm{E}\left[(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})'\right] = \mathrm{E}\left[\mathbf{L}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{L}\right] = \mathbf{L}'\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\mathbf{L} = \mathbf{L}\,\mathrm{V}\left[\boldsymbol{\varepsilon}\right]\,\mathbf{L}';$$

under homoskedasticity, $\mathrm{V}[\boldsymbol{\varepsilon}] = \sigma^2 I$, so

$$\mathrm{V}\left[\tilde{\boldsymbol{\beta}}\right] = \sigma^2\mathbf{L}'\mathbf{L}; \tag{3.36}$$

again, OLS is just a special case, so the variance of $\hat{\boldsymbol{\beta}}$ is easy to compute as $\mathrm{V}\left[\hat{\boldsymbol{\beta}}\right] = \sigma^2\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

The gist of the theorem lies in proving that the difference

$$\mathrm{V}\left[\tilde{\boldsymbol{\beta}}\right] - \mathrm{V}\left[\hat{\boldsymbol{\beta}}\right] = \sigma^2\mathbf{L}'\mathbf{L} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\left[\mathbf{L}'\mathbf{L} - (\mathbf{X}'\mathbf{X})^{-1}\right]$$

is positive semidefinite any time $\mathbf{L} \neq (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and therefore OLS is more efficient than $\tilde{\boldsymbol{\beta}}$. This is relatively easy: define $\mathbf{D} \equiv \mathbf{L}' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which has to be nonzero unless $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$. Therefore, $\mathbf{D}'\mathbf{D}$ must be positive semidefinite (see section 1.A.7).

$$\mathbf{D}'\mathbf{D} = \left[\mathbf{L}' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\left[\mathbf{L} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] = \mathbf{L}'\mathbf{L} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{L} - \mathbf{L}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1};$$

under unbiasedness, $\mathbf{L}'\mathbf{X} = \mathbf{X}'\mathbf{L} = I$, so the expression above becomes

$$\mathbf{D}'\mathbf{D} = \mathbf{L}'\mathbf{L} - (\mathbf{X}'\mathbf{X})^{-1} \tag{3.37}$$

and the claim is proven.

Having said this, let me add that the relevance of the Gauss-Markov theorem in modern econometrics is quite limited: the assumption that $\mathbf{X}$ is a fixed matrix makes sense in the context of a randomised experiment, but the data we have in economics are rarely compatible with this idea; the same goes, perhaps even

more strongly, for homoskedasticity. Moreover, one does not see why the linearity requirement should be important, aside from computational convenience; and a similar remark holds for unbiasedness, that is nice to have but not really important if our dataset is of a decent size and we can rely on consistency.

One may see why people insisted so much on the Gauss-Markov theorem in the early days of econometrics, when samples were small, computers were rare and statistical methods were borrowed from other disciplines with very few adjustments. Nowadays, it's just a nice exercise in matrix algebra.

3.A.4 Cross-validation and leverage

Given our usual regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.38}$$

we may indulge in the following thought exercise: "We have n datapoints, and we use all the available information to compute all the OLS-related statistics. But what if we had only $n-1$? We could pretend that the value of y_n was unavailable. What happens if we compute $\hat{\boldsymbol{\beta}}$ only using the first $n-1$ datapoints? How different would it be from its full-sample equivalent? And if we used $\hat{\boldsymbol{\beta}}$ and $\mathbf{x}_n$ to predict y_n, what should we expect?". As we will see, pursuing this idea will lead us to developing useful tools for identifying influential observations and testing the specification of our model.

Suppose we have n observations but we leave the i-th one aside, and introduce the following convention: the "$(-i)$" index means "excluding the i-th observation"; hence, $\mathbf{X}_{(-i)}$ is a $(n-1)\times k$ matrix, equal to $\mathbf{X}$ with the i-th row dropped, and the same interpretation holds for $\mathbf{y}_{(-i)}$.[38] The reason why we may want to do this is to check what happens to our model if a certain observation had not been available. There are several insights we can gain from doing so.

In order to perform the necessary calculations, it is useful to consider a model where you add to $\mathbf{X}$ a dummy variable identifying the i-th observation, that is an additional column $\mathbf{d}$, containing all zeros save for the i-th row, that contains 1. In pratice, our model becomes

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + d_i\lambda + \varepsilon_i = \mathbf{w}_i'\boldsymbol{\gamma} + \varepsilon_i. \tag{3.39}$$

For example, if $i = n$, $\mathbf{d}$ would be a vector of zeros with one 1 at the bottom and in matrix form the model would look like this:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_{(-i)} \\ y_i \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} \mathbf{X}_{(-i)} & \mathbf{0} \\ \mathbf{x}_i' & 1 \end{bmatrix} \qquad \boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\beta} \\ \lambda \end{bmatrix}.$$

Clearly, our original model 3.38 is just the special case $\lambda = 0$. Here are a few

[38] Note: this is not standard notation. I adopted it just for this section.

results that will be useful later on:[39]

$$\begin{aligned}
\mathbf{X}'\mathbf{M}_{\mathbf{d}} &= \begin{bmatrix}\mathbf{X}'_{(-i)} & \mathbf{0}\end{bmatrix} \\
\mathbf{X}'\mathbf{M}_{\mathbf{d}}\mathbf{X} &= \mathbf{X}'_{(-i)}\mathbf{X}_{(-i)} = \sum_{j\neq i}\mathbf{x}_j\mathbf{x}'_j \\
\mathbf{X}'\mathbf{M}_{\mathbf{d}}\mathbf{y} &= \mathbf{X}'_{(-i)}\mathbf{y}_{(-i)} = \sum_{j\neq i}\mathbf{x}_j y_j \\
\mathbf{d}'\mathbf{M}_{\mathbf{X}}\mathbf{d} &= m_i \\
\mathbf{d}'\mathbf{M}_{\mathbf{X}}\mathbf{y} = \mathbf{d}'\tilde{\mathbf{e}} &= \tilde{e}_i
\end{aligned}$$

where $\tilde{\mathbf{e}}$ are the OLS residuals for the full-sample model, that is equation (3.38); m_i is the i-th element on the diagonal of $\mathbf{M}_{\mathbf{X}}$, that is $1-\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$. Let's also define $h_i = 1-m_i = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$, the i-th element on the diagonal of $\mathbf{P}_{\mathbf{X}}$. It can be proven that $0 \leq m_i \leq 1$, so that the same holds for h_i too.[40]

Some readers may find the choice of symbols surprising: if the m_i values are the diagonal of $\mathbf{M}_{\mathbf{X}}$, then it would have been natural to use p_i for the diagonal of $\mathbf{P}_{\mathbf{X}}$. The reason for using h_i instead comes from calling $\mathbf{P}_{\mathbf{X}}$ the "hat matrix" (see section 1.4.1).

The Frisch-Waugh theorem (see section 1.4.4) makes it easy to compute the OLS estimates for model (3.39):

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left[\mathbf{X}'\mathbf{M}_{\mathbf{d}}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{M}_{\mathbf{d}}\mathbf{y} = \left[\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)}\right]^{-1}\mathbf{X}'_{(-i)}\mathbf{y}_{(-i)} \\
\hat{\lambda} &= (\mathbf{d}'\mathbf{M}_{\mathbf{X}}\mathbf{d})^{-1}\mathbf{d}'\mathbf{M}_{\mathbf{X}}\mathbf{y} = \frac{\tilde{e}_i}{m_i}
\end{aligned}$$

The $\hat{\boldsymbol{\beta}}$ vector is nothing but the OLS statistic you would have found after dropping the i-th observation. The $\hat{\lambda}$ parameter is more interesting: let's begin by considering the residuals of (3.39): $\hat{\mathbf{e}} = \mathbf{M}_{\mathbf{W}}\mathbf{y}$. Its i-th element, $\hat{e}_i$, is defined as

$$\hat{e}_i = y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}} - \hat{\lambda}.$$

This quantity is identically 0; to see why, note that $\mathbf{d}$ is an extraction vector (see section 3.3.1), so you can write

$$\hat{e}_i = \mathbf{d}'\hat{\mathbf{e}} = \mathbf{d}'\mathbf{M}_{\mathbf{W}}\mathbf{y} = 0;$$

since $\mathbf{d} \in \text{Sp}(\mathbf{W})$, then $\mathbf{d}'\mathbf{M}_{\mathbf{W}} = \mathbf{0}'$, and therefore $\mathbf{d}'\hat{\mathbf{e}} = \hat{e}_i = 0$. By putting the two equations above together, you get

$$\hat{\lambda} = y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}}.$$

[39] These are easy to prove, and provide a nice exercise on matrix algebra. Hint: start by computing $\mathbf{d}'\mathbf{d}$ and $\mathbf{d}\mathbf{d}'$.

[40] The proof is surprisingly easy: m_i lies on the diagonal of $\mathbf{M}_{\mathbf{X}}$; since $\mathbf{M}_{\mathbf{X}}$ is positive semi-definite, the diagonal elements cannot be negative. But the same holds for h_i and $\mathbf{P}_{\mathbf{X}}$. Since $h_i + m_i = 1$, the proof is complete.

which, in turn, means that $\hat{\lambda}$ is the prediction error you get if you try to predict the i-th observation by using all the other ones. Or, put another way, if you want to compute what the prediction error for y_i would be (based on the remaining observations) all you have to do is stick an appropriate dummy into your model and take its coefficient. Note that in fact there is an even easier way: since $\hat{\lambda} = \frac{\tilde{e}_i}{m_i}$, you may just as well run OLS on the full-sample model (3.38), save its residuals and the m_i series, and divide one by the other.

The **cross-validation criterion** is a model selection tool that is based on just that: you simulate the out-of-sample performance of your model by adding the squares of the n prediction errors you find by omitting each observation in turn:

$$CV = \sum_{i=1}^{n} e_{(-i)}^2 = \sum_{i=1}^{n} \left(\frac{\tilde{e}_i}{m_i} \right)^2$$

When you compare two models (say, A and B), it may well happen that model A yields a smaller sum of squared residuals than B, but B outperforms A in terms of the cross-validation criterion. Usually, this happens when A has a richer structure than B (in the OLS context, more regressors); in these cases, the canonical interpretation is that A is only apparently a better model than B: some of the apparently significant regressors catch in fact spurious regularities than cannot be expected to hold in general. In these cases, the term we customarily use is **overfitting**.

Data scientists are inordinately fond of the cross-validation concept, and in many cases they use sophisticated variations of this idea to pick the best forecasting model for a given problem.

The variant of the cross-validation method I just illustrated, where you exclude one observation at a time, has a lot in common with an old and established statistical technique called **jackknifing**, which in turn is a close relative of boostrapping (see Section 4.A.4).

In statistical learning and similar disciplines, the approach presented here is often generalised by excluding entire subsets of the entire dataset instead of one observation only, possibly choosing them in very elaborate ways. They call this **folding**.

In the light of the discussion above, there is something interesting we can say on the interpretation of the magnitude m_i and its complement to 1, $h_i = 1 - m_i$: from the definition of $\hat{\mathbf{e}}$ we have

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{d}\hat{\lambda} + \hat{\mathbf{e}};$$

by premultiplying the above by $\mathbf{M}_\mathbf{X}$, you get

$$\mathbf{M}_\mathbf{X}\mathbf{y} = \tilde{\mathbf{e}} = \mathbf{M}_\mathbf{X}\mathbf{d}\hat{\lambda} + \hat{\mathbf{e}}.$$

Therefore, since $\mathbf{M}_\mathbf{X}\hat{\mathbf{e}} = \mathbf{0}$,

$$\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \mathbf{d}'\mathbf{M}_\mathbf{X}\mathbf{d}\lambda^2 + \hat{\mathbf{e}}'\hat{\mathbf{e}}$$

so, finally,

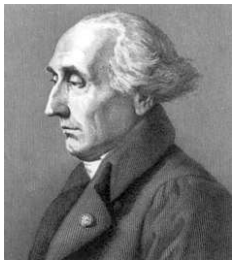
$$\tilde{e}_i^2 / m_i = \tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}.$$

Which means: if you compare the SSR for the two models you get by using the full sample or omitting the i-th observation, you get that their difference is always non-negative, and equals $\tilde{e}_i^2/m_i$. Clearly, if the difference is large, the results you get by adding/removing the i-th observation are dramatically different, so that data point deserves special attention.

The quantity $\tilde{e}_i^2/m_i$ may be large either because (a) $\tilde{e}_i$ is large in absolute value and/or (b) m_i is close to 0 (which implies that h_i is close to 1). This is why the h_i values are sometimes used as descriptive statistics to check for "influential observations", and are sometimes referred to as **leverage** values. Note that h_i only depends on the regressors $\mathbf{X}$, and not on $\mathbf{y}$. Therefore, large values of h_i indicate observations for which the combination of explanatory variables we have is uncommon enough to exert a substantial effect on the final estimates.

3.A.5 Derivation of RLS

As the reader doubtlessly already knows, the standard method for finding the extrema of a function subject to constraints is the so called "Lagrange multipliers method". For a full description of the method, the reader had better refer to one of the many existing texts of mathematics for economists[41], but here I'll give you a super-simplified account for your convenience.



JOSEPH LOUIS LAGRANGE

If you have to find maxima and/or minima of a function $f(\mathbf{x})$ subject to a system of constraints $g(\mathbf{x}) = \mathbf{0}$, you set up a function, called the **Lagrangean**, in which you sum the objective function to a linear combination of the constraints, like this:

$$\mathscr{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda' g(\mathbf{x}).$$

The elements of the vector λ are known as "Lagrange multipliers".

For example, the classic microeconomic problem of a utility-maximising consumer is represented as

$$\mathscr{L}(\mathbf{x}, \lambda) = U(\mathbf{x}) + \lambda \cdot \left(Y - \mathbf{p}'\mathbf{x}\right)$$

where $\mathbf{x}$ is the bundle of goods, $U(\cdot)$ is the utility function, Y is disposable income and $\mathbf{p}$ is the vector of prices. In this example, the only constraint you have is the budget constraint, so λ is a scalar.

The solution has to obey two conditions, known as the "first order conditions":

$$\frac{\partial \mathscr{L}}{\partial \mathbf{x}} = \mathbf{0} \qquad \frac{\partial \mathscr{L}}{\partial \lambda} = 0, \tag{3.40}$$

so in practice you differentiate the Lagrangean with respect to your variables and λ, and then check if there are any solutions to the system of equations you

[41] One I especially like is Dixit (1990), but for a nice introductory treatment I find Dadkhah (2011) hard to beat.

get by setting the partial derivatives to 0. If the solution is unique, you're all set. In the utility function example, applying equations (3.40) gives the standard microeconomic textbook solution to the problem:

$$\frac{\partial U(\mathbf{x})}{\partial \mathbf{x}} = \lambda \mathbf{p} \qquad Y = \mathbf{p}'\mathbf{x}.$$

in words: at the maximum, (a) marginal utilities are proportional to prices (or $\frac{\partial U}{\partial x_i} \Big/ \frac{\partial U}{\partial x_j} = \frac{p_i}{p_j}$ if you prefer) and (b) you should spend all your income.

In the case of RLS, the Lagrangean is[42]

$$\mathscr{L}(\boldsymbol{\beta}, \lambda) = \frac{1}{2}\mathbf{e}'\mathbf{e} + \lambda'(R\boldsymbol{\beta} - \mathbf{d}).$$

where of course $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. The derivative of $\mathscr{L}$ with respect to λ is just the constraint; as for the other one, since the derivative of $\mathbf{e}$ with respect to $\boldsymbol{\beta}$ is $-\mathbf{X}$, we can use the chain rule like in section 1.A.5, arranging all the products in an appropriate way so as to obtain column vectors, which gives

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\beta}} = -\mathbf{X}'\mathbf{e} + R'\lambda$$

so the first order condition with respect to $\boldsymbol{\beta}$ can be written as

$$\mathbf{X}'\tilde{\mathbf{e}} = R'\lambda, \tag{3.41}$$

where $\tilde{\mathbf{e}}$ is the vector that satisfies equation (3.41), defined as $\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$. By premultiplying (3.41) by $(\mathbf{X}'\mathbf{X})^{-1}$ we get

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}R'\lambda,$$

which of course implies

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}R'\lambda \tag{3.42}$$

So the constrained solution $\tilde{\boldsymbol{\beta}}$ can be expressed as the OLS vector $\hat{\boldsymbol{\beta}}$, plus a "correction factor", proportional to λ. If we premultiply (3.42) by R we get

$$\lambda = \left[R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - \mathbf{d}) \tag{3.43}$$

because $R\tilde{\boldsymbol{\beta}} = \mathbf{d}$ by construction. Interestingly, λ itself is proportional to $(R\hat{\boldsymbol{\beta}} - \mathbf{d})$, that is precisely the quantity we use for the construction of the W statistic in (3.18). Finally, equation (3.25) is obtained by combining (3.42) and (3.43):

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}R'\left[R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\boldsymbol{\beta}} - \mathbf{d}).$$

[42] Note that I divided the objective function by 2. Clearly, the solution is the same, but the algebra is somewhat simplified.

3.A.6 Asymptotic properties of the RLS estimator

Begin by (3.25)

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1} R' \left[R(\mathbf{X}'\mathbf{X})^{-1} R' \right]^{-1} (R\hat{\boldsymbol{\beta}} - \mathbf{d});$$

from this equation, it is quite easy to see that $\tilde{\boldsymbol{\beta}}$ is an affine function of $\hat{\boldsymbol{\beta}}$; this will be quite useful. Now define H as

$$H = \operatorname{plim}\left((\mathbf{X}'\mathbf{X})^{-1} R' \left[R(\mathbf{X}'\mathbf{X})^{-1} R' \right]^{-1} \right) = Q^{-1} R' \left[RQ^{-1} R' \right]^{-1} \tag{3.44}$$

normally, this is a $(k \times p)$ matrix with rank p; also note that HR is idempotent.

If $\hat{\boldsymbol{\beta}}$ is consistent, then

$$\tilde{\boldsymbol{\beta}} \xrightarrow{\text{p}} \boldsymbol{\beta} - H \cdot (R\boldsymbol{\beta} - \mathbf{d})$$

so if $R\boldsymbol{\beta} = \mathbf{d}$, consistency is guaranteed; otherwise, it isn't.

As for efficiency, let me briefly remind you what we mean by "efficiency":[43] an estimator a is more efficient than a competing estimator b if the difference $V(b) - V(a)$ is positive. If the two estimators are vectors, then the criterion generalises to the requirement that $V(\mathbf{b}) - V(\mathbf{a})$ is a positive semi-definite matrix (psd for short — see section 1.A.7, especially figure 1.4).

There are a few matrix algebra results that we are going to need here:[44]

1. if A is pd, then A^{-1} is pd too;
2. if A is psd, then $B'AB$ is also psd for any matrix B.

Using these, we will prove that the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is smaller (in a matrix sense) than that of $\hat{\boldsymbol{\beta}}$:

$$\operatorname{AV}\left[\tilde{\boldsymbol{\beta}}\right] = (I - HR) \cdot \operatorname{AV}\left[\hat{\boldsymbol{\beta}}\right] \cdot (I - R'H')$$

if $\operatorname{AV}\left[\tilde{\boldsymbol{\beta}}\right] = \sigma^2 Q^{-1}$, then note that

$$HRQ^{-1} = Q^{-1} R' \left[RQ^{-1} R' \right]^{-1} RQ^{-1}, \tag{3.45}$$

which is evidently symmetric, so $HRQ^{-1} = Q^{-1}R'H'$; furthermore, by using the fact that HR is idempotent, you get

$$HRQ^{-1} = HRHRQ^{-1} = HRQ^{-1}R'H'.$$

As a consequence, the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ can be written as

$$\begin{aligned} \operatorname{AV}\left[\tilde{\boldsymbol{\beta}}\right] &= \sigma^2 \left\{ Q^{-1} - HRQ^{-1} - Q^{-1}R'H' + HRQ^{-1}R'H' \right\} = \\ &= \sigma^2 \left[Q^{-1} - HRQ^{-1} \right] = \\ &= \operatorname{AV}\left[\hat{\boldsymbol{\beta}}\right] - \sigma^2 HRQ^{-1} \end{aligned}$$

[43] A slightly fuller discussion is at the end of section 2.3.2.

[44] They're both easy to prove; try!

the last thing we need do prove is that HRQ^{-1} is also positive semi-definite: for this, we'll use the right-hand side of (3.45).

Since Q is pd, then Q^{-1} is pd as well (property 1); therefore, $RQ^{-1}R'$ is also pd (property 2), and so is $\left[RQ^{-1}R'\right]^{-1}$ (property 1). Finally, by using property 1 again, we find that $Q^{-1}R'\left[RQ^{-1}R'\right]^{-1}RQ^{-1}$ is positive semi-definite and the result follows.

Chapter 4

Diagnostic testing in cross-sections

In order to justify the usage of OLS as an estimator, we made some assumptions in section 3.2. Roughly:

1. the data we observe are realisations of random variables such that it makes sense to assume that we are observing the same DGP in all the n cases in our dataset; or, more succinctly, there are no **structural breaks** in our dataset.

2. We can trust asymptotic results as a reliable guide to the distribution of our estimators, as the n observations we have are sufficiently homogeneous and sufficiently independent, so that the LLN and the CLT can be taken as valid;

3. the conditional expectation $\mathrm{E}\left[y|\mathbf{x}\right]$ exists and is linear: $\mathrm{E}\left[y|\mathbf{x}\right] = \mathbf{x}'\boldsymbol{\beta}$;

4. the conditional variance $\mathrm{V}\left[y|\mathbf{x}\right]$ exists and does not depend on $\mathbf{x}$ at all, so it's a positive constant: $\mathrm{V}\left[y|\mathbf{x}\right] = \sigma^2$.

Assumption number 2 may be inappropriate for two reasons: one is that our sample size is too small to justify asymptotic results as a reasonable approximation to the actual properties of our statistics; the other one is that our observation may not be identical, nor independent. The first case cannot really be tested formally; in most cases, the data we have are given and economists almost never enjoy the privileges of experimenters, who can have as many data points as they want (of course, given sufficient resources). Therefore, we just assume that our dataset is good enough for our purposes, and hope for the best. Certainly, intellectual honesty dictates that we should be quite wary of drawing conclusions on the basis of few data points, but there is not much more we can do. As for the second problem, we will defer the possible lack of independence to chapter 5, since the issue is most likely to arise with time-series data.

In the next section, we will consider a way of testing assumptions 1 (to some extent) and 3. If they fail, consistency of $\hat{\boldsymbol{\beta}}$ may be at risk. Conversely, assumption number 4 is crucial for our hypothesis testing apparatus, and will need some extra tools; this will be the object of section 4.2.

4.1 Diagnostics for the conditional mean

Our main tool in proving that $\hat{\boldsymbol{\beta}} \xrightarrow{\text{p}} \boldsymbol{\beta}$ was that $\mathrm{E}\left[y|\mathbf{x}\right] = \mathbf{x}'\boldsymbol{\beta}$ (see section 3.2.1). But this statement may be false. We will not explore the problem in its full generality: we'll just focus on two possible issues that often arise in practice.

1. the regression function is nonlinear, but can be approximated via a linear function (see the discussion in section 1.3.2). In the case of a scalar regressor,

$$\mathrm{E}\left[y_i|x_i\right] = m(x_i) \simeq \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots \beta_q x_i^q = \sum_{j=0}^{q} \beta_j x_i^j.$$

2. Our data comprise observations for which the DGP is partly different. That is, we have $j = 1 \ldots m$ separate sub-populations, for which

$$\mathrm{E}\left[y_i|\mathbf{x}_i\right] = \mathbf{x}_i'\boldsymbol{\beta}_j$$

where j is the class that observation i belongs to. For example, we have data on European and American firms, and the vector $\boldsymbol{\beta}$ is different on the two sides of the Atlantic (in this case, $m = 2$).

4.1.1 The RESET test

As I repeatedly said earlier, the hypothesis of linearity simply means that $\mathrm{E}\left[y_i|\mathbf{x}_i\right]$ can be written as a linear combination of observable variables. The short phrase we use is: the model has to be *linear in the parameters*, not necessarily *in the variables* (see Section 1.3.4 for a fuller discussion).

For example, suppose $\mathbf{x}_i$ is a scalar; it is perfectly possible to accommodate something like

$$\mathrm{E}\left[y_i|x_i\right] = \beta_1 x_i + \beta_2 x_i^2;$$

(to ease exposition, I am assuming here that the conditional mean has no constant term). Suppose that the expression above holds, but in the model we estimate the quadratic term x_i^2 is dropped. That is, we estimate a model like

$$y_i = \gamma x_i + u_i;$$

by OLS, so that we would obtain a statistic $\hat{\gamma}$ defined as

$$\hat{\gamma} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Clearly, there is no value of γ that can make u_i the difference between y_i and $\mathrm{E}[y_i|x_i]$, so we can't expect $\hat{\gamma}$ to have all the nice asymptotic properties that OLS has. In fact, it can be proven that (in standard cases) the statistic $\hat{\gamma}$ does have a limit in probability, but the number it tends to is neither β_1 nor a simple function of it, so technically there is no way we can use $\hat{\gamma}$ to estimate β_1 consistently.

The limit in probability of $\hat{\gamma}$ is technically known as a **pseudo-true value**, which is far too complex a concept for me to attempt an exposition here. The inquisitive reader may want to have a look at Cameron and Trivedi (2005), section 4.7 or (more technical) Gourieroux and Monfort (1995). The ultimate bible on this is White (1994).

In the present case, the remedy is elementary: add x_i^2 to the list of your regressors and, *voilà*, you get perfectly good CAN estimates of β_1 and β_2.[1] However, in a real-life case, where you have a vector of explanatory variables $\mathbf{x}_i$, things are not so simple. In order to have quadratic effects, you should include all possible cross-products between regressors. For example, a model like

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

would become

$$y_i = \beta_0 + \underbrace{\beta_1 x_i + \beta_2 z_i}_{\text{linear part}} + \underbrace{\beta_3 x_i^2 + \beta_4 x_i \cdot z_i + \beta_5 z_i^2}_{\text{quadratic part}} + \varepsilon_i$$

and it's very easy to show that the number of quadratic terms becomes rapidly unmanageable for a realistic model: if the original model has k regressors the quadratic one can have up to $\frac{k(k+1)}{2}$ additional terms.[2]. I don't think I have to warn the reader on how much of a headache it would be to incorporate cubic or quartic terms.

The **RESET test** (stands for **RE**gression **S**pecification **E**rror **T**est) is a way to check whether a given specification needs additional nonlinear effects or not. The intuition is simple and powerful: instead of augmenting our model with all the possible order 2 terms (squares and cross-products), we just use the square of the fitted values, that is instead of $x_i^2, x_i \cdot z_i$ and z_i^2 in the example above, we would use

$$\hat{y}_i^2 = \left(\hat{\beta}_1 x_i + \hat{\beta}_2 z_i\right)^2.$$

Clearly, a similar strategy could be extended to cubic terms; in the example above, we would replace the linear combination of $x_i^3, x_i^2 \cdot z_i, x_i \cdot z_i^2$ and z_i^3 with the simple scalar term $\hat{y}_i^3$. Then, we just check if the added terms are significant; since this is a test for addition of variables to a pre-existing model, the most convenient way to perform the test is by using the LM statistic (see section 3.5.1).

The procedure is then:

---

[1] Of course, you'd have to be careful in computing your marginal effects, but if you have read section 1.3.4, you know that, don't you?

[2] The reader is invited to work out what happens for various values of k.

1. Run OLS, save the residuals e_i and the fitted values $\hat{y}_i$;
2. generate squares and cubes $\hat{y}_i^2, \hat{y}_i^3$;
3. run the auxiliary regression

$$e_i = \gamma' \mathbf{x}_i + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3 + u_i;$$

4. compute $LM = n \cdot R^2 \overset{a}{\sim} \chi_2^2$

Example 4.1

Let us compute the RESET test to check the hedonic model used as an example in Section 3.4 for possible neglected nonlinearities; the auxiliary regression yields

```
Dependent variable: ehat

            coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const      195.825       81.6622      2.398    0.0166  **
  lsize       38.9198      17.1107      2.275    0.0230  **
  baths       -0.205739     0.0861412  -2.388    0.0170  **
  age         -0.0900385    0.0393760  -2.287    0.0223  **
  pool         3.97729      1.76245     2.257    0.0241  **
  yh2         -3.42319      1.40727    -2.433    0.0151  **
  yh3          0.103615     0.0399778   2.592    0.0096  ***

Mean dependent var   0.000000   S.D. dependent var   0.245999
Sum squared resid    152.9225   S.E. of regression   0.242381
R-squared            0.031428   Adjusted R-squared   0.029195
```

In this case, the LM statistic equals $n \cdot R^2 = 2610 \times 0.031 = 82.0263$*, which is much bigger than 5.99 (the 5% critical value for the* χ_2^2 *density); in fact, the p-value is a puny* $1.54243 \cdot 10^{-18}$*. Therefore, we reject the null and we conclude that the model has a specification problem.*

One final note: the usage of powers to model nonlinearity is widespread in applied econometrics, but it is by no means the only available choice. If you're interested, you may want to spend some time googling for "cubic splines" or "fractional polynomials", the former being hugely popular among data scientists; these techniques are useful for approximating arbitrary smooth function via linear combinations of observable variables, and are therefore perfectly suited for OLS estimation. If you're into even more exotic stuff, try "loess" or "Nadaraya-Watson".

4.1.2 Interactions and the Chow test

The problem of possible differences in the parameters between sub-samples is best illustrated in a simple setting: we have a scalar regressor x_i and two subpopulations. A dummy variable d_i tells us which group observation i belongs to. Suppose that the DGP could be described as follows:

$$\begin{array}{ll} \text{Subsample 1} & y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \text{Subsample 2} & y_i = \beta_0 + \beta_2 x_i + \varepsilon_i \end{array}$$

with $d_i = 0$ in subsample 1 and $d_i = 1$ in subsample 2. Note that the model could be rewritten as

$$\begin{aligned} y_i &= \beta_0 + \left[\beta_1 + (\beta_2 - \beta_1) d_i\right] x_i + \varepsilon_i = \\ &= \beta_0 + \beta_1 x_i + \gamma d_i \cdot x_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \gamma z_i + \varepsilon_i \end{aligned} \tag{4.1}$$

where $\gamma = \beta_2 - \beta_1$. Again, note that model (4.1) is perfectly fit for OLS estimation, since the product $z_i = d_i \cdot x_i$ is just another observable variable, which happens to be equal to x_i when $d_i = 1$ and 0 otherwise.

If the effect of x_i on y_i is in fact homogeneous across the two categories, then $\gamma = 0$; therefore, testing for the equality of β_1 and β_2 is easy: all you need to do is check whether the regressor z_i is significant. Explanatory variables of this kind, that you obtain by multiplying a regressor by a dummy variable, are often called **interactions** in the applied economics jargon. If the interaction term turns out to be significant, then the effect of x on y is different across the two subcategories, since the interaction term in your model measures how different the effect is across the two subgroups.

Clearly, you can interact as many regressors as you want: in the example above, you could also imagine that the intercept could be different across the two subpopulations as well, so the model would become something like

$$y_i = \beta_0 + \beta_1 x_i + \gamma_0 d_i + \gamma_1 d_i \cdot x_i + \varepsilon_i,$$

because interacting the constant by d_i just gives you d_i.

It should be noted that interactions can be viewed as including a peculiar form of nonlinearity, so you should keep this in mind when computing marginal effects. The marginal effect for x_i in equation (4.1), for example, would be

$$\frac{\partial \mathrm{E}\left[y_i | x_i\right]}{\partial x_i} = \beta + \gamma d_i,$$

that is, β if $d_i = 0$ and $\beta + \gamma$ if $d_i = 1$.

When you interact all the parameters by a dummy, then the test for equality of coefficients across the two subsamples is particularly simple, and amounts to what is known as the **Chow test**, since the SSR for the unrestricted model (that is, the one with all the interactions) is just the sum of the two separate regressions: if you have two subgroups, you can compute

1. the SSR for the OLS model on the whole sample (call it S_T);
2. the SSR for the OLS model using only the data in subsample 1 (call it S_1);
3. the SSR for the OLS model using only the data in subsample 2 (call it S_2).

Then, the Chow test is simply

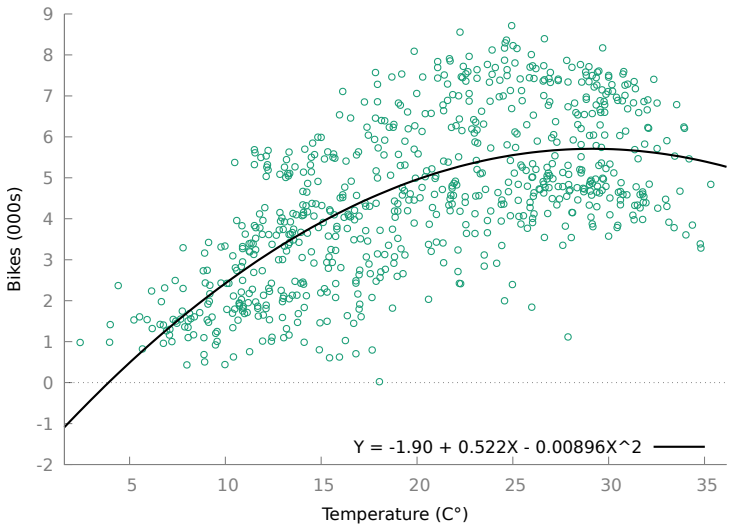
$$W = n \cdot \frac{S_T - (S_1 + S_2)}{S_1 + S_2} \tag{4.2}$$

because the SSR for the model with all the interactions is equal to the sum of the SSRs for the separate submodels. Of course, the appropriate number of degrees of freedom to use for the p-value would be k, the difference between the number of parameters in the unrestricted model ($k+k$) and those in the restricted one (k). The proof is contained in section 4.A.1, where I also generalise this idea to the case when you have more than 2 subsamples.

Example 4.2 (bike sharing and the weather)

Figure 4.1 comes from a dataset on bike sharing provided in Fanaee-T and Gama (2014) and depicts the relationship between the temperature on a given day (in Celsius, x_i in formulae) and the number of bikes rented (in thousands, y_i in formulae).

Figure 4.1: Relationship between temperature and bikes rented



Once you fit a quadratic model to the data, things appear to be basically OK: R^2 is not at all bad, and the estimated regression lines makes perfect sense, with the negative concavity indicating that most people want to ride a bike when the weather is warm, but not too hot.

Table 4.1: Various models for bike rentals

	Full sample	Sunny days	Cloudy/rainy	Full + interactions
const	-1.902^{***}	-2.659^{***}	-0.6698	-0.6698
	(-4.974)	(-6.281)	(-0.9444)	(-1.009)
x	0.5221^{***}	0.6563^{***}	0.3165^{***}	0.3165^{***}
	(12.70)	(14.54)	(4.005)	(4.279)
x^2	-0.008956^{***}	-0.01243^{***}	-0.003717^{*}	-0.003717^{*}
	(-8.895)	(-11.41)	(-1.835)	(-1.960)
d				-1.989^{**}
				(-2.494)
$d \cdot x$				0.3398^{***}
				(3.876)
$d \cdot x^2$				-0.008712^{***}
				(-3.940)
n	731	463	268	731
R^2	0.4532	0.5222	0.3965	0.5115
SSR	1498.035	779.7312	558.5052	1338.2364

Note: t-statistics in parenthesis.

However, we may surmise that what happens on sunny days may be different from rainy days. Fortunately, we also have the dummy variable d_i, which equals 1 if the weather on that day was sunny and 0 if it was cloudy or rainy. Splitting the sample in two gives the estimates in Table 4.1: the first column gives the estimates on the full sample (the same as in Figure 4.1). Column 2, instead, contains the estimates obtained using only the sunny days and column 3 only the ones for the bad weather days.

As you can see, the estimates for sunny days are numerically different from the ones for cloudy days. For example, the quadratic effect in column 3 seems to be much less significant than the one in column 2. However, the real question is: are they statistically *different? Or, in other words: is there a reason to believe that the relationship between the number of rented bikes and air temperature depends on the weather?*

In order to do so, we can run a Chow test. The mechanical way to do this would be adding to the base model all the interactions with the "sunny" dummy. The corresponding estimates are found in column 4 of Table 4.1. Note that the first three coefficients in column 4 are exactly equal to those in column 3,[3] *and that the coefficients in column 2 can be obtained by summing the correspon-*

[3]The standard errors are not: this is a side effect of the fact that model 3 and model 4 use different estimators for σ^2 and hence the two coefficient covariance matrices are different.

dent coefficients in column 4 to its "interacted" counterpart. For example, the coefficient for x_i reported in column 2 (0.6563) can be calculated as the two entries in column 4 for x and its interaction with d (0.3165 + 0.3398). Put differently, the interaction terms contain the differences between the "good weather" and "bad weather" coefficients. Another point worth noting is that the SSR for column 4 is exactly the sum for those in columns 2 and 3, as dictated by equation (4.2).

Of course, the fact that those interaction terms are individually significant would be enough to conclude that the null hypothesis of homogeneity between the two regimes has to be rejected. However, the Chow test is easy to compute using the SSRs:

$$W = 731 \times \frac{1498.035 - (779.7312 + 558.5052)}{779.7312 + 558.5052} = 87.2888,$$

where W is a rather astronomical value for a χ^2_3 distribution (the corresponding p value is 8.37114e-19), so we have to reject the null: we conclude that the regression function for sunny days is different from the one for rainy days. ⌟

Historically, the Chow test has mostly been used with time-series data, where each row of the dataset refers to a certain time period and the rows are consecutive. For example, data on the economy of a certain country (GDP, interest rate etc.) in which each row refers to a quarter, so for example the dataset starts in 1980q1, the next row is 1980q2, and so on. Regressions on data of this kind present the user with special issues, that have to be analysed separately, and we will do so in Chapter 5. However, it can be seen rather easily that the Chow test lends itself very naturally to testing whether a model remains stable before and after a certain event: just imagine that in equation (4.1) d_i equals 0 up to a certain point in time and $d_i = 1$ after that. It is for this reason that the Chow test is sometimes referred to as the **structural stability** test. Rejection of the Chow test would in this case point to something we economists often call **structural break** or **regime change**, obvious examples being the introduction of the single currency in the Euro Area, the COVID pandemic, etc.

In a time series context, assuming that the putative date for the break is known *a priori* may be unwarranted. In some case, we may suppose that a structural break has occurred at some point, without knowing exactly when. For these situations, some clever test procedures are available (one is the so-called **CUSUM test**), as well as methods for estimating the timing of the break. These, however, are too advanced for this book, since they require a fairly sophisticated inferential apparatus.

4.2 Heteroskedasticity and its consequences

As the reader might recall, the homoskedasticity assumption was a fundamental ingredient in the derivation of the asymptotic covariance matrix of the OLS estimator (see 3.2). While the linearity assumption $\mathrm{E}\left[y|\mathbf{x}\right] = \mathbf{x}'\boldsymbol{\beta}$ makes consistency almost automatic ($\mathrm{E}\left[\varepsilon \cdot \mathbf{x}\right] = 0$ implies $\hat{\boldsymbol{\beta}} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}$ if some kind of LLN can be invoked), it's impossible to derive the parallel result for asymptotic normality

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathrm{p}} \mathcal{N}\left(\mathbf{0}, \sigma^2 Q^{-1}\right)$$

without assuming homoskedasticity, that is $\mathrm{E}\left[\varepsilon^2|\mathbf{x}\right] = \sigma^2$.

Like all assumptions, homoskedasticity is easier to justify in certain circumstances than others. If data come from controlled experiments, ε_i can often be interpreted as the disturbance term that contaminated the i-th experiment; it is normally safe to think that ε_i should be independent from $\mathbf{x}_i$, the conditions under which the experiment was performed. Therefore, if $\varepsilon_i \perp\!\!\!\perp \mathbf{x}_i$, no moments of ε_i can depend on $\mathbf{x}_i$, and homoskedasticity follows.

This is almost never the case in economics, where virtually all data come from non-experimental settings. This is particularly true for cross-sectional data, where we collect data about individuals who did not take part in an experiment at all. When we estimate a wage equation, our dependent variable y_i (typically, the log wage for individual i) will be matched against a vector of explanatory variables $\mathbf{x}_i$ that contain a description of that individual (education, age, work experience and so on), and ε_i is simply defined as the deviation of y_i from its conditional expectation, so in principle there is no reason to think that it should enjoy any special properties except $\mathrm{E}\left[\varepsilon_i|\mathbf{x}_i\right] = 0$, which holds by construction under the linearity hypothesis.

Therefore, assuming that ε_i has a finite second moment, in general all we can say is that $\mathrm{E}\left[\varepsilon_i^2|\mathbf{x}_i\right]$ is some function of $\mathbf{x}_i$:

$$\mathrm{E}\left[\varepsilon_i^2|\mathbf{x}_i\right] = h(\mathbf{x}_i) = \sigma_i^2, \tag{4.3}$$

where the function $h(\cdot)$ is of unknown form (but certainly non-linear, since σ_i^2 can never be negative). Since the variances σ_i^2 may be different across observations, we use the term **heteroskedasticity**.

The reader may recall (see page 82) that this function is known as the "skedastic" function, and in principle one could try to carry out an inferential analysis of the $h(\mathbf{x}_i)$ function very much like we do with the regression function. However, in this section we will keep to the highest level of generality and simply allow for the possibility that the sequence $\sigma_1^2, \sigma_2^2 \cdots, \sigma_n^2$ contains potentially different numbers, without committing to a specific formula for $h(\mathbf{x}_i)$.

Contrary to what many people think, heteroskedasticity is not a property of the *data*, but only of the *model* we use, since it depends on the conditioning set you use. For example, assume that $\mathrm{E}[y_i|x_i]$ is a constant, so the linear model we would use is

$$y_i = \beta_0 + \varepsilon_i,$$

but the variance of ε_i depends on x_i (for example, $\sigma_i^2 = \mathbf{x}_i^2$). If you estimate a model in which the only regressor is the constant, the model is homoskedastic.

If, however, you estimate the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

that is a perfectly valid representation of the data, since the true value of β_1 is 0, then the model becomes heteroskedastic, since the variance of ε_i is a function of the explanatory variables.

Having said this, it is very common for applied economists to say "the data are heteroskedastic", when you can't get rid of heteroskedasticity in any meaningful model you may think of.

To simplify notation, in this section all expectation operators will be implicitly understood as conditional on $\mathbf{X}$. In other words, we will treat $\mathbf{X}$ as if it were a matrix of constants. Therefore,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \mathrm{V}[\boldsymbol{\varepsilon}] = \mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] = \Sigma$$

If our model is heteroskedastic, then Σ is a diagonal matrix, where elements along the main diagonal need not be equal to each other, and it would look like this:[4]

$$\begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix}$$

The variance of $\hat{\boldsymbol{\beta}}$ can be simply computed as

$$\mathrm{V}\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{V}\left[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \tag{4.4}$$

and clearly if $\Sigma = \sigma^2 I$ we go back to $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. But under heteroskedasticity the $\hat{V} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ matrix may be very far from the actual asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ (shown in equation (4.4)), even asymptotically; therefore, our test statistics are very unlikely to be χ^2-distributed under H_0, which makes our p-values all wrong and inference impossible.

In order to see how the situation can be remedied, it's instructive to consider a case of limited practical relevance, but that provides a few insights that may help later: the case when Σ is known.

[4] In fact, some of our considerations carry over to more general cases, in which Σ is a generic symmetric, positive semi-definite matrix, but let's not complicate matters.

4.2.1 If Σ were known

If the matrix Σ were observable, then the σ_i^2 variances would be known (they are just the diagonal elements of Σ), and getting rid of heteroskedasticity would be easy: define

$$\dot{y}_i = \frac{y_i}{\sigma_i} \qquad \dot{\mathbf{x}}_i = \frac{1}{\sigma_i}\mathbf{x}_i \qquad u_i = \frac{\varepsilon_i}{\sigma_i}$$

and the model becomes

$$\dot{y}_i = \dot{\mathbf{x}}_i'\boldsymbol{\beta} + u_i \tag{4.5}$$

but clearly $\mathrm{V}[u_i] = 1$ by construction,[5] so you can happily run OLS on the transformed variables, since $\dot{y}_i$ and $\dot{\mathbf{x}}_i$ would both be observable. The resulting estimator

$$\tilde{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} \dot{\mathbf{x}}_i \dot{\mathbf{x}}_i'\right]^{-1} \sum_{i=1}^{n} \dot{\mathbf{x}}_i \dot{y}_i = (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\dot{\mathbf{y}}$$

could also be written as

$$\tilde{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\mathbf{x}_i \mathbf{x}_i'\right]^{-1} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\mathbf{x}_i y_i = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y} \tag{4.6}$$

and is called **GLS**. It can be proven that GLS is more efficient than OLS (the proof is in subsection 4.A.2), and that its covariance matrix equals

$$\mathrm{V}\left[\tilde{\boldsymbol{\beta}}\right] = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$$

Since GLS is just OLS on suitably transformed variables, all standard properties of OLS in the homoskedastic case remain valid, so for example you could test hypotheses by the usual techniques.[6]

Some readers may find it intriguing to know that GLS has more or less the geometrical interpretation of OLS that I described in Section 1.4, once a more general definition of "distance" is adopted. GLS arises if ordinary Euclidean distance is generalised to

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)'\Sigma^{-1}(\mathbf{x}_1 - \mathbf{x}_2)}$$

which obviously becomes Euclidean distance if $\Sigma = I$. (The fact that OLS equals GLS if Σ is a scalar multiple of I is a trivial consequence.) You can apply all the usual concepts of projections etc, with the only difference that the space you're considering is somewhat "distorted".

It is interesting to note that, in the case we are considering here, Σ is diagonal, and therefore the operation that makes GLS equivalent to OLS on the transformed data can be written very simply as in equation (4.5). However, it can easily be proven that the formula (4.6) applies far more generally: all that is

[5] Readers would hopefully not feel offended if I reminded them that a straightforward application of equation (2.7) yields $\mathrm{V}[X/b] = \mathrm{V}[X]/b^2$.

[6] In fact, we wouldn't even have to observe Σ, as long as we had a matrix which is *proportional* to Σ by a (possibly unknown) scalar factor. If we had $\Omega = c \cdot \Sigma$, where c is an unknown positive scalar, we could use Ω instead of Σ in equation (4.6), since the scalar c would cancel out.

required is that Σ is a proper covariance matrix, that is, symmetric and positive definite.

Of course, in ordinary circumstances Σ is unknown, but we could use this idea to explore alternative avenues:

1. In some cases, you may have reason to believe that σ_i^2 should be roughly proportional to some observable variable. For example, if y_i is an average from some sampled values and n_i is the size of the i-th sample, it would be rather natural to conjecture that $\sigma_i^2 \simeq K n_i^{-1}$, where K is some constant. Therefore, by dividing all the observables by $\sqrt{n_i}$ you get an equivalent representation of the model, in which heteroskedasticity is less likely to be a problem, since in the transformed model all variances should be roughly equal to K. The resulting estimator is sometimes called **WLS** (for *Weighted Least Squares*), because you "weight" each observation by an observable quantity w_i. In our example, $w_i = \sqrt{1/n_i}$.

2. The idea above can be generalised: one could try to reformulate the model in such a way that the heteroskedasticity problem might be attenuated. For example, it is often the case that, rather than a model like

$$Y_i = \alpha_0 + \alpha_1 X_i + u_i,$$

   a formulation in natural logs, like

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$$

   not only leads to a more natural interpretation of the parameters (since β_1 can be read as an elasticity), but also alleviates heteroskedasticity problems.

3. Even more generally, it can be proven that, if we have $\hat{\Sigma} \xrightarrow{\text{p}} \Sigma$, we can use it in the so-called "feasible" version of GLS, or **FGLS** for short:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{y};$$

   in principle, this can be accomplished by setting an explicit functional form for the conditional variance (the function $h(\cdot)$ in 4.3). It can be done, but in most cases it's much more difficult computationally: the resulting estimator cannot be written in closed form as an explicit function of the observables, but only in implicit form as the minimiser of the least squares function. This in turn, involves computational techniques that are standard nowadays, but are far beyond the scope of an introductory treatment like this one.

   In some cases, however, Σ can be assumed to be a function of a small set of parameters, which can be consistently estimated separately. In that case, FGLS is a perfectly sensible option. One example will be provided in section 7.4.

In many cases, however, neither strategy is possible, so we may have to do with OLS; the next section illustrates how you can make good use of OLS even under heteroskedasticity.

4.2.2 Robust estimation

As the previous section should have made clear, heteroskedasticity doesn't affect consistency of OLS, which therefore remains a perfectly valid estimator (it wouldn't be as efficient as GLS, but this is something we can live with). The real problem is that using the "regular" estimator for $\mathrm{V}\left[\hat{\boldsymbol{\beta}}\right]$, that is

$$\hat{V} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

for hypothesis testing yields statistics that are not asymptotically χ^2-distributed, so all our p-values would be wrong. On the other hand, if we could use the correct variance for OLS (given in equation (4.4) that I'm reporting here for your convenience)

$$\mathrm{V}\left[\hat{\boldsymbol{\beta}}\right] = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Sigma\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1},$$

or anything asymptotically equivalent, inference would be perfectly standard. This seems impossible, given that Σ is unobservable: the middle matrix in the equation above could be written as

$$\mathbf{X}'\Sigma\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_n] \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \sum_{i=1}^{n} \sigma_i^2 \mathbf{x}_i \mathbf{x}_i' \tag{4.7}$$

and it would seem that in order to compute an asymptotically equivalent expression we would need the σ_i^2 variances (or at the very least consistent estimates).

However, although Σ does in fact contain n distinct unknown elements, the size of $\mathbf{X}'\Sigma\mathbf{X}$ is $k \times k$,[7] a fixed number of elements about which, in principle, we may hope to say something as $n \to \infty$. In other words, even if we can't estimate consistently the individual variances σ_i^2, we may be able to estimate consistently the individual elements of the matrix $\mathbf{X}'\Sigma\mathbf{X}$.

This is the basic idea that was put forward in White (1980): first, observe that under heteroskedasticity, OLS is still consistent, so

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\xrightarrow{\mathrm{p}} \boldsymbol{\beta} \quad \Longrightarrow \\ e_i - \varepsilon_i = (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) - (y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{x}_i'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) &\xrightarrow{\mathrm{p}} \mathbf{x}_i'\mathbf{0} = 0 \quad \Longrightarrow \\ e_i \xrightarrow{\mathrm{p}} \varepsilon_i \quad \Longrightarrow \quad e_i^2 &\xrightarrow{\mathrm{p}} \varepsilon_i^2 : \end{aligned}$$

the difference between the OLS residuals e_i and the disturbances ε_i should be "small" in large samples, and likewise for their squares.

[7] In fact, it's a symmetric matrix, so the number of its distinct elements is $k(k+1)/2$.

Next, define a random variable η_i as

$$\eta_i = \varepsilon_i^2 - \mathrm{E}\left[\varepsilon_i^2 | \mathbf{x}_i\right]$$

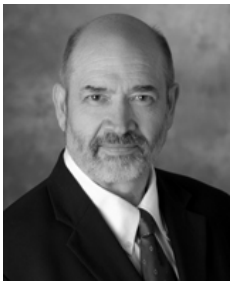
and by a similar argument to that used in Section 3.1, you get that

$$\varepsilon_i^2 = \sigma_i^2 + \eta_i$$

where $\mathrm{E}\left[\eta_i | \mathbf{x}_i\right] = 0$ by definition. Therefore, in large samples, you can approximate σ_i^2 by $e_i^2 - \eta_i$.

If you substitute this into (4.7), you get

$$\mathbf{X}'\Sigma\mathbf{X} \simeq \sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i' - \sum_{i=1}^{n} \eta_i \mathbf{x}_i \mathbf{x}_i'.$$



HAL WHITE

The two elements of the right-hand side are interesting, because the former is observable, while the latter can be easily proven[8] to be a sum of zero-mean variables, which should converge in probability to [0] if divided by n, where I'm using the [0] notation for "a matrix full of zeros".

As a consequence, we'd expect that the average of $e_i^2 - \sigma_i^2$ should be a small quantity, so that

$$\frac{1}{n}\sum_{i=1}^{n}(e_i^2 - \sigma_i^2)\mathbf{x}_i\mathbf{x}_i' \xrightarrow{\mathrm{p}} [0].$$

Now rewrite (4.4) as

$$\mathrm{V}\left[\hat{\boldsymbol{\beta}}\right] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}$$

Therefore, asymptotically you can estimate $\mathrm{V}\left[\hat{\boldsymbol{\beta}}\right]$ via

$$\widetilde{V} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} e_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{4.8}$$

In fact, many variants have been proposed since White's 1980 paper, that seem to have better performance in finite samples, and most packages use one of the later solutions. The principle they are based on, however, is the original one.

A clever variation on the same principle that goes under the name of **cluster-robust** estimation has become very fashiobnable in recent years. I'm not going to describe it in this book, but you should be aware that in some circles you will

[8] It's easy, really: $\mathrm{E}\left[\eta_i | \mathbf{x}_i\right] = 0$ means that $\mathrm{E}\left[\eta_i \mathbf{x}_i \mathbf{x}_i'\right] = \mathrm{E}\left[\mathrm{E}\left[\eta_i | \mathbf{x}_i\right]\mathbf{x}_i\mathbf{x}_i'\right] = [0]$.

be treated like the village idiot if you don't use "clustering". In some cases, people do this just because it's cool and trendy. In some contexts, however, cluster-robust inference is quite appropriate and should be considered as a very useful tool; for example, with panel datasets, which I'll describe in Chapter 7, giving a summary treatment of clustering in Section 7.3.4. For more details, read Cameron and Miller (2010), Cameron and Miller (2015) and MacKinnon et al. (2023).

An even more radical solution for dealing with heteroskedasticity has become quite popular over the recent past because of the enormous advancement of our computing capabilities: it's called the **bootstrap**. In many respects, the bootstrap is a very ingenious solution for performing inference with estimators whose covariance matrix could be unreliable, for various reasons. In a book like this, giving a full account of the bootstrap is far too ambitious a task, and I'll just give you a cursory description in section 4.A.4. Nevertheless, the reader ought to be aware that "bootstrapped standard errors" are becoming more and more widely used in the applied literature.

```
Heteroskedasticity-robust standard errors, variant HC0

             coefficient   std. error    t-ratio     p-value
  ------------------------------------------------------------
  const       8.85359      0.0557726     158.7      0.0000    ***
  lsize       1.03696      0.0270429      38.34     1.85e-255 ***
  baths      -0.00515142   0.0150608      -0.3420   0.7323
  age        -0.00238675   0.000300502    -7.943    2.92e-15  ***
  pool        0.106793     0.0239646       4.456    8.69e-06  ***

Mean dependent var   11.60193   S.D. dependent var   0.438325
Sum squared resid    157.8844   S.E. of regression   0.246187
R-squared            0.685027   Adjusted R-squared   0.684544
F(4, 2605)           929.7044   P-value(F)           0.000000
Log-likelihood      -42.58860   Akaike criterion     95.17721
Schwarz criterion    124.5127   Hannan-Quinn         105.8041
```

Table 4.2: Example: houses prices in the US (with robust standard errors)

Example 4.3

The hedonic model presented in section 3.4 was re-estimated with robust standard errors, and the results are shown in Table 4.2.

As the reader can check, all the figures in Table 4.2 are exactly the same as those in Table 3.1, except for those that depend on the covariance matrix of the parameters: the standard errors (and therefore, the t-statistics and their p values) and the overall specification test. In this case, I instructed gretl to use White's original formula, but this is not the software's default choice (although results would change but marginally).

4.2.3 White's test

Is it possible to test for homoskedasticity? Yes. In fact, many tests exist, and they all have in common the fact that, under H_0, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$ (that is, homoskedasticity). In this section, I will focus on one of the mostly widely used, also due to Hal White: other similar tests (that I will not describe here) go after the name of **Breusch-Pagan** and **Koenker** test.

White's idea is both simple and powerful: if ε_i is homoskedastic, then both estimators of the parameters covariance matrix are consistent, so $\hat{V} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ and its robust counterpart $\widetilde{V}$ should be similar in large samples. Otherwise, the two matrices should diverge from one another. Therefore, one can indirectly spot the problem by comparing the two matrices.[9]

If we re-write $\hat{V}$ as

$$\hat{V} = (\mathbf{X}'\mathbf{X})^{-1}\left(\hat{\sigma}^2\mathbf{X}'\mathbf{X}\right)(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\hat{\sigma}^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1},$$

and compare this expression to (4.8), it's clear to see that any difference between the two variance estimators comes from the matrix in the middle, which equals $\sum_{i=1}^{n}\hat{\sigma}^2\mathbf{x}_i\mathbf{x}_i'$ for $\hat{V}$ and $\sum_{i=1}^{n}e_i^2\mathbf{x}_i\mathbf{x}_i'$ for $\widetilde{V}$. Therefore, the difference between them

$$\frac{1}{n}\sum_{i=1}^{n}(e_i^2 - \hat{\sigma}^2)\mathbf{x}_i\mathbf{x}_i'$$

is the quantity of interest. We need a test for the hypothesis that the probability limit of the expression above is a matrix of zeros. If it were so, then the two estimators would converge to the same limit, and therefore the two estimators would coincide asymptotically; this, of course, wouldn't happen under heteroskedasticity. Therefore, the null hypothesis of White's test is *homoskedasticity*.

Note that the alternative hypothesis is left unspecified: that is, the alternative hypothesis is simply the there is at least one variance σ_i^2 that differs from the other ones. This has two implications, one good and one bad. The good one is that this is a fairly general test and is not specific to any assumption we may make on the skedastic function $h(\mathbf{x}_i)$. The bad one is that the test is "non-constructive": if the null is rejected the test gives us no indication on what to do.

It would seem that performing such a test is difficult; fortunately, an asymptotically equivalent test is easy to compute by means of an auxiliary regression:

$$e_i^2 = \gamma_0 + \mathbf{z}_i'\gamma + u_i;$$

the vector $\mathbf{z}_i$ can be defined, technically, as

$$\mathbf{z}_i = \text{vech}\left(\mathbf{x}_i\mathbf{x}_i'\right).$$

[9] A generalisation of the same principle is known among econometricians as the **Hausman test**, after the great Jerry Hausman. More on this in Section 6.4.

The definition of the vech $(\cdot)$ operator is given in Subsection 4.A.3, but in practice, $\mathbf{z}_i$ contains the non-duplicated cross-products of $\mathbf{x}_i$, that is all combinations of the kind $x_{li} \cdot x_{mi}$ (with $l, m = 1 \dots k$); some of them could cause collinearity, so they must be dropped from the auxiliary regression (see below for an example). Of course, if $\mathbf{x}_i$ contains a constant term, then $\mathbf{z}_i$ would contain all the elements of $\mathbf{x}_i$, as the products $1 \cdot x_{mi}$.

Like in all auxiliary regression, we don't really care about its results; running it is just a computationally convenient way to calculate the test statistic we need, namely

$$LM = n \cdot R^2.$$

Under the null of homoskedasticity, this statistic will be asymptotically distributed as χ^2_p, where p is the size of the vector $\mathbf{z}_i$.

For example: suppose that $\mathbf{x}_i$ contains:

1. the constant;
2. two continuous variables x_i and w_i;
3. a dummy variable d_i

The cross products could be written as per the following "multiplication table"

	1	x_i	w_i	d_i
1	1	x_i	w_i	d_i
x_i	x_i	x_i^2	$x_i \cdot w_i$	$x_i \cdot d_i$
w_i	w_i	$x_i \cdot w_i$	w_i^2	$w_i \cdot d_i$
d_i	d_i	$x_i \cdot d_i$	$w_i \cdot d_i$	d_i

where I indicated the elements to keep by shading the corresponding cell in grey. Of course the lower triangle is redundant, because it reproduces the upper one, but the element in the South-East corner must be dropped too: since d_i is a dummy variable, it only contains zeros and ones, so its square d_i^2 contains the same entries as d_i itself; clearly, inserting both d_i and d_i^2 into $\mathbf{z}_i$ would make the auxiliary regression collinear.

Therefore, the vector $\mathbf{z}_i$ would contain

$$\mathbf{z}_i' = [x_i, \quad w_i, \quad d_i, \quad x_i^2, \quad x_i \cdot w_i, \quad x_i \cdot d_i, \quad w_i^2, \quad w_i \cdot d_i]$$

so the auxiliary regression would read

$$\begin{aligned} e_i^2 &= \gamma_0 + \gamma_1 x_i + \gamma_2 w_i + \gamma_3 d_i + \\ &+ \gamma_4 x_i^2 + \gamma_5 x_i w_i + \gamma_6 x_i d_i + \\ &+ \gamma_7 w_i^2 + \gamma_8 w_i d_i + u_i \end{aligned}$$

where u_i is the error term of the auxiliary regression. In this case, $p = 8$.[10]

[10] It's easy to prove that, if you have k regressors, then $p \leq \frac{k(k+1)}{2} - 1$.

Example 4.4

Running White's heteroskedasticity test on the hedonic model for houses (see section 3.4) yields:

```
White's test for heteroskedasticity
OLS, using observations 1-2610
Dependent variable: uhat^2

             coefficient    std. error     t-ratio    p-value
  ------------------------------------------------------------
  const       0.806236      0.160351       5.028     5.30e-07 ***
  lsize      -0.710993      0.134434      -5.289     1.33e-07 ***
  baths       0.0936181     0.0514158      1.821     0.0688   *
  age         0.00464135    0.00119047     3.899     9.91e-05 ***
  pool        0.267089      0.110699       2.413     0.0159   **
  sq_lsize    0.159277      0.0307527      5.179     2.40e-07 ***
  X2_X3      -0.0409056     0.0240448     -1.701     0.0890   *
  X2_X4      -0.00163418    0.000527404   -3.099     0.0020   ***
  X2_X5      -0.164525      0.0498979     -3.297     0.0010   ***
  sq_baths    0.00360873    0.00513648     0.7026    0.4824
  X3_X4       0.000204179   0.000252621    0.8082    0.4190
  X3_X5       0.0653657     0.0290985      2.246     0.0248   **
  sq_age     -4.93515e-08   4.95338e-06   -0.009963  0.9921
  X4_X5       0.00245662    0.000753929    3.258     0.0011   ***

  Unadjusted R-squared = 0.054056

Test statistic: TR^2 = 141.085963,
with p-value = P(Chi-square(13) > 141.085963) = 0.000000
```

The cross-products are: (a) the original regressors first (because the original model has a constant) and (b) all the cross-products, except for the square of `pool`*, which is a dummy variable. The total number of regressors in the auxiliary model is 14 including the constant, so the degrees of freedom for our test statistic is* $14-1=13$.

Since the LM statistic is 141.1 *(which is a huge number, compared to the* χ^2_{13} *distribution), the null hypothesis of homoskedasticity is strongly rejected. Therefore, the standard errors presented in Table 4.2 are a much better choice than those in Table 3.1.*

4.2.4 So, in practice...

In practice, when you estimate a model in which heteroskedasticity is a possible problem (in practice, every time you have cross-sectional data), you should in principle strive for maximal efficiency, and you can do so by employing the following algorithm, graphically depicted in Figure 4.2.

1. Start with OLS on a tentative model

2. Perform White's test; if it doesn't reject H_0, fine. Otherwise
3. can you reformulate the model so as to achieve homoskedasticity? If you can, try a different formulation and start back from the top. Otherwise,
4. see if you can use FGLS. If you can, do it; otherwise
5. stick to OLS with robust standard errors.

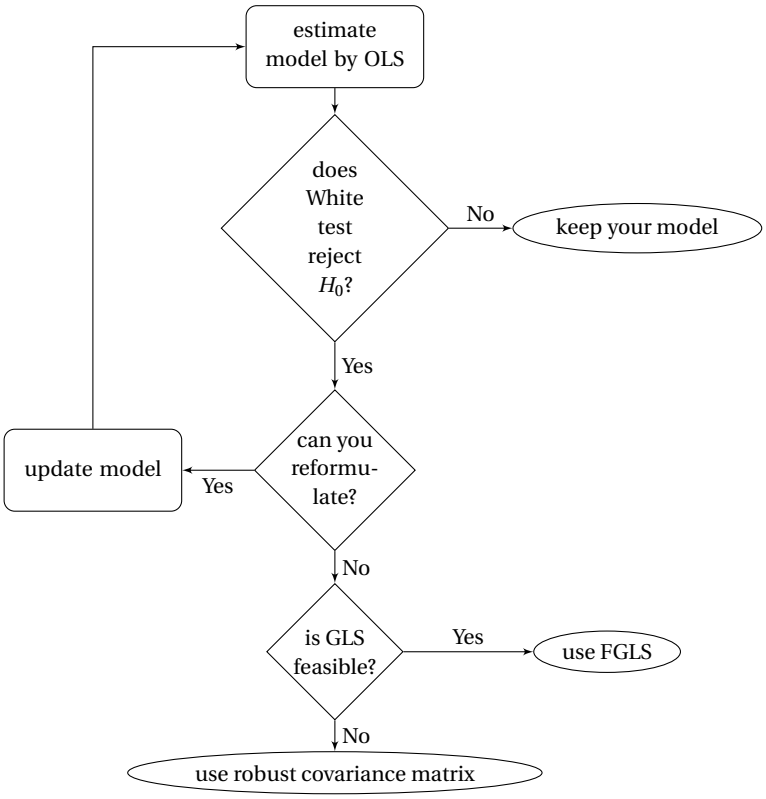


Figure 4.2: Heteroskedasticity flowchart

The things you can do at points 3 and 4 are many: for example, you can try transforming your dependent variable and/or use weighting; for more details, go back to section 4.2.1.

Note, however, that this algorithm often ends at point 5; this is so common that many people, in the applied economics community, don't even bother checking for heteroskedasticity and start directly from there.[11] This is especially true in some cases, where you know from the outset what the situation is. The

[11] In fact, some researchers show sometimes an inclination to disregard specification issues in hope that robust inference will magically take care of everything, which is of course not the case. For an insightful analysis, see King and Roberts (2015).

so-called **linear probability model** (often abbreviated as **LPM**) is a notable example.

The LPM is what you get when your dependent variable is a dummy. So for example you may want to set up a model where y_i is the employment status of an individual, so $y_i = 1$ if the i-th person has a job and $y_i = 0$ otherwise. Contrary to what happens in most cases, we know exactly what the distribution of the dependent variable is: it's a Bernoulli random variable:

$$y_i = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1 - \pi_i \end{cases} \tag{4.9}$$

The linearity hypothesis implies that $\mathrm{E}\left[y_i|\mathbf{x}_i\right] = \pi_i = \mathbf{x}_i'\boldsymbol{\beta}$, since the expected value of a Bernoulli rv is, by construction, the probability of success. This is quite weird already, because π is a probability, and therefore has to be between 0 and 1, whereas if it really was a linear function of the $\mathbf{x}$ variables, you could always imagine to find an observation for which the predicted probability is outside the $[0,1]$ interval. Many applied econometricians are OK with that: they concede that the linearity assumption is inappropriate after all, but assume that it shouldn't be a problem in practice, and use it as a convenient approximation.[12]

But then, you also have that for a Bernoulli rv $\mathrm{V}\left[y_i\right] = \pi_i \cdot (1 - \pi_i)$, and therefore

$$\mathrm{V}\left[y_i|\mathbf{x}_i\right] = \pi_i = \mathbf{x}_i'\boldsymbol{\beta} \cdot (1 - \mathbf{x}_i'\boldsymbol{\beta}_i)$$

so the conditional variance cannot be constant unless the conditional mean is constant too. The vector of parameters $\boldsymbol{\beta}$ enjoys a special nature, being the vector of parameters that determine both the conditional mean and the conditional variance. In theory, it is possible to estimate $\boldsymbol{\beta}$ by an elaborate FGLS strategy, but in these cases practitioners always just use OLS with robust standard errors.

4.A Assorted results

4.A.1 Proof that full interactions are equivalent to split-sample estimation

Suppose you have m categories in which you can split your sample and that all the parameters in your model are liable to be different between the m subsamples.[13] Then, you can write the model as

$$y_i = \sum_{j=1}^{m} (d_{ji} \cdot \mathbf{x}_i)'\boldsymbol{\beta}_j + \varepsilon_i \tag{4.10}$$

[12] Models that overcome this questionable approach have existed for a long time: you'll find a thorough description of *logit* and *probit* models in any decent econometrics textbook, but for some bizarre reason they are going out of fashion.

[13] The classic Chow test occurs when $m = 2$; in order to study the argument below, I suggest you to start with the special case $m = 2$ and generalise later.

where $d_{ji} = 1$ if observation i belongs to sub-population j, and 0 otherwise.

This model can be written in matrix notation as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{X}_m \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix} \tag{4.11}$$

where $\mathbf{y}_j$ is the segment of the $\mathbf{y}$ vector containing the observations for the j-th subsample, and so forth. If you apply the OLS formula to equation (4.11), you get

$$\begin{aligned} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m \end{bmatrix} &= \left(\begin{bmatrix} \mathbf{X}_1' & 0 & \dots & 0 \\ 0 & \mathbf{X}_2' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{X}_m' \end{bmatrix} \cdot \begin{bmatrix} \mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{X}_m \end{bmatrix} \right)^{-1} \times \\ &\times \begin{bmatrix} \mathbf{X}_1' & 0 & \dots & 0 \\ 0 & \mathbf{X}_2' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{X}_m' \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2'\mathbf{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{X}_m'\mathbf{X}_m \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbf{X}_1'\mathbf{y}_1 \\ \mathbf{X}_2'\mathbf{y}_2 \\ \vdots \\ \mathbf{X}_m'\mathbf{y}_m \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}_2 \\ \vdots \\ (\mathbf{X}_m'\mathbf{X}_m)^{-1}\mathbf{X}_m'\mathbf{y}_m \end{bmatrix} \end{aligned}$$

So clearly each $\hat{\boldsymbol{\beta}}_j$ coefficient can be calculated by an OLS regression using the data for subsample j only. Therefore, the residuals for subsample j are $\mathbf{e}_j = \mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j$. As a consequence,

$$\mathbf{e}'\mathbf{e} = \sum_{j=1}^{m} \mathbf{e}_j{}'\mathbf{e}_j,$$

which in words reads: the SSR for model (4.10) is the same as the sum of the SSRs you get for the m separate submodels. Equation (4.2) is a simple special case when $m = 2$; the corresponding generalisation for a generic m is

$$W = n \cdot \frac{S_T - \sum_{j=1}^{m} S_j}{\sum_{j=1}^{m} S_j} \tag{4.12}$$

and the degrees of freedom for the test equals $k \cdot (m-1)$.

Now note that if you take q to be the "reference" category[14], you can rewrite equation (4.10) as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sum_{j \neq q} (d_{ji}\mathbf{x}_i'\boldsymbol{\gamma}_j) + \varepsilon_i$$

where $\boldsymbol{\gamma}_j = \boldsymbol{\beta}_j - \boldsymbol{\beta}_q$ by a simple generalisation of the argument at the start of section 4.1.2. As a consequence, you can compare the model above with the model where all the $\boldsymbol{\gamma}_j$ vectors are 0 by comparing the SSR for the restricted model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ (call it $\mathbf{e}'\mathbf{e}$) against the sum of the SSRs of the m separate submodels (call them $\mathbf{e}'\mathbf{e}_j$, with $j = 1 \dots m$), and they corresponding Wald-type statistic would be exactly equation (4.12).

4.A.2 Proof that GLS is more efficient than OLS

In order to prove that $\mathrm{V}\left[\hat{\boldsymbol{\beta}}\right] - \mathrm{V}\left[\tilde{\boldsymbol{\beta}}\right]$ is psd, I'll use the properties on psd matrices that I listed in section 3.A.6, plus a few more

1. if A and B are invertible and $A - B$ is psd, then $B^{-1} - A^{-1}$ is also psd;
2. if A is psd, there always exists a matrix H such that $A = HH'$;[15]
3. all idempotent matrices are psd.

Therefore, to check the relative efficiency of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, we'll perform an equivalent check on $\Delta \equiv \mathrm{V}\left[\tilde{\boldsymbol{\beta}}\right]^{-1} - \mathrm{V}\left[\hat{\boldsymbol{\beta}}\right]^{-1}$ (by property 1 above). To prove that Δ is psd, start from its definition:

$$\Delta \equiv \mathrm{V}\left[\tilde{\boldsymbol{\beta}}\right]^{-1} - \mathrm{V}\left[\hat{\boldsymbol{\beta}}\right]^{-1} = \mathbf{X}'\Sigma^{-1}\mathbf{X} - (\mathbf{X}'\mathbf{X})(\mathbf{X}'\Sigma\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X});$$

since Σ is pd, we can write it as $\Sigma = HH'$ (by property 2), so that $\Sigma^{-1} = (H')^{-1}H^{-1}$:

$$\begin{aligned} \Delta &= \mathbf{X}'(H')^{-1}H^{-1}\mathbf{X} - (\mathbf{X}'\mathbf{X})(\mathbf{X}'HH'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \\ &= (H^{-1}\mathbf{X})'\left[I - H'\mathbf{X}(\mathbf{X}'HH'\mathbf{X})^{-1}\mathbf{X}'H\right](H^{-1}\mathbf{X}). \end{aligned}$$

Now define $\mathbf{W} = H'\mathbf{X}$ and re-express Δ as:

$$\Delta = (H^{-1}\mathbf{X})'\left[I - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\right](H^{-1}\mathbf{X}) = (H^{-1}\mathbf{X})'\mathbf{M}_{\mathbf{W}}(H^{-1}\mathbf{X}),$$

since $\mathbf{M}_{\mathbf{W}}$ is idempotent, it is psd (property 3); but then, the same is true of $(H^{-1}\mathbf{X})'\mathbf{M}_{\mathbf{W}}(H^{-1}\mathbf{X})$; therefore, the claim follows.

Note that under heteroskedasticity Σ is assumed to be diagonal, but the above proof holds for any non-singular covariance matrix Σ.

[14] I will not offend the reader's intelligence by writing the obvious double inequality $1 \leq q \leq m$.

[15] Note: H is not unique, but that doesn't matter here. By the way, it is also true that if a matrix H exists such that $A = HH'$, then A is psd, but we won't use this result here.

4.A.3 The "vec" and "vech" operators

In some cases, it can be useful to reshape the contents of a matrix so as to transform it into a vector. The "vec" operator does just that: it stacks the columns of a matrix below one another. For example,

$$\text{vec}\left(\begin{bmatrix} a & c \\ b & d \end{bmatrix}\right) = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

or more generally

$$\text{vec}\left(\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_k \end{bmatrix}\right) = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_k \end{bmatrix}.$$

The "vech" operator works in a similar way, but is generally applied to symmetric matrices: the difference from "vec" is that the redundant elements are not considered. For example:

$$\text{vech}\left(\begin{bmatrix} x & y \\ y & z \end{bmatrix}\right) = \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

More generally, if A is an $n \times n$ symmetric matrix, vech (A) is a vector holding the $\frac{n(n+1)}{2}$ elements on and below its diagonal.

4.A.4 The bootstrap

For a reliable account, get hold of Efron and Hastie (2016) (Bradley Efron is none other than the inventor of the technique), or MacKinnon (2006) for a more econometrics-oriented approach. Here, I'm just giving you a basic intuition on what the bootstrap is. Suppose you have an estimator

$$\hat{\theta} = T(\mathbf{X}),$$

where $\mathbf{X}$ is a data matrix with n rows. Clearly, in order to perform inference, you need to have an idea of what the distribution of the random variable $\hat{\theta}$ is. Asymptotically, the CLT may be of help, but perhaps your sample size is not large enough to trust the asymptotic approximation given by the CLT; and even if you're willing to take the asymptotic distribution as an acceptable approximation, the covariance matrix of $\hat{\theta}$ may be unknown, or difficult to compute.

Of course, given your data $\mathbf{X}$ you can compute $\hat{\theta}$ just once, but if you could observe many different datasets with the same distribution, then you could compute your estimator many times and get an idea of the distribution of your statistic by looking at the different values of $\hat{\theta}$ you get each time.

The idea is to use your observed data $\mathbf{X}$ to produce, with the aid of computer-generated pseudo random numbers, H alternative datasets $\mathbf{X}_h$, with $h = 1 \dots H$, and compute your estimator for each of them, so you end up with a collection $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_H$. This procedure is what we call bootstrapping,[16] and the H realisations you get of your statistic are meant to give you an idea of the actual, finite-sample distribution of the statistic itself.

Then, one possible way of computing $V(\hat{\theta})$ is just to take the sample variance of the bootstrap estimates:

$$\begin{aligned} \bar{\theta} &= \frac{1}{H} \sum_{h=1}^{H} \hat{\theta}_h \\ \widetilde{V}(\hat{\theta}) &= \frac{1}{H} \sum_{h=1}^{H} \left(\hat{\theta}_h - \bar{\theta}\right)^2 \end{aligned}$$

How do you generate your artificial datasets $\mathbf{X}_h$? There is a myriad of ways to do this, but when the observations are iid,[17] the simplest solution is just to resample from the rows of $\mathbf{X}$ with replacement, as exemplified in Table 4.3; the example uses the scripting language of gretl, but it should be relatively easy to translate this into any language that you like better.[18] Note that the rows are picked with replacement, which means that you have near-certainty that some of the rows of $\mathbf{X}$ will be present in your "fake" dataset $\mathbf{X}_h$ more than once and some others won't be there at all.

You may find it puzzling, but a simple argument should give you an idea of why this is done. Suppose you have only 3 data points; x_1, x_2 and x_3. If your data are iid, then each of your observations is equally likely, so you could have observed, with the same probability, each of the following 27 datasets:

$$\begin{aligned} \mathbf{X}_1 &= (x_1, x_1, x_1) \\ \mathbf{X}_2 &= (x_1, x_1, x_2) \\ \mathbf{X}_3 &= (x_1, x_1, x_3) \\ \mathbf{X}_4 &= (x_1, x_2, x_1) \\ \mathbf{X}_5 &= (x_1, x_2, x_2) \\ \mathbf{X}_6 &= (x_1, x_2, x_3) \\ &\vdots \\ \mathbf{X}_{26} &= (x_3, x_3, x_2) \\ \mathbf{X}_{27} &= (x_3, x_3, x_3) \end{aligned}$$

[16] According to Efron, "[i]ts name celebrates Baron Munchausen's success in pulling himself up by his own bootstraps from the bottom of a lake" (Efron and Hastie, 2016, p. 177), although the story is reportedly a little different. However, the name was chosen to convey the idea of the accomplishment of something apparently impossible without external help.

[17] When data are not independent, things get a bit more involved.

[18] Warning: the algorithm in Table 4.3 wouldn't be a computationally efficient way to get the job done. It's just meant to illustrate the procedure in the most transparent way possible.

ad it's only by chance that you observed $\mathbf{X}_6$ instead of any of the others. The number 27 comes from the fact that the number of possible datasets is n^n, so in this case $3^3 = 27$. Clearly, the estimator $\hat{\theta}_h$ can be computed for each of the 27 cases and various descriptive statistics can be computed easily. In realistic cases, computing $\hat{\theta}_h$ for each possible sample is impossible, since n^n is astronomical: therefore, we just randomly extract H samples and use those.

```
# allocate space for H estimates (H is the number of bootstrap replications)
matrix thetas = zeros(H, 1)

# generate H simulated datasets and corresponding estimators
loop h = 1 .. H
    Xh = zeros(n, k)              # start with a matrix of zeros

    loop i = 1 .. n               # for each row of our dataset
        k = randgen1(i, 1, n)     # pick a random number between 1 and n
        # put the k-th row of the true data into the i-th row of
        # the simulated data
        Xh[i, ] = X[k, ]
    endloop

    # now compute the estimator on the generated data Xh and store it
    thetas[h] = estimator(Xh)     # this would be the T(X) function
endloop

# compute the variance of the simulated thetas
V = mcov(thetas)
```

Note: this is not meant to run "out of the box". The script above assumes that a few objects, such as the scalars `n` or `H`, or the function `estimator()` have already been defined.

Table 4.3: Elementary example of bootstrap

Chapter 5

Dynamic Models

5.1 Dynamic regression

In cross sectional datasets, it is quite natural to assume that the most useful information set on which to condition the distribution of y_i is $\mathbf{x}_i$. Why should we consider, for the conditional distribution of y_i, the information available for individual j as relevant (with $i \neq j$)? In some cases, there could be something to this; perhaps individuals i and j have some unobservable feature in common, but in most cross-sectional datasets this shouldn't be something to worry about.

This argument does not apply to time-series datasets. Here, we have two fundamental differences from cross-sectional datasets:

1. Data have a natural ordering.

2. At any given point in time, we can take as known what happened in the past (and, possibly, at present time), but the future remains unknown.

This means that, if we want to condition y_t to something[1], we may proceed as in chapter 3 and consider $\mathrm{E}\left[y_t|\mathbf{x}_t\right]$, but this is unlikely to be a good idea, especially in the light of a feature that most economics time series display, that is, they are very *persistent*.

Persistence is a loose term we use for describing the quality, that time series often possess, whereby contiguous observations look more like each other than distant ones. In other words, persistence is the observable consequence of the fact that most phenomena evolve gradually through time.[2] In fact, you may think of time series as something with "memory" of the past. The information embodied in a time series dataset is not only in the numbers it contains, but

[1] Attention: for this chapter, I'm going to switch to a slightly different notation convention than what I used in the previous chapters. Since we're dealing with time series, I will use the symbols t and T instead of i and n, so for example the dependent variable has values $y_1, \ldots, y_t, \ldots, y_T$.

[2] In fact, the econometric treatment of time series data has become, since the 1980s, such a vast and complex subject that we may legitimately treat time-series econometrics as a relatively autonomous scientific field (with financial econometrics as a notable sub-field).

also in the *sequence* in which they come, as if the data told you a story. If you scramble the ordering of the rows in a cross-sectional dataset, the information remains intact; in a time series dataset, most of it is gone.

For example: figure 5.1 shows log of real GDP and log of private consumption in the Euro area between 1995 and 2019 (y and c, respectively).[3] By looking at the plot, it just makes sense to surmise that c_{t-1} may contain valuable information about c_t, even more than y_t does.

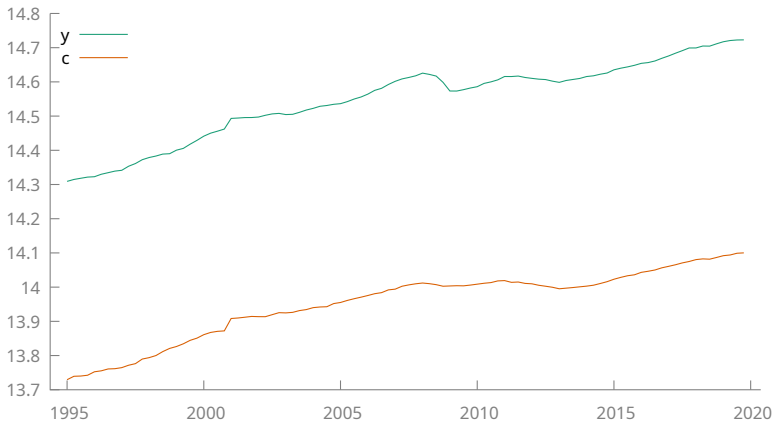


Figure 5.1: Consumption and income in the Euro area (in logs)

Therefore:

- the choice of $\mathbf{x}_t$ as the conditioning set for $\mathrm{E}[y_t|\mathbf{x}_t]$ says implicitly that information on what happened before time t is not of our interest (which is silly);
- since observations are very unlikely to be independent, there is no ground for assuming that covariance matrix of $y_t - \mathrm{E}[y_t|\mathbf{x}_t]$ is diagonal.

In the early days of econometrics, this situation was treated in pretty much the same way as we did with heteroskedasticity in section 4.2, that is, by considering a model like

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t \tag{5.1}$$

and working out solutions to deal with the fact that $\mathrm{E}[\varepsilon\varepsilon'] = \Sigma$ is not a diagonal matrix (although the elements on the diagonal might well be constant).

The presence of non-zero entries outside the diagonal was commonly called the "autocorrelation" or "serial correlation" problem. In order to define this concept,[4] let us begin by defining what the **autocovariance** of a sequence of random

[3] Source: Eurostat.

[4] You may also want to take a look at Section 5.A.2.

variables is: suppose you have T random variables observed through time

$$z_1, z_2, \ldots, z_t, \ldots, z_T.$$

The covariance between z_t and z_s is an autocovariance, since it's the covariance of a random variable "with itself at a different time", so to speak. Clearly, if this quantity is different from 0, the two random variables z_t and z_s cannot be independent. If we standardise this covariance as

$$\rho_{t,s} = \frac{\text{Cov}[z_s, z_t]}{\sqrt{\text{V}[z_t]\,\text{V}[z_s]}}$$

we have something called **autocorrelation**. In most cases, it makes sense to assume that the correlation between z_s and z_t is only a function of how far they are from each other; that is, assume that $\rho_{t,s}$ is just a function of $|t-s|$; if this is the case, the quantity $\text{Corr}[z_{t-1}, z_t] = \text{Corr}[z_t, z_{t+1}] = \ldots$ is called **first-order autocorrelation** or **autocorrelation of order 1**. Generalisation is straightforward.

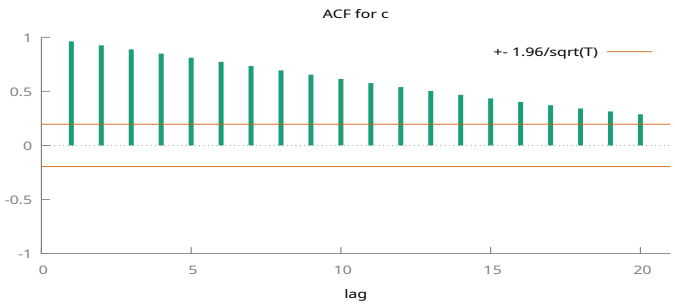


Figure 5.2: Sample autocorrelation for the log consumption series

Example 5.1

Figure 5.2 displays the sample autocorrelations for the log consumption series shown in Figure 5.1. As you can see, the numbers are very different from 0. For example, the first 3 sample correlations equal

$$\begin{aligned}
\hat{\rho}_1 &= \text{Corr}[z_t, z_{t-1}] = 0.9627 \\
\hat{\rho}_2 &= \text{Corr}[z_t, z_{t-2}] = 0.9259 \\
\hat{\rho}_3 &= \text{Corr}[z_t, z_{t-3}] = 0.8881
\end{aligned}$$

and it would be hard to argue that the random variables contained in this time series are independent.

Clearly, if the autocorrelation between ε_t and ε_s is nonzero for some t and s,

Σ cannot be diagonal, so GLS solutions have been devised[5], and a clever generalisation of White's robust estimator (due to Whitney Newey and Kenneth West) is also available, but instead of "fixing" OLS, a much better strategy is to rethink our conditioning strategy. That is, instead of employing clever methods to perform acceptable inference on equation (5.1), we'd be much better off if we redefined our object of interest altogether.

What we want to do is using all the possibly relevant available information as our conditioning set; to this end, define the **information set** at time t as[6]

$$\Im_t = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_t, y_1, y_2, \ldots, y_{t-1}\};$$

(note that $\Im_t$ includes $\mathbf{x}_t$). For example, in order to build a model where consumption is the dependent variable and the only explanatory variable is income (a dynamic consumption function, if you will), it may make sense to condition consumption on the whole information set $\Im_t$.

Therefore, the conditioning operation will be done by using all the variables relevant for the distribution of y_t that can be assumed to be known at that time. Clearly, that includes the current value of $\mathbf{x}_t$, but also the past of both y_t and $\mathbf{x}_t$. Possibly, even future variables that are known with certainty at time t ; variables such as these are normally said to be **deterministic**. Apart from the constant term ($x_t = 1$), popular examples include time trends (eg $x_t = t$), seasonal dummy variables (eg $x_t = 1$ if t is the month of May), or more exotic choices, such as the number of days in a given month, that is known in advance. Note that (this will be very important) $\Im_t$ is an element of a sequence where $\Im_{t-1} \subseteq \Im_t \subseteq \Im_{t+1}$; in other words, the sequence of information sets is increasing.[7]

Now consider the conditional expectation $\mathrm{E}[y_t|\Im_t]$; even under the linearity assumption, this object may have two potentially troublesome characteristics:

1. since the sequence $\Im_t$ is increasing, $\mathrm{E}[y_t|\Im_t]$ may contain information that goes indefinitely back into the past, and

2. $\mathrm{E}[y_t|\Im_t]$ could be different for each t.

If none of the above is true, things are much simplified; under the additional assumption of linearity of the conditional mean,

$$\mathrm{E}[y_t|\Im_t] = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{q} \boldsymbol{\beta}_i' \mathbf{x}_{t-i},$$

[5] For the readers who are into the history of econometrics: the so-called **Cochrane-Orcutt** estimator and its refinements are totally forgotten today, but they were a big thing back in the 1960s and 1970s.

[6] To be rigorous, we should define the information set by using a technical tool called a σ-field. This ensures that $\Im_t$ contains all possible functions of the elements listed above (Δy_{t-1}, for example). But in an introductory treatment such as this, I'll just use the reader's intuition and use $\Im_t$ as "all the things we knows at time t".

[7] Or, if you will, we are assuming that we always learn and never forget.

where p and q are *finite numbers*. Although in principle $\Im_t$ contains all the past, no matter how remote, only the most recent elements of $\Im_t$ actually enter the conditional expectation. A slightly more technical way of expressing the same concept is: we are assuming that there is a subset of $\Im_t$ (call it $\mathscr{F}_t$), that contains only recent information, such that conditioning on $\Im_t$ or $\mathscr{F}_t$ makes no difference:

$$\mathrm{E}\left[y_t|\Im_t\right] = \mathrm{E}\left[y_t|\mathscr{F}_t\right], \tag{5.2}$$

where $\Im_t \supset \mathscr{F}_t$. In practice, $\mathscr{F}_t$ is the relevant information at time t.

The linearity assumption makes the regression function of y_t on $\Im_t$ a **difference equation**, that is, a relationship in which an element of a sequence y_t is determined by a linear combination of its own past and the present and past of another sequence $\mathbf{x}_t$;[8] if we proceed in a similar way as in chapter 3, and define $\varepsilon_t \equiv y_t - \mathrm{E}\left[y_t|\Im_t\right]$, we can write the so-called **ADL model**:

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{q} \boldsymbol{\beta}_i' \mathbf{x}_{t-i} + \varepsilon_t. \tag{5.3}$$

The ADL acronym is for *Autoregressive Distributed Lags* (some people prefer the **ARDL** acronym): in many cases, we call the above an ADL(p,q) model, to make it explicit that the conditional mean contains p lags of the dependent variable and q lags of the explanatory variables.

Of course, it would be very nice if we could estimate the above parameters via OLS. Clearly, the first few observations would have to be discarded, but once this is done, we may construct our $\mathbf{y}$ and $\mathbf{X}$ matrices as[9]

$$\begin{bmatrix} y_{p+1} \\ y_{p+2} \\ y_{p+3} \\ \vdots \end{bmatrix} = \begin{bmatrix} y_p & y_{p-1} & \cdots & y_1 & \mathbf{x}_{p+1}' & \mathbf{x}_p' & \cdots & \mathbf{x}_{p-q+1}' \\ y_{p+1} & y_p & \cdots & y_2 & \mathbf{x}_{p+2}' & \mathbf{x}_{p+1}' & \cdots & \mathbf{x}_{p-q+2}' \\ y_{p+2} & y_{p+1} & \cdots & y_3 & \mathbf{x}_{p+3}' & \mathbf{x}_{p+2}' & \cdots & \mathbf{x}_{p-q+3}' \\ & & & & \vdots & & & \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \\ \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_q \end{bmatrix} + \begin{bmatrix} \varepsilon_{p+1} \\ \varepsilon_{p+2} \\ \varepsilon_{p+3} \\ \vdots \end{bmatrix} =$$

$$\mathbf{y} = \mathbf{W}\gamma + \boldsymbol{\varepsilon}$$

where $\mathbf{w}_t$ is defined as

$$\mathbf{w}_t' = [y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \mathbf{x}_t', \mathbf{x}_{t-1}', \ldots, \mathbf{x}_{t-q}'],$$

and

$$\gamma' = [\alpha_1, \alpha_2, \ldots, \alpha_p, \boldsymbol{\beta}_0', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q'].$$

[8] Note: this definition works for our present purposes, but in some cases you may want to consider non-linear relationships, or cases which involve *future* values.

[9] I assumed for simplicity that $p \geq q$; of course, potentially collinear deterministic terms would have to be dropped.

Given this setup, clearly the OLS statistic can be readily computed with the usual formula $(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}$, but given the nature of the conditioning, one may wonder if OLS is a CAN estimator of the α and β parameters. As we will see in section 5.3, the answer is positive, under certain conditions.

Before we focus on the possible inferential difficulties, however, it is instructive to consider another problem. Even if the parameters of the conditional expectation $\mathrm{E}[y_t|\Im_t]$ were known and didn't have to be estimated, *how do we interpret them*?

5.2 Manipulating difference equations

Given the difference equation

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{q} \beta_i' \mathbf{x}_{t-i}$$

we may ask ourselves: what is the effect of $\mathbf{x}$ on y *after a given period*? That is: how does $\mathbf{x}_t$ affect y_{t+h}? Since the coefficients α_i and β_i do not depend on t, we may rephrase the question as: what is the impact on y_t of something that happened h periods ago, that is $\mathbf{x}_{t-h}$? Clearly, if $h = 0$ we have a quantity that is straightforward to interpret, that is the instantaneous impact of $\mathbf{x}_t$ in y_t, but much is to be gained by considering magnitudes like

$$d_h = \frac{\partial y_t}{\partial \mathbf{x}_{t-h}} = \frac{\partial y_{t+h}}{\partial \mathbf{x}_t}; \tag{5.4}$$

the d_h parameters take the name of **dynamic multipliers**, or just multipliers for short. In order to find a practical and general way to compute them, we will need a few extra tools. Read on.

5.2.1 The lag operator

Time series are nothing but sequences of numbers, with a natural ordering given by time. In many cases, we may want to manipulate sequences by means of appropriate operators. The **lag operator** is generally denoted by the letter L by econometricians (statisticians prefer B — savages!); it's an operator that turns a sequence x_t into another sequence, that contains the same objects as x_t, but shifted back by one period.[10] If you apply L to a constant, the result is the same constant. In formulae,

$$Lx_t = x_{t-1}$$

Repeated application of the L operator n times is indicated by L^n, and therefore $L^n x_t = x_{t-n}$. By convention, $L^0 = 1$. The L operator is *linear*, which means that,

[10] In certain cases, you might want to use the **lead operator**, usually notated as F, which is defined as the inverse to the lag operator ($Fx_t = x_{t+1}$, or $F = L^{-1}$). I'm not using it in this book, but its usage is very common in economic models with rational expectations.

if a and b are constant, then $L(ax_t + b) = aLx_t + b = ax_{t-1} + b$. These simple properties have the nice consequence that, in many cases, we can manipulate the L operator algebraically as if it was a number. This trick is especially useful when dealing with *polynomials* in L. Allow me to exemplify:

Example 5.2

Call b_t the money you have at time t, and s_t the difference between the money you earn and the money you spend between $t-1$ and t (in other words, your savings). Of course,

$$b_t = b_{t-1} + s_t.$$

Now the same thing with the lag operator::

$$b_t = Lb_t + s_t \rightarrow b_t - Lb_t = (1-L)b_t = \Delta b_t = s_t$$

The Δ operator, which I suppose not unknown to the reader, is defined *as $(1-L)$, that is a polynomial in L of degree 1. The above expression simply says that the variation in the money you have is your net saving.*

Example 5.3

Call q_t the GDP for the Kingdom of Verduria in quarter t. Obviously, yearly GDP is given by

$$y_t = q_t + q_{t-1} + q_{t-2} + q_{t-3} = (1 + L + L^2 + L^3)q_t$$

Since $(1 + x + x^2 + x^3)(1 - x) = (1 - x^4)$, if you "multiply" the equation above[11] *by $(1 - L)$ you get*

$$\Delta y_t = (1 - L^4)q_t = q_t - q_{t-4};$$

The variation in yearly GDP between quarters is just the difference between the quarterly figures a year apart from each other.

A polynomial $P(x)$ may be evaluated at any value, but two cases are of special interest. Obviously, if you evaluate $P(x)$ for $x = 0$ you get the "constant" coefficient of the polynomial, since $P(0) = p_0 + p_1 \cdot 0 + p_2 \cdot 0 + \cdots = p_0$; instead, if you evaluate $P(1)$ you get the sum of the polynomial coefficients:

$$P(1) = \sum_{j=0}^{n} p_j 1^j = \sum_{j=0}^{n} p_j.$$

This turns out to be quite handy when you apply a lag polynomial to a constant, since

$$P(L)\mu = \sum_{j=0}^{n} p_j \mu = \mu \sum_{j=0}^{n} p_j = P(1)\mu.$$

[11] To be precise, we should say: 'if you apply the $(1 - L)$ operator to the expression above'.

There are two more routine results that come in very handy: the first one has to do with inverting polynomials of order 1. It can be proven that, if $|\alpha| < 1$,

$$(1-\alpha L)^{-1} = (1+\alpha L+\alpha^2 L^2+\cdots) = \sum_{i=0}^{\infty} \alpha^i L^i; \tag{5.5}$$

the other one is that a polynomial $P(x)$ is invertible if and only if all its roots are greater than one in absolute value:

$$\frac{1}{P(x)} \quad \text{exists iff} \quad P(x)=0 \Rightarrow |x|>1. \tag{5.6}$$

The proofs are in subsection 5.A.1.

Example 5.4 (The Keynesian multiplier)

Let me illustrate a possible use of polynomial manipulation by a very old-school macro example: the simplest possible version of the Keynesian multiplier idea. Suppose that

$$Y_t = C_t + I_t; \tag{5.7}$$
$$C_t = \alpha Y_{t-1}; \tag{5.8}$$

where Y_t is GDP, C_t is aggregate consumption and I_t is investment; $0<\alpha<1$ is the marginal propensity to consume.

By combining the two equations,

$$Y_t = \alpha Y_{t-1} + I_t \rightarrow (1-\alpha L)Y_t = I_t.$$

therefore, by applying the first degree polynomial $A(L)=(1-\alpha L)$ to the Y_t sequence (national income), you get the time series for investments, simply because $I_t = Y_t - C_t = Y_t - \alpha Y_{t-1}$.

If you now invert the $A(L)=(1-\alpha L)$ operator,

$$Y_t = (1+\alpha L+\alpha^2 L^2+\cdots)I_t = \sum_{i=0}^{\infty} \alpha^i I_{t-i}:$$

aggregate demand at time t can be seen as a weighted sum of past and present investment. Suppose that investment goes from 0 to 1 at time 0. This brings about a unit increase in GDP via equation (5.7); but then, at time 1 consumption goes up by α, by force of equation (5.8), so at time 2 it increases by α^2 and so on. Since $0<\alpha<1$, the effect dies out eventually.

If investments were constant through time, then $I_t = \bar{I}$; therefore, $A(L)Y_t = \bar{I}$ becomes

$$Y_t = \frac{1}{A(L)}\bar{I} = \frac{1}{A(1)}\bar{I} = \frac{\bar{I}}{1-\alpha}$$

where the second equality comes from the fact that $\bar{I}$ is constant. The rightmost expression is nothing but the familiar "Keynesian multiplier" formula. ⌟

A word of caution: in many cases, it's OK to manipulate L algebraically as if it was a number, but sometimes it's not: the reader should always keep in mind that the expression Lx_t does not mean 'L times x_t', but 'L applied to x_t'. The following example should, hopefully, convince you.

Example 5.5

Given two sequence x_t and y_t, define the sequence z_t as $z_t = x_t \cdot y_t$. Obviously, $z_{t-1} = x_{t-1}y_{t-1}$; however, one may be tempted to argue that

$$z_{t-1} = x_{t-1}y_{t-1} = Lx_tLy_t = L^2x_ty_t = L^2z_t = z_{t-2}$$

which is obviously absurd.

5.2.2 Dynamic multipliers

When considering an ADL model, the problem that we are ultimately after is: how do we interpret its parameters? Let's start from a difference equation like the following:

$$A(L)y_t = B(L)x_t$$

where the degrees of the $A(L)$ and $B(L)$ polynomials are p and q, respectively. If the polynomial $A(L)$ is invertible, the difference equation is said to be **stable**. In this case, we may define $D(L) = A(L)^{-1}B(L) = B(L)/A(L)$; as a rule, $D(L)$ is of infinite order (although not necessarily so):

$$y_t = D(L)x_t = \sum_{i=0}^{\infty} d_i x_{t-i}.$$

This is all we need for dealing with our problem, if you consider that the dynamic multipliers as defined in equation (5.4),

$$d_i = \frac{\partial y_t}{\partial x_{t-i}} = \frac{\partial y_{t+i}}{\partial x_t},$$

are simply the coefficients of the $D(L)$ polynomial.[12] It is possible to calculate them analytically by inverting the $A(L)$ polynomial, but doing so is neither instructive nor enjoyable. On the contrary, the same effect can be achieved by using a nice recursive algorithm.

The impact multiplier d_0 is easy to find, since $d_0 = D(0) = B(0)/A(0)$, which simply equals β_0 (since $A(0) = 1$). All other multipliers can be found by means of (5.4), which can be used to express d_i in terms of d_{i-1}, d_{i-2} etc. Once you have d_0, the rest of the sequence follows.

[12] As we will see in section 5.3.1, invertibility of $A(L)$ is not only required for the calculation of the multipliers, but also for the CAN property of OLS.

Let me show you a practical example. For an ADL(1,1) model,

$$y_t = \alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1}, \tag{5.9}$$

use the definition of a multiplier as a derivative and write

$$\begin{aligned} d_0 &= \frac{\partial y_t}{\partial x_t} = \frac{\partial}{\partial x_t}\left(\alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1}\right) = \beta_0 \\ d_1 &= \frac{\partial y_t}{\partial x_{t-1}} = \frac{\partial}{\partial x_{t-1}}\left(\alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1}\right) = \alpha \frac{\partial y_{t-1}}{\partial x_{t-1}} + \beta_1 = \alpha d_0 + \beta_1, \end{aligned}$$

where we used the property

$$\frac{\partial y_{t-1}}{\partial x_{t-1}} = \frac{\partial y_t}{\partial x_t} = d_0$$

in such a way that d_1 is expressed as a function of d_0; similarly,

$$d_2 = \frac{\partial y_t}{\partial x_{t-2}} = \frac{\partial}{\partial x_{t-2}}\left(\alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1}\right) = \alpha \frac{\partial y_{t-1}}{\partial x_{t-2}} = \alpha d_1$$

and so on, recursively.

A nice and cool way to express the above is by saying that the multipliers can be calculated through a difference equation *with the same polynomials as the original one*; the sequence of multipliers obeys the relationship

$$A(L) d_i = B(L) u_i, \tag{5.10}$$

where u_i is a sequence that contains 1 for u_0, and 0 everywhere else. This makes it easy to calculate the multipliers numerically, given the polynomial coefficients, via appropriate software.

Example 5.6 (Multiplier calculation)
Take for example the following difference equation:

$$y_t = 0.2 y_{t-1} + 0.4 x_t + 0.3 x_{t-2}.$$

In this case, $A(L) = 1 - 0.2L$ *and* $B(L) = 0.4 + 0.3L^2$*. The inverse of* $A(L)$ *is*

$$A(L)^{-1} = 1 + 0.2L + 0.04L^2 + 0.008L^3 + \cdots = \sum_{i=0}^{\infty} 0.2^i L^i;$$

therefore,

$$\frac{B(L)}{A(L)} = (0.4 + 0.3L^2) \times (1 + 0.2L + 0.04L^2 + 0.008L^3 + \cdots).$$

The two polynomials can be multiplied directly, as in

$$\begin{aligned}\frac{B(L)}{A(L)} &= 0.4\times(1+0.2L+0.04L^2+0.008L^3+\cdots)+\\ &\quad +0.3L^2\times(1+0.2L+0.04L^2+0.008L^3+\cdots)=\\ &= 0.4+0.08L+0.016L^2+0.0032L^3+\cdots+\\ &\quad +0.3L^2+0.06L^3+0.012L^4+0.0024L^5\cdots=\\ &= 0.4+0.08L+0.316L^2+0.0632L^3+\cdots\end{aligned}$$

but it's really boring. The recursive approach is much quicker:

$$\begin{aligned}d_0 &= B(0)/A(0)=0.4/1=0.4\\ d_1 &= 0.2\cdot d_0=0.08\\ d_2 &= 0.2\cdot d_1+0.03=0.016+0.3=0.316\\ d_3 &= 0.2\cdot d_2=0.0632\end{aligned}$$

and so on.

In certain cases, the multipliers d_i may all have the same sign. If so, the sequence $\pi_i=\frac{d_i}{c}$ has all the characteristics of a discrete probability distribution: all the π_i coefficients are non-negative and sum to 1.
Therefore, it makes sense to compute quantities such as the mean lag or the median lag. For example, the mean lag can be defined as $m=\sum_{i=0}^{\infty} i\cdot\pi_i$, and can be given a nice interpretation as the "average" time span it takes x_t to affect y_t.
Note, however, that in general the sequence d_i may well include positive and negative numbers, and the long-run multiplier c could even be 0; in those cases, the notion itself of mean lag is meaningless.

5.2.3 Interim and long-run multipliers

If you go back to the definition of the multipliers, (5.4), that is $d_i=\frac{\partial y_t}{\partial x_{t-i}}=\frac{\partial y_{t+i}}{\partial x_t}$, it is quite natural to interpret the magnitude d_h as the effect of something that happened h periods ago on what we see today. The implicit idea in this definition is that the source of the dynamic behaviour in our system is a one-off event.

In many cases, instead, we could be interested in computing the effect on y_t of a *permanent* change in x_t. Clearly, at time 0 the effect will be equal to the impact multiplier d_0, but after one period the instantaneous effect will overlap with the lagged one, so the effect will be equal to d_0+d_1. By induction, we may define a new sequence of multipliers as

$$c_j=d_0+d_1+\cdots+d_j=\sum_{i=0}^{j}d_i. \tag{5.11}$$

These are called **interim multipliers** and measure the effect on y_t of a *permanent* change in x_t that took place j periods ago. In order to see what happens in the long run after a permanent change, we may also want to consider the **long-run multiplier** $c = \lim_{j\to\infty} c_j$. Calculating c is much easier than what may seem, since

$$c_j = \sum_{i=0}^{\infty} d_i = D(1);$$

that is: c is the number you get by evaluating the polynomial $D(z)$ in $z = 1$; since $D(z) = B(z)/A(z)$, c can be easily computed as $c = D(1) = \frac{B(1)}{A(1)}$.

Example 5.7 (interim multipliers)
Let's go back to the difference equation we used in example 5.6:

$$y_t = 0.2y_{t-1} + 0.4x_t + 0.3x_{t-2}.$$

Interim multipliers are easily computed:

$$\begin{aligned} c_0 &= d_0 = 0.4 \\ c_1 &= d_0 + d_1 = c_0 + d_1 = 0.48 \\ c_2 &= d_0 + d_1 + d_2 = c_1 + d_2 = 0.796 \end{aligned}$$

and so on. The limit of this sequence (the long-run multiplier) is also easy to compute:

$$c = D(1) = \frac{B(1)}{A(1)} = 0.7/0.8 = 0.875$$

Et voilà.

The long-run multiplier c is very important, because it describes the relationship between y_t and x_t in **steady state**. The concept of steady state is of paramount importance in econometrics, because is the closest you get to what you refer to as "equilibrium" in theoretical economics: by "equilibrium", we usually mean that there is no internal force that pushes the state of the system away from where it currently is. Therefore, if a system is in equilibrium, all the variables that describe it will remain stable through time until an external shock occurs.

When the dynamic behaviour of a system is described by a difference equation, the concept of steady state can be explained as follows: suppose we fix x_t at a certain value that stays the same forever. Is there a limit value for y_t? It can be shown that the limit exists as long as the difference equation is stable; if this condition is met, then the system admits a steady state. The steady state is a long-run equilibrium: in steady state, neither y_t nor x_t change until external shocks come from outside the system to perturb it.

Mathematically, if the system is in steady state, both variables are invariant through time, so in steady state $y_t = Y$ and $x_t = X$ (note that Y and X bear no subscript); as a consequence,

$$A(L)y_t = B(L)x_t \Rightarrow A(L)Y = B(L)X \Rightarrow A(1)Y = B(1)X \Rightarrow Y = \frac{B(1)}{A(1)}X = cX,$$

where we used the property that, if X is a constant sequence, $L^n X = X$. Therefore, the system is *not* in equilibrium any time $y_t \neq cx_t$. As we will see, this trivial observation will become important later.

5.3 Inference on OLS with time-series data

At this point, we know how to interpret the coefficients of a difference equation. An ADL model (equation (5.3), reproduced here for the reader's convenience in lag-polynomial notation)

$$A(L)y_t = B(L)\mathbf{x}_t + \varepsilon_t \tag{5.12}$$

is basically a difference equation plus an error term; therefore, the coefficients of the two polynomials $A(L)$ and $B(L)$ are unobservable, but perhaps we could find a CAN estimator.

We showed in section 5.1 how OLS can be applied to a dynamic model by defining the $\mathbf{X}$ matrix and the $\mathbf{y}$ vector appropriately. The question now is: is OLS a CAN estimator of the ADL parameters? The answer is positive, if certain conditions are satisfied.

5.3.1 Martingale differences

Define $\mathbf{w}_t$ like at the end of section 5.1, as

$$\mathbf{w}_t' = [y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \mathbf{x}_t', \mathbf{x}_{t-1}', \ldots, \mathbf{x}_{t-q}'],$$

so that we can write our dynamic model as

$$y_t = \mathbf{w}_t'\boldsymbol{\gamma} + \varepsilon_t$$

where of course $\boldsymbol{\gamma}' = [\alpha_1, \alpha_2, \ldots, \alpha_p, \boldsymbol{\beta}_0', \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q']$.

The first important requirement is that the second moments of $\mathbf{w}_t$ exist and that

$$T^{-1}\sum_{t=1}^{T}\mathbf{w}_t\mathbf{w}_t' \xrightarrow{\text{p}} Q$$

where Q is invertible. The conditions under which we can expect this to happen are quite tricky to lay down formally. Here, I'll just say that in order for everything to work as expected, it is sufficient that our observed data are realisations

of *covariance-stationary* and *ergodic* stochastic processes.[13] For a summary description of what this means, I have written subsection 5.A.2 at the end of this chapter. If you can't be bothered, just take this to mean that all moments up to the fourth order of all the observables exist and are stable through time.

On top of this, a fundamental ingredient for OLS being a CAN estimator of the parameters in equation (5.12) is that ε_t be a **martingale difference** sequence (or **MDS** for short).[14]

Roughly speaking, a MDS is a sequence of random variables whose expected value (conditional to a certain information set) meets certain requirements:

- a **martingale** with respect to $\Im_{t-1}$ is a sequence of random variables X_t such that $\mathrm{E}[X_t|\Im_{t-1}] = X_{t-1}$;
- If X_t is a martingale, then ΔX_t is a MDS: $\mathrm{E}[\Delta X_t|\Im_{t-1}] = 0$

Of course, if we could condition y_t on $\Im_t$ then ε_t would be a MDS by construction:

$$\mathrm{E}[\varepsilon_t|\Im_t] = \mathrm{E}\left[y_t - \mathrm{E}\left[y_t|\Im_t\right]|\Im_t\right] = \mathrm{E}\left[y_t|\Im_t\right] - \mathrm{E}\left[y_t|\Im_t\right] = 0$$

(this is essentially the same argument we used in section 3.1). But of course we can't use $\Im_t$ in practice; however, we're assuming (see equation 5.2) that there exists a subset $\mathscr{F}_t \subset \Im_t$ such that $\mathrm{E}\left[y_t|\Im_t\right] = \mathrm{E}\left[y_t|\mathscr{F}_t\right]$, so $\mathscr{F}_t$ (which is usable, because it's finite) is just as good. So if you condition y_t on $\mathscr{F}_t$, the quantity $\varepsilon_t = y_t - \mathrm{E}\left[y_t|\mathscr{F}_t\right]$ is a MDS and all is well.

However, what happens if you use a conditioning set $\mathscr{G}_t$ that is "too small"? That is, that doesn't contain $\mathscr{F}_t$? In that case, the difference $u_t = y_t - \mathrm{E}\left[y_t|\mathscr{G}_t\right]$ is *not* a MDS with respect to $\Im_t$: if $\mathrm{E}\left[y_t|\mathscr{G}_t\right] \neq \mathrm{E}\left[y_t|\mathscr{F}_t\right]$, then

$$\mathrm{E}[u_t|\Im_t] = \mathrm{E}\left[y_t - \mathrm{E}\left[y_t|\mathscr{G}_t\right]|\Im_t\right] = \mathrm{E}\left[y_t|\mathscr{F}_t\right] - \mathrm{E}\left[y_t|\mathscr{G}_t\right] \neq 0.$$

On the contrary, it is easy to prove that in the opposite case, when you condition on a subset of $\Im_t$ that is *larger* than $\mathscr{F}_t$, no problems arise.

This remark is extremely important in practice because the order of the polynomials $A(L)$ and $B(L)$ (p and q, respectively) are not known: what happens if we get them wrong? Well, if they are larger than the "true" ones, then our conditioning set contains $\mathscr{F}_t$, and all is well. But if they're smaller, the disturbance term of our model is not a MDS, and all inference collapses. For example, if $p = 2$ and $q = 3$, then $\mathscr{F}_t$ contains $y_{t-1}, y_{t-2}, x_t, x_{t-1}, x_{t-2}$ and x_{t-3}. Any set of regressors that doesn't include at least these renders inference invalid.

[13] I'm being very vague and unspecific here: if you want an authoritative source on the asymptotics for dynamic models, you'll want to check chapters 6 and 7 in Davidson (2000).

[14] MDSs arise quite naturally in inter-temporal optimisation problems, so their usage in economic and finance models with uncertainty is very common. In this contexts, an MDS is, so to speak, something that cannot be predicted in any way from the past. For a thorough discussion, see Hansen and Sargent (2013), chapter 2.

Therefore, ε_t is a MDS, if we pick p and q large enough. The obvious implication is that $\mathrm{E}[\varepsilon_t|\mathbf{w}_t] = 0$ (since $\mathbf{w}_t$ is contained in $\Im_{t-1}$). If we also add a homoskedasticity assumption $\mathrm{E}\left[\varepsilon_t^2|\mathbf{w}_t\right] = \sigma^2$, then we have a set of results that parallel completely those in section 3.2. To put it simply, everything works exactly the same way as in cross-sectional models: the martingale property ensures that $\mathrm{E}[\mathbf{w}_t \cdot \varepsilon_t|\Im_t] = \mathbf{0}$, and therefore $\hat{\boldsymbol{\gamma}} \xrightarrow{\mathrm{p}} \boldsymbol{\gamma}$; additionally,

$$\sqrt{T}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mathbf{0}, \sigma^2 Q^{-1}\right).$$

In practice, the whole testing apparatus we set up for cross sectional datasets remains valid; the t statistic, the W statistic, everything. Nice, isn't it? In addition, since the dynamic multipliers are continuous and differentiable functions of the ADL parameters $\boldsymbol{\gamma}$, we can simply compute the multipliers from the estimated parameters $\hat{\boldsymbol{\gamma}}$ and get automatically CAN estimators of the multipliers.[15]

The homoskedasticity assumption is not normally a problem, except for financial data at high frequencies (eg daily); for those cases, you get a separate class of models, the most notable example of which is the so-called **GARCH** model, which I will not consider here, but are extremely important in the field of financial econometrics. In case we want to stick with OLS, robust estimation is perfectly viable.

5.3.2 Testing for autocorrelation and the general-to-specific approach

Basically, we need a test for deciding, on the basis of the OLS residuals, whether ε_t is a MDS or not. Because if it were not, the OLS estimator would not be consistent for the ADL parameters, let alone have the asymptotic distribution we require for carrying out tests. As I argued in the previous section, ε_t cannot be a MDS if we estimate a model in which the orders p and q that we use for the two polynomials $A(L)$ and $B(L)$ are too small.

Most tests hinge on the fact that a MDS cannot be autocorrelated:[16] for the sake of brevity, I don't prove this here, but the issue is discussed in section 5.A.3 if you're interested. Therefore, in practice, the most important diagnostic check on a dynamic regression model is checking for autocorrelation: if we reject the null of no autocorrelation, then ε_t cannot be a MDS.

All econometric software pays tribute to tradition by reporting a statistic invented by James Durbin and Geoffrey Watson in 1950, called **DW statistic** in their honour. Its support is, by construction, the interval between 0 and 4, and ideally it should be close to 2. It is practically useless, because it only checks for autocorrelation of order 1, and there are several cases in which it doesn't work

[15] Unfortunately, the function linking multipliers and parameters is nonlinear, so you need the delta method to compute their asymptotic variance. See section 2.3.2.

[16] If I were insufferably pedantic, I would say "a MDS with finite second moments".

(notably, when lags of the dependent variable are among the regressors); therefore, nobody uses it anymore, although all software packages routinely print it out as a homage to tradition.

The **Godfrey test** (also known as the Breusch-Godfrey test, or the LM test for autocorrelation) is much better:

$$A(L)y_t = B(L)\mathbf{x}_t + \gamma_1 e_{t-1} + \gamma_2 e_{t-2} + \cdots + \gamma_h e_{t-h} + \varepsilon_t$$

where e_t is the t-th OLS residual and h is known as the order of the test. There is no precise rule for choosing h; the most important aspect to consider is "how long is the period we can reasonably expect to consider long enough for dynamic effects to show up?". When dealing with macro time series, a common choice is 2 years. That is, it is tacitly assumed that nothing can happen now, provoke no effects for two years, and then suddenly do something.[17] Therefore, you would use $h = 2$ for yearly data, $h = 8$ for quarterly data, and so on. But clearly, this is a very subjective criterion, so take it with a pinch of salt and be ready to adjust it to your particular dataset.

This test, being a variable addition test, is typically implemented as an LM test (see section 3.5.1) and is asymptotically distributed (under H_0) as χ^2_h. In practice, you carry out an auxiliary regression of the OLS residuals e_t against $\mathbf{w}_t$ and h lags of e_t; you multiply R^2 by T and you're done.

The Godfrey test is the cornerstone of the so-called **general-to-specific** estimation strategy: since the polynomial orders p and q are not known in practice, one has to make a guess. There are three possible situations:

1. your guess is exactly right; you're a lucky bastard.

2. Your guess is wrong because you overestimated p and/or q: in this case, your model contains the "true" one and the disturbance term will still be a MDS; hence, the probability of the Godfrey test rejecting the null hypothesis is 5%. The only slight inconvenience is that you're using too many parameters. This is not a problem, however, because asymptotic inference is valid and you can trim your model down by using ordinary specification tests (see section 3.3).

3. Your guess is wrong because you underestimated one of p and q: your model does not contain the "true" one and the disturbance term will not be a MDS. In this case, the Godfrey test should reject the null.

So the idea of the general-to-specific approach is: start from a large model, possibly ridiculously oversized. Then you can start refining it by ordinary hypothesis tests, running diagnostics[18] at each step to make sure your reduction was not too aggressive.

[17] "*Mi ha detto mio cuggino che sa un colpo segreto...*", EELST.

[18] The most important test to run at this stage is of course the Godfrey test, but other diagnostics, such as the RESET test for example, won't hurt.

5.4 An example, perhaps?

If we were to ignore the points I raised at the beginning of this chapter, we could simply use the data depicted in Figure 5.1 to estimate the parameters of what an economist in the 1970s would have called a "consumption function" and regress consumption at time t on a time trend and DGP at time t. If we did, we'd obtain a "static model"

$$c_t = \beta_0 + \beta_1 t + \beta_2 y_t + \varepsilon_t,$$

whose output is in Table 5.1.

```
OLS, using observations 1995:1-2019:4 (T = 100)
Dependent variable: c

             coefficient    std. error    t-ratio   p-value
  ------------------------------------------------------------
  const      -0.372424      0.391519      -0.9512   0.3439
  time       -0.000436889   0.000107381   -4.069    9.64e-05 ***
  y           0.986243      0.0272713     36.16     4.18e-58 ***

Mean dependent var   13.95257   S.D. dependent var   0.100926
Sum squared resid    0.007333   S.E. of regression   0.008695
R-squared            0.992728   Adjusted R-squared   0.992578
F(2, 97)             6621.001   P-value(F)           1.9e-104
Log-likelihood       334.1324   Akaike criterion    -662.2647
Schwarz criterion   -654.4492   Hannan-Quinn        -659.1016
rho                  0.866455   Durbin-Watson        0.254935

Breusch-Godfrey test for autocorrelation up to order  4
TR^2 = 78.249011, with p-value =  4.08e-16
```

Table 5.1: Static regression example

Superficially, it would look as if the static model is a rather good one: R^2 looks great, but this is common with trending data (as macro time series typically are). The important thing is that $\hat{\rho} = 0.866$ and the Godfrey test rejects the null hypothesis with a vengeance. In an equation like the one above, there is no way the disturbance term ε_t can be thought of as an MDS. Therefore, not only inference is invalid. There's much more than can be said about income affects consumption *through time*.

If instead we enlarge the information set to $\Im_t$, the model we come up with is an ADL(1,2) model. In practice:

$$c_t \simeq k + \alpha c_{t-1} + \beta_0 y_t + \beta_1 y_{t-1} + \beta_2 y_{t-2};$$

table 5.2 contains the OLS estimates, that is $\hat{\alpha} = 0.894$, $\hat{\beta}_0 = 0.589$, and so on. Also note that the time trend, which appeared to be highly significant in the static model, drops out in the dynamic model.

In this case the Godfrey test cannot reject the null, so we may be confident that inference is correct. The next thing we want to do now is interpreting the output from an economic point of view.

```
Model 2: OLS, using observations 1995:3-2019:4 (T = 98)
Dependent variable: c

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        0.220191    0.0643345     3.423    0.0009   ***
  c_1          0.893634    0.0378199    23.63     4.39e-41 ***
  y            0.588653    0.0648987     9.070    1.90e-14 ***
  y_1         -0.612108    0.109440     -5.593    2.23e-07 ***
  y_2          0.110455    0.0655676     1.685    0.0954   *

Mean dependent var   13.95702   S.D. dependent var   0.096924
Sum squared resid    0.001048   S.E. of regression   0.003357
R-squared            0.998850   Adjusted R-squared   0.998800
F(4, 93)             20190.32   P-value(F)           1.0e-135
Log-likelihood       421.7841   Akaike criterion    -833.5681
Schwarz criterion   -820.6433   Hannan-Quinn        -828.3403
rho                  0.079847   Durbin's h           0.852445

Breusch-Godfrey test for autocorrelation up to order 4:
TR^2 =  3.484938, with p-value =  0.48
```

Table 5.2: Dynamic regression example

Example 5.8 (Multipliers for the Euro consumption function)

The calculation of the sequence of multipliers for the model in Table 5.2 can be undertaken by using equation (5.10); the estimates of two polynomials we need are

$$\begin{aligned} \widehat{A(L)} &= 1 - 0.893634L \\ \widehat{B(L)} &= 0.588653 - 0.612108L + 0.110455L^2, \end{aligned}$$

so in this case we have

$$d_i = 0.893634 d_{i-1} + 0.588653 u_i - 0.612108 u_{i-1} + 0.110455 u_{i-2},$$

where $u_0 = 1$ and $u_i = 0$ for $i \neq 0$. Therefore, d_0 equals

$$d_0 = 0.588653$$

while for d_1 and d_2 we have

$$\begin{aligned} d_1 &= 0.893634 \cdot d_0 - 0.612108 = -0.0860674 \\ d_2 &= 0.893634 \cdot d_1 + 0.110455 = 0.0335424, \end{aligned}$$

and so on. With a little effort (and appropriate software), you get the following results:

i	d_i	c_i
0	0.588653	0.588653
1	-0.0860674	0.502585
2	0.0335424	0.536128
3	0.0299747	0.566103
4	0.0267864	0.592889
5	0.0239372	0.616826
6	0.0213911	0.638217
⋮	⋮	⋮

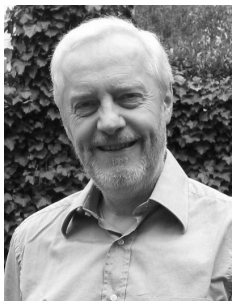
Where I also added a column for the interim (cumulated) multipliers. Moreover, you have that $A(1) = 1 - 0.893634 = 0.106366$, $B(1) = 0.087$, *and therefore the long-run multiplier equals* $\mathbf{c} = 0.87/0.106366 = 0.81793$.

5.5 The ECM representation

As I argued in section 5.2.2, the best way to interpret the parameters of an ADL model is by computing the dynamic multipliers (and possibly cumulating them). The multipliers that are presumably of most interest from an economic viewpoint are (a) the impact multiplier d_0 (because it measures what happens instantaneously) and (b) the long run-multiplier (because it measures what happens when all adjustment has taken place).

Both are easy to compute, since $d_0 = B(0)/A(0)$ and $c = B(1)/A(1)$. Nevertheless, there is a way to rewrite an ADL model in such a way that these quantities are even more evident: the so-called **ECM representation**. This device amounts, essentially, to expressing a difference equation in a slightly modified form, so that certain quantities appear more clearly. In fact, this is an example of the "reparametrisation" trick I described in section 1.4.3: since the difference equation that underlies the statistical model is exactly the same, just written in a different way.

As for what the acronym means… it's a long story. Sir David Hendry, who is considered the father of ECM (or at least, one of the fathers) is adamant on *Equilibrium Correction Mechanism*, which is probably the most precise way to express the concept. Unfortunately, this is not the original choice. Back in the day, when the phrase was introduced in the deservedly famous article by Davidson et al. (1978), the original expansion was *Error Correction Model*, and most people I know (including myself) keep using the old name.


DAVID HENDRY

To illustrate how the ECM representation works, let's start from the simple case of an ADL(1,1) (here $\mathbf{x}_t$ is a vector):

$$y_t = \alpha y_{t-1} + \beta_0' \mathbf{x}_t + \beta_1' \mathbf{x}_{t-1};$$

from the definition of the Δ operator, evidently $y_t = y_{t-1} + \Delta y_t$ and $\mathbf{x}_t = \mathbf{x}_{t-1} + \Delta\mathbf{x}_t$. After substitution,

$$\Delta y_t = (\alpha - 1) y_{t-1} + \beta_0' \Delta\mathbf{x}_t + (\beta_0 + \beta_1)' \mathbf{x}_{t-1}$$

which, after rearranging terms, yields

$$\Delta y_t = \beta_0' \Delta\mathbf{x}_t + (\alpha - 1) \left[y_{t-1} - \frac{(\beta_0 + \beta_1)'}{1 - \alpha} \mathbf{x}_{t-1} \right] \tag{5.13}$$

Which means: the time variation in y_t (on the left-hand side) may come from variation in $\mathbf{x}_t$, with response β_0 (the impact multiplier); however, even if $\Delta\mathbf{x}_t = 0$ there may be some variation in y_t if the term in square brackets is non-zero. This term can also be written as

$$y_{t-1} - \mathbf{c}'\mathbf{x}_{t-1}$$

where $\mathbf{c} = \frac{\beta_0 + \beta_1}{1-\alpha}$, that is the long-run multiplier vector. In practice, the above expression, commonly referred to as **ECM term**, gives you the difference (at $t-1$) between the actual value y_{t-1} and the value that (given $\mathbf{x}_{t-1}$) the dependent variable should have taken if the system had been in equilibrium.

If $|\alpha| < 1$, then $(\alpha - 1)$ is negative: if the ECM term is positive (so y_{t-1} was larger than its equilibrium value), then Δy_t will be negative, so y_t would tend to get closer to equilibrium. Evidently, this situation is reversed when the ECM term is negative, so if $(\alpha - 1) < 0$, the dynamic system has an inherent tendency to go back to a steady state. To be more precise, the number $1 - \alpha$ can be seen as the fraction of disequilibrium that get re-absorbed in one period, so that the closer α is to 0, the faster adjustment occurs.

You can always go from the ADL representation to the ECM representation (and back), for polynomials $A(L)$ e $B(L)$ of any order: for the algebra-loving reader, the formal proof is in section 5.A.4. In general, however, if

$$A(L) y_t = m_t + B(L)\mathbf{x}_t + \varepsilon_t$$

where the order of $A(L)$ is p and the order of $B(L)$ is q, then the ECM representation is

$$H(L)\Delta y_t = m_t + K(L)\Delta\mathbf{x}_t - A(1) y_{t-1} + B(1)\mathbf{x}_{t-1} + \varepsilon_t,$$

where the orders of $H(L)$ and $K(L)$ are $q-1$ and $p-1$, respectively. For example:

Example 5.9 (ECM Representation)
Let's take another look at the difference equation I used in example 5.7:

$$y_t = 0.2 y_{t-1} + 0.4 x_t + 0.3 x_{t-2}$$

and compute the ECM representation. The quickest way to do this is to re-express all the terms relative to time $t-1$*:*

$$\begin{aligned} y_t &= y_{t-1} + \Delta y_t \\ x_t &= x_{t-1} + \Delta x_t \\ x_{t-2} &= x_{t-1} - \Delta x_{t-1}; \end{aligned}$$

now substitute

$$y_{t-1} + \Delta y_t = 0.2y_{t-1} + 0.4(x_{t-1} + \Delta x_t) + 0.3(x_{t-1} - \Delta x_{t-1})$$

and collect

$$\Delta y_t = -0.8y_{t-1} + 0.7x_{t-1} + 0.4\Delta x_t - 0.3\Delta x_{t-1}$$

so finally

$$\Delta y_t = 0.4\Delta x_t - 0.3\Delta x_{t-1} - 0.8\left[y_{t-1} - 0.875x_{t-1}\right];$$

the impact multiplier is 0.4, the long-run multiplier is 0.875; the fraction of disequilibrium that re-adjusts each period is 0.8.

Note that the ADL model and the ECM are not two different models, but are simply two ways of expressing the same difference equation. As a consequence, you can use OLS on either and get the same residuals. The only difference between them is that the ECM forms makes it more immediate for the human eye to calculate the parameters that are most likely to be important for the dynamic properties of the model: that is the long-run multipliers and the convergence speed. On the other hand, the ADL form allows for simple (and, most importantly, *mechanical*) calculation of the whole sequence of dynamic multipliers.

```
OLS, using observations 1995:3-2019:4 (T = 98)
Dependent variable: dc

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       0.220191     0.0643345     3.423    0.0009   ***
  dy          0.588653     0.0648987     9.070    1.90e-14 ***
  dy_1       -0.110455     0.0655676    -1.685    0.0954   *
  c_1        -0.106366     0.0378199    -2.812    0.0060   ***
  y_1         0.0870005    0.0333187     2.611    0.0105   **

Mean dependent var   0.003681   S.D. dependent var   0.004886
Sum squared resid    0.001048   S.E. of regression   0.003357
R-squared            0.547335   Adjusted R-squared   0.527866
F(4, 93)             28.11251   P-value(F)           2.61e-15
Log-likelihood       421.7841   Akaike criterion    -833.5681
Schwarz criterion   -820.6433   Hannan-Quinn        -828.3403
```

Table 5.3: Dynamic regression in ECM form

Example 5.10 (ECM on real data)
The ECM representation of the model shown in table 5.2 is easily computed after

performing the following substitutions:

$$\begin{aligned} c_t &= c_{t-1} + \Delta c_t \\ y_t &= y_{t-1} + \Delta y_t \\ y_{t-2} &= y_{t-1} - \Delta y_{t-1} \end{aligned}$$

Hence,

$$\Delta c_t = k + (\alpha - 1) c_{t-1} + \beta_0 \Delta y_t + \left(\beta_0 + \beta_1 + \beta_2\right) y_{t-1} - \beta_2 \Delta y_{t-1} + \varepsilon_t,$$

that is

$$\Delta c_t = k + \beta_0 \Delta y_t - A(1) \left[c_{t-1} - \mathbf{c} y_{t-1}\right] - \beta_2 \Delta y_{t-1} + \varepsilon_t;$$

so, after substituting the estimated numerical values (and rounding results a little),

$$\Delta c_t = 0.220 + 0.589 \Delta y_t - 0.110 \Delta y_{t-1} - 0.106 \left[c_{t-1} - 0.818 y_{t-1}\right] + \varepsilon_t.$$

Note, however, that this representation could have been calculated directly by applying OLS to the model in ECM form: it is quite clear from Table 5.3 that what gets estimated is the same model in a different form. *Not only the parameters for each representation can be calculated from the other one: the objective function (the SSR) is identical for both models (and equals 0.001048); clearly, the same happens for all the statistics based on the SSR. The only differences (eg the R^2 index) come from the fact that the model is transformed in such a way that the dependent variable is not the same (it's c_t in the ADL form and Δc_t in the ECM form).*

5.6 Hypothesis tests on the long-run multiplier

In some cases, it may be of interest to test hypotheses on c, such as $H_0 : c = k$. One way to do this could be to use the estimator of c provided by the OLS estimates

$$\hat{c} = \frac{\widehat{B}(1)}{\widehat{A}(1)}$$

and then working out its asymptotic distribution, but this is complicated by the fact that $\hat{c}$ is a nonlinear function of the estimated parameters,[19] so the delta method (see section 2.3.2, particularly equation (2.14)) would be required. A much simpler way comes from observing that

$$c = k \Longleftrightarrow B(1) - k \cdot A(1) = 0$$

which is a linear test and, as such, falls under the $R\beta = \mathbf{d}$ jurisdiction.

[19] You have $\widehat{A}(1)$ in the denominator, so for example in an ADL(1,1) $c = \frac{\beta_0 + \beta_1}{1-\alpha}$, and the Jacobian term would be $J = \frac{1}{1-\alpha}[1 \quad 1 \quad c]$.

It may be worth mentioning here that tests of this type behave in the ordinary way only if the assumptions we made in section 5.3.1 are valid. There are some important cases when this may not be true, notably when the data we are working with are generated by non-stationary DGPs.

The test is particularly easy when $k = 1$, which is a common hypothesis to test, since it implies, if true, that the two variables under considerations are proportional to each other in the long run. In this case, the hypothesis becomes

$$H_0 : \alpha_1 + \cdots + \alpha_p + \beta_0 + \cdots + \beta_q = 1$$

that can be tested quite easily.

The test is even easier if you start from the estimates of the model in ECM form: all you have to do is set up a test that involves just 2 parameters, since the parameter for y_{t-1} is just $-\widehat{A}(1)$ (note the minus sign) and the parameter for x_{t-1} is $\widehat{B}(1)$.

Example 5.11

Suppose that we have the following estimates:

$$\hat{y}_t = 0.75 y_{t-1} + 0.53 x_t - 0.24 x_{t-1}$$

with the following covariance matrix:

$$\hat{V}_{ADL} = 0.001 \times \begin{bmatrix} 5 & 0.5 & -2 \\ & 5 & 4 \\ & & 5 \end{bmatrix}$$

The hypothesis $c = 1$ implies $\alpha + \beta_0 + \beta_1 = 1$. Therefore, a Wald test can be set up with $R = [1 \quad 1 \quad 1]$ and $\mathbf{d} = 1$ (see section 3.3.2 for details). Therefore

$$\begin{aligned} R\hat{\boldsymbol{\beta}} - \mathbf{d} &= [1 \quad 1 \quad 1] \begin{bmatrix} 0.75 \\ 0.53 \\ -0.24 \end{bmatrix} - 1 = 0.04 \\ R \cdot \hat{V}_{ADL} \cdot R' &= 0.001 \times 20 = 0.02 \\ W &= \frac{0.04^2}{0.02} = 0.08 \end{aligned}$$

which leads of course to accepting H_0, since its p-value is way larger than 5% ($P(\chi^2_1 > 0.08) = 0.777$). The same test could have been performed even more easily from the ECM representation:

$$\widehat{\Delta y}_t = 0.53 \Delta x_t - 0.25 y_{t-1} + 0.29 x_{t-1}$$

with the associated covariance matrix

$$\hat{V}_{ECM} = 0.001 \times \begin{bmatrix} 5 & 0.5 & 9 \\ & 5 & -1.5 \\ & & 18 \end{bmatrix}$$

In this case the hypothesis can be written as $H_0 : B(1) - A(1) = 0$*, so for the ECM form*

$$\begin{aligned} R\hat{\boldsymbol{\beta}} - \mathbf{d} &= \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.53 \\ -0.25 \\ 0.29 \end{bmatrix} = 0.04 \\ R \cdot \hat{V}_{ECM} \cdot R' &= 0.001 \times 20 = 0.02 \end{aligned}$$

and of course the W *statistic is the same as above.* ∎

5.7 Forecasting and Granger causality

One of the cool things you can do with an ADL model is forecasting. Here's how it works: suppose we have data that goes from $t = 1$ to $t = T$, and that our model of choice is an ADL(1,1). What we can say, on these premises, about y_{T+1}, that we haven't yet observed? The random variable y_{T+1} can be represented as

$$y_{T+1} = \alpha y_T + \beta_0 x_{T+1} + \beta_1 x_T + \varepsilon_{T+1}; \tag{5.14}$$

of all the objects that appear on the right-hand side of the equation, the only ones that are known with certainty at time T are y_T and x_T. Suppose we also know for certain what the future value x_{T+1} will be, and call it $x_{T+1} = \check{x}_{T+1}$. Therefore, since ε_t is a martingale difference sequence, its conditional expectation with respect to $\Im_{T+1}$ is 0,[20] so

$$\begin{aligned} \mathrm{E}\left[y_{T+1}|\Im_T\right] &= \alpha y_T + \beta_0 \check{x}_{T+1} + \beta_1 x_T \\ \mathrm{V}\left[y_{T+1}|\Im_T\right] &= \sigma^2 \end{aligned}$$

Following the same logic as in Section 3.7, we can use the conditional expectation as predictor and the estimated values for the parameters instead of the true ones. Therefore, our prediction will be

$$\hat{y}_{T+1} = \hat{\alpha} y_T + \hat{\beta}_0 \check{x}_{T+1} + \hat{\beta}_1 x_T$$

and a 95% confidence interval can be constructed as

$$\hat{y}_{T+1} \pm 1.96 \times \hat{\sigma}$$

where it is implicitly assumed that ε_t is normal and uncertainty about the parameters is ignored.

Now, there are two points I'd like to draw your attention on. First: in order to predict y_{T+1} we need x_{T+1}; but then, we could generalise this idea and imagine

[20] We're sticking to the definition of $\Im_{T+1}$ I introduced in Section 5, so $\Im_{T+1}$ includes x_{T+1} but not y_{T+1}; later in this section, we'll use a different convention.

that we could make guesses about $x_{T+2}, x_{T+3}, \ldots$ as well. What keeps us from predicting y_t farther into the future? To cut a long story short, performing multi-step forecasts is rather easy if you use your own predictions in lieu of the future values of y_t and proceed recursively. In other words, once you have $\hat{y}_{T+1}$ you can push equation (5.14) one step ahead in time and write

$$y_{T+2} = \alpha y_{T+1} + \beta_0 x_{T+2} + \beta_1 x_{T+1} + \varepsilon_{T+2};$$

next, we operate in a similar way as we just did, using the conditional expectation as predictor

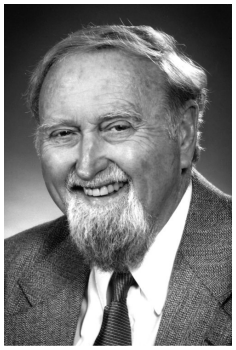
$$\hat{y}_{T+2} = \hat{\alpha}\hat{y}_{T+1} + \hat{\beta}_0 \check{x}_{T+2} + \hat{\beta}_1 \check{x}_{T+1},$$

repeating the process with the obvious adjustments for $T+3$, $T+4$ etc. It can be proven (nice exercise for the reader) that the variance you should use for constructing confidence interval for multi-step forecasts would be in this case

$$\mathrm{V}\left[\hat{y}_{T+k}\right] = \frac{1-(\alpha^2)^k}{1-\alpha^2}\sigma^2.$$

Extending the formulae above to the general ADL(p,q) case is trivial but boring, and I'll just skip it.

The second point I want to make comes by considering the possibility that the β_0 and β_1 coefficients were 0 in equation (5.14). In this case, there would be no need to conjecture anything about x_t, in order to forecast y_t. In other words, x_t has no predicting power for y_t. This is a hypothesis we may want to test.


CLIVE GRANGER

As a general rule, in the context of dynamic regression models, it is difficult to formulate hypotheses of economic interest that can be tested through restrictions on coefficients, since the coefficients of the $A(L)$ and $B(L)$ polynomials normally don't have a natural economic interpretation *per se*, and this is why we compute multipliers.

However, there are exceptions: we just saw one of them in the previous section. Another one is the so-called **Granger-causality** test, after the great Clive Granger, Nobel Prize winner in 2003.[21] The idea on which the test is built is that, whenever A causes B, the cause should come, in time, before the effect. Therefore, if A does *not* cause B, it should have no effect on the quantity we normally use for prediction, i. e. the conditional expectation.

The only difference with the ADL models we've considered so far is that, since we're dealing with predictions about the future, we will want to base our

[21] C. W. Granger is one of the founding fathers of modern time series econometrics; his most famous brainchild, that earned him the Nobel Prize, is a concept called *cointegration*, that I will skip in this book, but is absolutely indispensable if you want to engage in applied macroeconomics.

inference on an information set that collects everything that is known at time $t-1$, namely

$$\Im^*_{t-1} = \{y_{t-1}, y_{t-2}, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots\};$$

note that, contrary to the concept of information set $\Im_t$ we used so far (defined in section 5.1), $\Im^*_t$ does *not* include $\mathbf{x}_{t+1}$; in practice, it collects all information on y_t and $\mathbf{x}_t$ that is available up to time t. Forecasting, therefore, amounts to finding

$$\hat{y}_{T+1|T} = \mathrm{E}\left[y_{T+1}|\Im^*_T\right].$$

The subscript "$T+1|T$" is customarily read as "at time $T+1$, based on the information available at time T".

There is no doubt that the discerning reader has spotted, by now, a fundamental difference between the information set $\Im^*_{T-1}$ that we are using here and the information set $\Im_T$ we use in the rest of this chapter: the latter includes $\mathbf{x}_t$, while the former does not.
Since $\Im^*_{T-1} \subset \Im_T$, predictions on y_t made using $\Im^*_{T-1}$ are obviously going to be less accurate, but have the advantage of being possible one period earlier. Moreover, this makes also possible to forecast $\hat{\mathbf{x}}_{T+1|T} = \mathrm{E}\left[\mathbf{x}_{T+1|T}|\Im^*_T\right]$. This seemingly innocent remark paves the way to multi-step forecasts, where we use the predictions for T to forecast $T+1$, which in turn we use for forecasting $T+2$, and so on.

This is the principle used in the so-called **VAR model**, which is probably the main empirical tool in modern macroeconometrics. If you're curious, check out Lütkepohl (2005).

As a consequence, our ADL model

$$A(L)y_t = B(L)x_t + \varepsilon_t$$

will not include $\mathbf{x}_t$, but only its lags:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \boldsymbol{\beta}_1' \mathbf{x}_{t-1} + \boldsymbol{\beta}_2' \mathbf{x}_{t-2} + \cdots + \varepsilon^*_t$$

where ε^*_t is defined as $y_t - \mathrm{E}\left[y_{T+1}|\Im^*_T\right]$. Clearly, this can also be written as an ordinary ADL model in which $B(0) = 0$. The idea that x_t does not cause y_t is equivalent to the idea $B(L) = 0$; since a polynomial is 0 if and only if all its coefficients are, it is easy to formulate the hypothesis of no-Granger-causality as

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$$

which is of course a system of linear restrictions, that we can handle just fine via the $R\boldsymbol{\beta} = \mathbf{d}$ machine we described in Section 3.3.2.

In the late 1960s, when this idea was introduced, it was hailed as a breakthrough in economic theory, because for a while this seemed to provide a data-based way to ascertain causal links. For example, a hotly debated point among macroeconomists in the 1970 and 80s was: is there a causality direction between the quantity of money and GDP in an economy? If there is, the repercussions on economic policy (notably, on the effectiveness of monetary policy) are huge. In

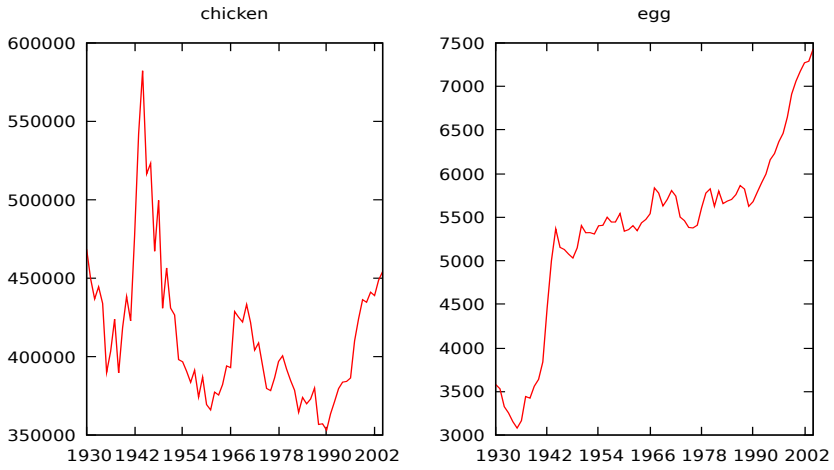


Figure 5.3: Thurman & Fisher data on chickens and eggs

those days, the concept of Granger-causality seemed to provide a convincing answer.[22]

Example 5.12

In a humorous article, Thurman and Fisher (1988) collected data on the production of chickens and eggs from 1930 to 2004, that are depicted in Figure 5.3.

After taking logs, we estimate by OLS the following 2 equations:

$$c_t = m_1 + \alpha_1 c_{t-1} + \alpha_2 c_{t-2} + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \varepsilon_t \tag{5.15}$$

$$e_t = \mu_1 + \gamma_1 c_{t-1} + \gamma_2 c_{t-2} + \lambda_1 e_{t-1} + \lambda_2 e_{t-2} + \eta_t \tag{5.16}$$

where c_t is the log of chickens at time t and e_t is the log of eggs.

The hypothesis that chickens don't Granger-cause eggs is $H_0 : \gamma_1 = \gamma_2 = 0$; the opposite hypothesis, that eggs don't Granger-cause chickens is $H_1 : \beta_1 = \beta_2 = 0$. The results are in Table 5.4. As can be seen, the hypothesis of absence of Granger-causality is rejected in the egg → chicken direction, but not the other way round; hence, the perennial question "what comes first?" has finally found an answer: it's the egg that comes first.

There are a few issues that may be raised here: one is statistical, and pertains to the fact that the test is relative to a certain conditioning set. You may see this as a variation on the same theme I discussed in Section 3.8, especially example 3.3. It may well be that A turns out to be Granger-causal for B in a model, and the reverse happens in another model, in which some other variables are included

[22] Readers who are into the history of economics and econometrics might want to take a look at Sims (1972).

Dependent variable: l_chicken

	Coefficient	Std. Error	t-ratio	p-value
const	2.1437	0.7715	2.7788	0.0070
l_chicken_1	0.4037	0.1389	2.9054	0.0049
l_chicken_2	0.4362	0.1320	3.3037	0.0015
l_egg_1	0.8627	0.2011	4.2906	0.0001
l_egg_2	−0.8724	0.1999	−4.3642	0.0000

Mean dependent var	12.92290	S.D. dependent var	0.100867
Sum squared resid	0.125953	S.E. of regression	0.043038
R^2	0.828060	Adjusted R^2	0.817946

Granger-causality test egg → chicken: $F(2,68) = 9.57089$, p-value = 0.000217573

Dependent variable: l_egg

	Coefficient	Std. Error	t-ratio	p-value
const	0.8816	0.5292	1.6660	0.1003
l_chicken_1	−0.1196	0.0953	−1.2548	0.2139
l_chicken_2	0.0695	0.0906	0.7673	0.4456
l_egg_1	1.5302	0.1379	11.0961	0.0000
l_egg_2	−0.5570	0.1371	−4.0624	0.0001

Mean dependent var	8.577181	S.D. dependent var	0.204279
Sum squared resid	0.059259	S.E. of regression	0.029520
R^2	0.980277	Adjusted R^2	0.979117

Granger-causality test chicken → egg: $F(2,68) = 1.28907$, p-value = 0.282174

Table 5.4: Granger causality tests between chickens and eggs

or excluded. This is why, in some cases, people perform Granger-causality tests on models in which the only variables considered are the ones that come directly into play. I'll leave it to the reader to judge whether this approach leads to results that have a sensible statistical interpretation.

Another one is more substantial in nature, and has to do with the fact that in economics it may well be that the cause comes *after* the effect, because expectations play a major role in human behaviour; people may do something at a certain time in view of something that they expect to happen in the future. In fact, standard economic theory assumes that agents are rational and forward-looking: they base *all* their choices on the expectations they have about the future.

There are many examples I could give you, but I'll simply hint at a widely used one: if people anticipate that a company is going to go bust, everyone will sell that stock, causing its price to drop. If one should mechanically assess causality from time precedence, it would be legitimate to say that the drop in the stock price drove the company bankrupt, rather than the other way around. The problem here is that in this case the statistical concept of Granger causality does not agree very much with the notion of causality we use in everyday life (and is arguably what we care about in economics). In fact, it is much more accurate to consider the Granger-causality test as a device for assessing predictive power; whether predictive power can be considered a sign of a causal chain depends on the circumstances.

5.A Assorted results

5.A.1 Inverting polynomials

Let us begin by noting that, for any $a \neq 1$,

$$\sum_{i=0}^{n} a^i = \frac{1-a^{n+1}}{1-a}, \tag{5.17}$$

which is easy to prove: call

$$S = \sum_{i=0}^{n} a^i = 1 + a + a^2 + \cdots + a^n; \tag{5.18}$$

of course

$$a \cdot S = a + a^2 + \cdots + a^{n+1} \tag{5.19}$$

and therefore, by subtracting (5.19) from (5.18), $S(1-a) = 1 - a^{n+1}$, and hence equation (5.17).

If a is a small number ($|a| < 1$), then $a^n \to 0$, and therefore $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$. By setting $a = \alpha L$, you may say that, for $|\alpha| < 1$, the inverse of $(1-\alpha L)$ is $(1 + \alpha L + \alpha^2 L^2 + \cdots)$, that is

$$(1-\alpha L)(1 + \alpha L + \alpha^2 L^2 + \cdots) = 1,$$

or, alternatively,

$$\frac{1}{1-\alpha L} = \sum_{i=0}^{\infty} \alpha^i L^i$$

provided that $|\alpha| < 1$. Now consider a n-th degree polynomial $P(x)$:

$$P(x) = \sum_{j=0}^{n} p_j x^j$$

If $P(0) = p_0 = 1$, then $P(x)$ can be written as the product of n first-degree polynomials as follows:[23]

$$P(x) = \prod_{j=1}^{n} \left(1 - \frac{1}{\lambda_j} x\right) \tag{5.20}$$

where the numbers λ_j are the roots of $P(x)$: if $x = \lambda_j$, then $1 - \frac{1}{\lambda_j} x = 0$ and consequently $P(x) = 0$. Therefore, if $P(x)^{-1}$ exists, it must satisfy

$$\frac{1}{P(x)} = \prod_{j=1}^{n} \left(1 - \frac{1}{\lambda_j} x\right)^{-1};$$

but if at least one of the roots λ_j is smaller than 1 in modulus,[24] then $1/|\lambda_j|$ is larger than 1 and, as a consequence, $\left(1 - \frac{1}{\lambda_j} x\right)^{-1}$ does not exist, and neither does $P(x)^{-1}$.

Example 5.13

Consider the polynomial $A(x) = 1 - 1.2x + 0.32x^2$; is it invertible? Let's check its roots:

$$A(x) = 0 \Longleftrightarrow x = \frac{1.2 \pm \sqrt{1.44 - 1.28}}{0.64} = (1.2 \pm 0.4)/0.64$$

so $\lambda_1 = 2.5$ and $\lambda_2 = 1.25$. Both are larger than 1 in modulus, so the polynomial is invertible. Specifically,

$$A(x) = (1 - \lambda_1^{-1} x)(1 - \lambda_2^{-1} x) = (1 - 0.4x)(1 - 0.8x)$$

and

$$\begin{aligned} \frac{1}{A(x)} &= (1 - 0.4x)^{-1}(1 - 0.8x)^{-1} = (1 + 0.4x + 0.16x^2 + \cdots)(1 + 0.8x + 0.64x^2 + \cdots) \\ &= 1 + 1.2x + 1.12x^2 + 0.96x^3 + 0.7936x^4 + \cdots \end{aligned}$$

[23] If you don't believe me, google for "Fundamental theorem of algebra".

[24] Warning: the roots may be complex, but this is not particularly important. If z is a complex number of the form $z = a + bi$ (where $i = \sqrt{-1}$), then $|z| = \sqrt{a^2 + b^2}$.

In practice: if the sequence a_t is defined as the result of the application of the operator $P(L)$ to the sequence u_t, that is $a_t = P(L)u_t$, then reconstructing the sequence u_t from a_t is only possible if $P(L)$ is invertible. In this case,

$$u_t = P(L)^{-1}a_t = \frac{1}{P(L)}a_t.$$

5.A.2 Basic concepts on stochastic processes

This section is just meant to give you a rough idea of some of the concept I hinted at in section 5.3.1; if you want the real thing, go for Brockwell and Davis (1991).

Suppose you have an infinitely long sequence of random variables

$$\ldots, x_{t-1}, x_t, x_{t+1}, \ldots$$

where the index t is normally taken to mean "time" (although not necessarily). This sequence is a **stochastic process**.[25] When we observe a time series, we observe a part of the realisation of a stochastic process (also called a *trajectory* of the process). Just in the same way as the DGP for the toss of a coin can be thought of as the machine that nature uses for giving us a binary number that we cannot predict, a stochastic process is a machine that nature uses for giving us an infinitely long trajectory through time, and what we observe is just a short segment of it. This idea may be unintuitive at start (it certainly was for me, back in the day), but I find it very useful.

If we take two different elements of the sequence, say x_s and x_t (with $s \neq t$), we could wonder what their joint distribution is. The two fundamental properties of the joint distribution that we are interested in are:

1. is the joint distribution stable through time? That is, is the joint distribution of (x_s, x_t) the same as (x_{s+1}, x_{t+1}) ?

2. Is it likely that x_s and x_t become independent (or nearly so) if $|t-s|$ is large?

Property number 1 refers to the idea that the point in time when we observe the process should be irrelevant: the probability distribution of the data we see today (x_s, x_t) should be the same as the one for an observer in the past (x_{s-100}, x_{t-100}) or in the future (x_{s+100}, x_{t+100}). This gives rise to the concept of **stationarity**. A stochastic process is said to be **weakly stationary**, or **covariance stationary**, or **second-order stationary** if the covariance between x_s and x_t (also known as **autocovariance**) exists and is independent of time. In formulae:

$$\gamma_h = \text{Cov}[x_t, x_{t+h}]$$

[25] It's not inappropriate to think of stochastic processes as infinite-dimensional random variables. Using the same terminology as in section 2.2.1, we may think of the sequence $\ldots, x_{t-1}(\omega), x_t(\omega), x_{t+1}(\omega), \ldots$ as the infinite-dimensional outcome of *one* point in the state space $\omega \in \Omega$.

note that γ_h, the autocovariance of order h, is a function of h only, not of t; of course, γ_0 is just $\mathrm{V}[x_t]$. If this is the case, the internal structure of correlation between points in time is often described via the **autocorrelation** sequence (or autocorrelation function, often abbreviated as **ACF**), defined as

$$\rho_h = \frac{\gamma_h}{\gamma_0}.$$

Property number 2, instead, is what we realistically imagine should happen when we observe many phenomena through time: if s and t are very far apart, what happened at time s one should contain little or no information on what happened at time t. For example: the temperature at Cape North on May 29th, 1453 at 12am should contain no useful information on the temperature at Cape North *right now*. This intuition can be translated into maths in a number of different ways. A common one is **ergodicity**. While a formal definition of ergodicity would require a hefty investment in measure theory, if a process is covariance stationary, ergodicity amounts to *absolute summability* of its autocovariances. The property

$$\sum_{i=0}^{\infty} |\gamma_i| = M < \infty$$

ensures that $\lim_{h\to\infty} |\gamma_h| = 0$ (so correlation between distant events should be negligible), but most importantly, that the sample mean of an observed stochastic process is a consistent estimator of the true mean of the process:

$$\frac{1}{T}\sum_{t=1}^{T} x_t \xrightarrow{\mathrm{p}} \mathrm{E}[x_t].$$

Note that the above expression can be considered as one of the many versions of the Law of Large Numbers, applicable when observations are not necessarily independent.

It goes without saying that, in the same way as we can define multivariate random variables, it is perfectly possible to define multivariate stochastic processes, that is, sequences of random vectors: modern macroeconometrics is primarily built upon these objects. A large part of the statistical analysis of time series is based on the idea that the time series we observe are realisations of stationary and ergodic processes (or can be transformed to this effect).

How do you adapt statistical inference to such a context? The main idea underlying most approaches is to describe the a DGP in such a way that the whole autocovariance structure of a stochastic process (the sequence $\gamma_0, \gamma_1, \gamma_2, ...$) can be expressed as a function of a finite set of parameters θ; if the process is stationary and ergodic, then maybe the available data $x_1, \ldots, x_T$ can be used to construct CAN estimators of θ. **ARIMA models** are one of the most celebrated instances of this approach, and the literature that has developed after their introduction in the late 1960s is truly gigantic. If you're interested, Brockwell and Davis (1991) is an excellent starting point.

5.A.3 Why martingale difference sequences are serially uncorrelated

Here's a rapid proof: if ε_t is a MDS with respect to $\Im_t$, then

$$\mathrm{E}[\varepsilon_t|\Im_t] = \mathrm{E}\left[\varepsilon_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, y_{t-1}, y_{t-2}, \ldots\right] = 0.$$

Now, observe that ε_{t-1} is defined as $\varepsilon_{t-1} = y_{t-1} - \mathrm{E}\left[y_{t-1}|\Im_{t-1}\right]$; since both elements of the right-hand side of the equation are contained in $\Im_t$, then $\varepsilon_{t-1} \in \Im_t$; moreover, $\Im_t \supseteq \Im_{t-1} \supseteq \Im_{t-2} \ldots$, so clearly all lags of ε_t are all contained in $\Im_t$ (see footnote 6 in this chapter). As a consequence, we can use the law of iterated expectations (2.8) as follows:

$$\begin{aligned}
\mathrm{Cov}[\varepsilon_t, \varepsilon_{t-k}] &= \mathrm{E}[\varepsilon_t \cdot \varepsilon_{t-k}] = \mathrm{E}[\mathrm{E}[\varepsilon_t \cdot \varepsilon_{t-k}]|\Im_t] = \\
&= \mathrm{E}[\mathrm{E}[\varepsilon_t|\Im_t] \cdot \varepsilon_{t-k}] = \mathrm{E}[0 \cdot \varepsilon_{t-k}] = \\
&= 0
\end{aligned}$$

A second argument, perhaps more intuitive, rests directly on the definition of a MDS: if ε_t is a MDS with respect to $\Im_t$, then its expectation conditional on $\Im_{t-k}$ (for $k > 0$) must also be 0, because $\Im_{t-k}$ is a subset of $\Im_t$. But that means that the expectation of any *future* element ε_{t+k} conditional on the present information set $\Im_t$ is 0. In formulae:

$$\left.\begin{array}{r} \mathrm{E}[\varepsilon_t|\Im_t] = 0 \\ \Im_{t-k} \subseteq \Im_t \text{for } k > 0 \end{array}\right\} \Longrightarrow \mathrm{E}[\varepsilon_t|\Im_{t-k}] = \mathrm{E}[\varepsilon_{t+k}|\Im_t] = 0$$

This is tantamount to saying that ε_t is effectively unpredictable. But then, if $\mathrm{Cov}[\varepsilon_t, \varepsilon_{t-k}] \neq 0$, ε_t wouldn't be totally unpredictable, because there would be some information in the past about the future. Therefore, the autocorrelations of a MDS must be 0 for any k.

5.A.4 From ADL to ECM

Let's begin with a preliminary result (which I'm not going to prove):

If $P(x)$ is a polynomial whose degree is $n > 0$ and a is a scalar, you can always find a polynomial $Q(x)$, whose degree is $(n-1)$, such that

$$P(x) = P(a) + Q(x)(a - x);$$

if $n = 0$, obviously $Q(x) = 0$.

For example, the reader is invited to check that, if we choose $a = 1$, the polynomial $P(x) = 0.8x^2 - 1.8x + 1.4$ can be written as

$$P(x) = 0.4 + (1 - 0.8x)(1 - x)$$

where $P(1) = 0.4$ and $Q(x) = 1 - 0.8x$.

Now consider $P(L)$, a polynomial in the lag operator of degree $n \geq 1$, and apply the result above twice in a row, once with $a = 0$ and then with $a = 1$:

$$P(L) = P(0) - Q(L) \cdot L \tag{5.21}$$
$$Q(L) = Q(1) + P^*(L)(1 - L) \tag{5.22}$$

If $n = 1$, evidently $P^*(L) = 0$. Otherwise, the order of $Q(L)$ is $(n-1)$ and the order of $P^*(L)$ is $(n-2)$. If you evaluate equation (5.21) in $L = 1$, you have $P(1) = P(0) - Q(1)$, so that equation (5.22) becomes

$$Q(L) = P(0) - P(1) + P^*(L)(1 - L)$$

and therefore, using equation (5.21) again,

$$P(L) = P(0) - \left[P(0) - P(1) + P^*(L)(1 - L)\right] \cdot L = P(0)\Delta + P(1)L - P^*(L)\Delta \cdot L.$$

The actual form of the $P^*(L)$ polynomial is not important: all we need is knowing that it exists, so that the decomposition of $P(L)$ we just performed is always possible. As a consequence, every sequence $P(L)z_t$ can be written as:

$$P(L)z_t = P(0)\Delta z_t + P(1)z_{t-1} - P^*(L)\Delta z_{t-1}.$$

Now apply this result to both sides of the ADL model $A(L)y_t = B(L)\mathbf{x}_t + \varepsilon_t$:

$$\Delta y_t + A(1)y_{t-1} - A^*(L)\Delta y_{t-1} = B(0)\Delta \mathbf{x}_t + B^*(L)\Delta \mathbf{x}_{t-1} + B(1)\mathbf{x}_{t-1} + \varepsilon_t;$$

(note that $A(0) = 1$ by construction). After rearranging terms, you obtain the ECM representation proper:

$$\Delta y_t = B(0)\Delta \mathbf{x}_t + A^*(L)\Delta y_{t-1} + B^*(L)\Delta \mathbf{x}_{t-1} - A(1)\left[y_{t-1} - \mathbf{c}'\mathbf{x}_{t-1}\right] + \varepsilon_t$$

where $\mathbf{c}' = \frac{B(1)}{A(1)}$ contains the long-run multipliers. In other words, the variation of y_t over time is expressed as the sum of three components:

1. the external unpredictable shock ε_t;
2. a short-run transitory component: $B(0)\Delta \mathbf{x}_t + A^*(L)\Delta y_{t-1} + B^*(L)\Delta \mathbf{x}_{t-1}$; the first coefficient, $B(0)$, gives you the instantaneous effect of $\mathbf{x}_t$ on y_t;
3. a long-run component whose base ingredient is the long-run multiplier $\mathbf{c}$.

Chapter 6

Instrumental Variables

The arguments I presented in chapter 3 should have convinced the reader that OLS is an excellent solution to the problem of estimating linear models of the kind

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is defined as $\mathbf{y} - \mathrm{E}[\mathbf{y}|\mathbf{X}]$, with the appropriate adjustments for dynamic models; the derived property $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$ is the key ingredient for guaranteeing consistency of OLS as an estimator of $\boldsymbol{\beta}$. With some extra effort, we can also derive asymptotic normality and have all the hypothesis testing apparatus at our disposal.

In some cases, however, this is not what we need. What we have implicitly assumed so far is that the parameters of *economic* interest are the same as the *statistical* parameters that describe the conditional expectation (or functions thereof, like for example marginal effects or multipliers in dynamic models).

Sometimes, this might not be the case. As anticipated in section 3.6, this happens when the model we have in mind contains explanatory variables that, in common economics parlance, are said to be **endogenous**. In the next section, I will give you a few examples where the quantities of interpretative interest are not computable from the regression parameters. Hence, it should come as no surprise that OLS is not a usable tool for this purpose: this is why we'll want to use a different estimator, known as **instrumental variables** estimator, or **IV** for short.

6.1 Examples

6.1.1 Measurement error

Measurement error is what you get when one or more of your explanatory variable are measured imperfectly. Suppose you have the simplest version of a linear model, where everything is a scalar:

$$y_i = x_i^* \beta + \varepsilon_i \tag{6.1}$$

where $\mathrm{E}\left[y_i|x_i^*\right] = x_i^*\beta$ and β is our parameter of interest. The problem is that we do not observe x_i^* directly; instead, all we have is a version of x_i^* that is contaminated by some measurement error:

$$x_i = x_i^* + \eta_i \tag{6.2}$$

where η_i is a zero-mean random variable, independent of x_i^* and ε_i, with variance $\sigma_\eta^2 > 0$; clearly, the larger $\sigma_\eta^2 > 0$ is, the worse is the quality of our measurement for the variable of interest x_i^*. One may think that, since η_i is, so to speak, "neutral", setting up a model using x_i instead of x_i^* would do no harm. Instead, this is not the case: unfortunately, OLS regression y_i on x_i won't give you a consistent estimator of β. This is quite easy to prove: combine the two equations above to get

$$y_i = x_i\beta + \left(\varepsilon_i - \beta\eta_i\right) = x_i\beta + u_i \tag{6.3}$$

so

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i(x_i\beta + u_i)}{\sum_{i=1}^n x_i^2} = \beta + \frac{\sum_{i=1}^n x_i u_i}{\sum_{i=1}^n x_i^2}$$

From the assumptions above, you get

$$\begin{aligned} \mathrm{E}\left[x_i u_i\right] &= \mathrm{E}\left[(x_i^* + \eta_i)(\varepsilon_i - \beta\eta_i)\right] = \mathrm{E}\left[x_i^*\varepsilon_i\right] - \beta\mathrm{E}\left[x_i^*\eta_i\right] + \mathrm{E}\left[\eta_i\varepsilon_i\right] - \beta\mathrm{E}\left[\eta_i^2\right] = \\ &= -\beta\sigma_\eta^2 \end{aligned}$$

If we define $Q = \mathrm{E}\left[x_i^2\right]$, clearly

$$\hat{\beta} \xrightarrow{\mathrm{p}} \beta - \frac{\beta\sigma_\eta^2}{Q} = \beta\left(1 - \frac{\sigma_\eta^2}{Q}\right) \neq \beta$$

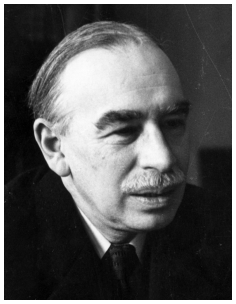
It can be proven that $0 < \sigma_\eta^2 < Q$,[1] so two main conclusions can be drawn from the equation above: first, the degree of inconsistency of OLS is proportional to the size of measurement error σ_η^2 relative to Q; second, the asymptotic bias is such that $|\mathrm{plim}\left(\hat{\beta}\right)| < |\beta|$; that is, the estimated effect is smaller than the true one. This is often called **attenuation**.

As the rest of this chapter should hopefully make clear, the reason why OLS doesn't work as we'd like it to work lies in the fact that equation (6.3) does not split y_i into a conditional expectation and a disturbance term. It can be shown that the regression function $\mathrm{E}\left[y_i|x_i\right]$ is not equal to $x_i\beta$: if it were, $\mathrm{E}\left[x_i u_i\right]$ would be zero, but we just showed it isn't. Since OLS is programmed to estimate the parameters of a conditional expectation, you can't expect it to come up with anything else.

[1] Come on, it's easy, do it by yourself.

This argument came out as important in an economic theory controversy in the 1950s about the consumption function. In those days, orthodoxy was Keynes' idea that

> [T]he fundamental psychological law [...] is that men are disposed, as a rule and on the average, to increase their consumption as their income increases but not by as much as the increase in the income.[2]



JOHN MAYNARD KEYNES

In formulae, this was translated as

$$C = C_0 + cY.$$

with $0 < c < 1$. As the reader knows, c is the "marginal propensity to consume", that is a key ingredient in mainstream Keynesian macroeconomics.

In the 1950s, few people would dissent from the received wisdom: one of them was Milton Friedman, who would argue that c should not be less than 1 (at least in the long run), since the only thing income is good for is buying things. Over the span of your life, it would be silly to save money unconditionally: a rational individual with perfect foresight should die penniless.[3]

Back in the day, economists thought of measuring c by running regressions on the consumption function, and regularly found estimates that were significantly smaller than 1. Friedman, however, put forward a counter-argument, based on the "permanent income" concept: consumption is not based on current income, but rather on a concept of income that takes into account your expectations about the future. For example, if you knew with certainty that you're going to inherit a disgustingly large sum from a moribund distant uncle, you would probably start squandering money today (provided, of course, you find somebody willing to lend you money), far beyond your level of *current* income.

In this case, your observed actual income x_i does not coincide with your permanent income x_i^* (which is unobservable), and estimated values of c lower than 1 could well be the product of attenuation.

6.1.2 Simultaneous equation systems

Simultaneous equation systems make for another nice example. Inclined as I am to put econometric concepts in a historical context, I would love to inflict on the reader a long, nostalgic account about the early days of econometrics, the great Norwegian pioneer Trygve Haavelmo and Lawrence Klein[4] and the Cowles Commission, but this is not the place for it. Suffice it to say that estimation of

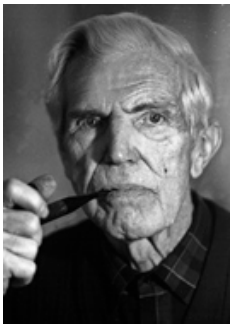
[2] Keynes, J.M. (1936) *The General Theory of Employment, Interest and Money*

[3] "Avaritia vero senilis quid sibi velit, non intellego; potest enim quicquam esse absurdius quam, quo viae minus restet, eo plus viatici quaerere?" Marcus Tullius Cicero, *De senectute*.

[4] Klein and Haavelmo got the Nobel Prize for their work in 1980 and 1989, respectively.

systems of equations is the first autonomous contribution of econometrics to the general arsenal of statistical tools.

The reason why the estimation of parameters in simultaneous systems may be tricky is relatively easy to see by focusing on the distinction between parameters of interest and parameters of the conditional mean.



TRYGVE HAAVELMO

Consider one of the simplest examples of simultaneous equation system in economics: a two-equation linear model of supply and demand for a good.

$$q_t = \alpha_0 - \alpha_1 p_t + u_t \tag{6.4}$$
$$p_t = \beta_0 + \beta_1 q_t + v_t, \tag{6.5}$$

equation (6.4) is the demand equation (quantity at time t as a function of price at time t), (6.5) is supply (price as a function of quantity); the two disturbance terms u_t and v_t represent random shocks to the two curves. For example, u_t could incorporate random fluctuations in demand due to shifting customer preferences, fluctuations in disposable income and so forth; v_t, instead could be non-zero because of productivity shifts due to random events (think for example weather for agricultural produce). Assume that $\mathrm{E}[u_t] = \mathrm{E}[v_t] = 0$.

If you considered the two equations separately, one may think of estimating their parameters by using OLS, but this would be a big mistake, since the "systematic part" of each of the two equations is not a conditional expectation.

An easy way to convince yourself is simply to consider that if $\mathrm{E}[q_t|p_t]$ is upward (downward) sloping, the correlation between q_t and p_t must be positive (negative), and therefore there's no way the reverse conditional expectation $\mathrm{E}[p_t|q_i]$ can be downward (upward) sloping. Since the demand function goes down and the supply function goes up, at least one of them cannot be a conditional expectation.

However, a more rigorous proof can be given: take the demand curve (6.4): if the expression $(\alpha_0 - \alpha_1 p_t)$ were in fact the conditional expectation of q_t to p_t, then $\mathrm{E}[u_t|p_t]$ should be 0. Now substitute (6.4) into equation (6.5):

$$\begin{aligned} p_t &= \beta_0 + \beta_1(\alpha_0 - \alpha_1 p_t + u_t) + v_t \\ &= (\beta_0 + \beta_1\alpha_0) - (\beta_1\alpha_1)p_t + (v_t + \beta_1 u_t) \Rightarrow \\ (1 + \beta_1\alpha_1)p_t &= (\beta_0 + \beta_1\alpha_0) + (v_t + \beta_1 u_t) \Rightarrow \qquad (6.6) \\ p_t &= \pi_1 + \eta_t, \qquad (6.7) \end{aligned}$$

where the constant π_1 is $\frac{\beta_0+\beta_1\alpha_0}{1+\beta_1\alpha_1}$ and $\eta_t = \frac{v_t+\beta_1 u_t}{1+\beta_1\alpha_1}$ is a zero-mean random variable. The covariance between p_t and u_t is easy to compute:

$$\begin{aligned} \mathrm{Cov}[p_t, u_t] &= \mathrm{E}[p_t \cdot u_t] = \mathrm{E}[u_t\pi_1 + u_t\eta_t] = 0 + \mathrm{E}[u_t \cdot (v_t + \beta_1 u_t)] = \\ &= \mathrm{Cov}[v_t, u_t] + \beta_1 V(u_t) \end{aligned}$$

Now, unless the covariance between v_t and u_t happens to be *exactly equal* to $-\beta_1 V(u_t)$ (and there is no reason why it should), $\mathrm{Cov}\left[p_t, u_t\right] \neq 0$. Borrowing on the definition I gave in Section 3.6, the variable p_t is clearly endogenous.

But if $\mathrm{Cov}\left[p_t, u_t\right] \neq 0$, then $\mathrm{E}\left[u_t | p_t\right]$ can't be 0 either; therefore, $(\alpha_0 - \alpha_1 p_t)$ can't be $\mathrm{E}\left[q_t | p_t\right]$, and as a consequence there's no way that OLS applied to equation (6.4) (that is, regressing quantity on a constant and price) could be a consistent estimator of α_0 and α_1.

To be more specific: even assuming that $\mathrm{E}\left[q_t | p_t\right]$ is a linear function like

$$\mathrm{E}\left[q_t | p_t\right] = \gamma_0 + \gamma_1 p_t,$$

OLS gives you an excellent estimate of the coefficients γ_0 and γ_1; unfortunately, they are not the same thing as α_0 and α_1.

Of course, the same argument in reverse could be applied to the supply equation so regressing p_t on q_t won't give you good estimates of β_0 and β_1, either. This example will be generalised in section 6.6.2.

6.2 The IV estimator

In a standard linear model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$. As we argued in chapter 3, the assumption $\mathbf{x}'\boldsymbol{\beta} = \mathrm{E}\left[y | \mathbf{x}\right]$ is crucial for the consistency of the OLS statistic as an estimator of $\boldsymbol{\beta}$; in fact, you could see this assumption as a *definition* of $\boldsymbol{\beta}$, in that $\boldsymbol{\beta}$ is the only vector for which the following equation is true:

$$\mathrm{E}\left[\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] = \mathbf{0}. \tag{6.8}$$

The OLS statistic $\hat{\boldsymbol{\beta}}$, instead, is implicitly defined by the relationship

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{6.9}$$

which corresponds to the first-order conditions for the minimisation of the sum of squared residuals (see section 1.3.2, especially equation (1.10)); note that equation (6.9) can be seen as the sample equivalent of equations (6.8). The fact that the OLS statistic $\hat{\boldsymbol{\beta}}$ works quite nicely as an estimator of its counterpart $\boldsymbol{\beta}$ just agrees with common sense.

If, on the contrary, the parameter of interest $\boldsymbol{\beta}$ satisfies an equation other than (6.8), then we may proceed by analogy and use, as an estimator, a statistic $\tilde{\boldsymbol{\beta}}$ that satisfies the corresponding sample property. In this chapter, we assume we have a certain number of observable variables $\mathbf{W}$ for which

$$\mathrm{E}\left[\mathbf{W}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] = \mathbf{0}. \tag{6.10}$$

The corresponding statistic will then be implicitly defined by

$$\mathbf{W}'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{0}. \tag{6.11}$$

The variables $\mathbf{W}$ are known as **instrumental variables**, or, more concisely, **instruments**. The so-called "simple" IV estimator can then be defined as follows: if we had a matrix $\mathbf{W}$, of the same size as $\mathbf{X}$, satisfying (6.10), then we may define a statistic $\tilde{\boldsymbol{\beta}}$ such that (6.11) holds:

$$\mathbf{W}'(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})=\mathbf{0} \qquad \Longrightarrow \qquad \mathbf{W}'\mathbf{X}\cdot\tilde{\boldsymbol{\beta}}=\mathbf{W}'\mathbf{y}.$$

In a parallel fashion, the difference $\boldsymbol{\varepsilon}=\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$ is not defined as the difference between $\mathbf{y}$ and it conditional mean, but rather as a zero-mean random variable which describes how much $\mathbf{y}$ deviates from its "standard" value, as described by the "structural" relationship $\mathbf{X}\boldsymbol{\beta}$. A term that we use in this context is **structural disturbance**, or just "disturbance" when no confusion arises. Given this definition, there is no guarantee that the structural disturbance should be orthogonal to the regressors.

Since $\mathbf{W}$ has as many columns as $\mathbf{X}$, then the matrix $\mathbf{W}'\mathbf{X}$ is square; if it's also invertible, then $\tilde{\boldsymbol{\beta}}$ is

$$\tilde{\boldsymbol{\beta}}=(\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{y}; \tag{6.12}$$

this is sometimes called the "simple" IV estimator.

The actual availability of the variables $\mathbf{W}$ may be a problem, sometimes. In fact, collecting observable data that can be used as instrument is a bit of an art, although in many cases the choice of instruments is dictated by economic intuition. In Section 6.6, we will look at the examples provided in Section 6.1 and suggest possible solutions. However, before doing so, it is convenient to consider a generalisation.

6.2.1 The generalised IV estimator

What if the number of columns of $\mathbf{W}$ (call it m) was different from number of columns from $\mathbf{X}$ (call it k)? Of course, the matrix $\mathbf{W}'\mathbf{X}$ wouldn't be square and therefore not invertible. While there's no remedy for the case $m<k$, one may argue that in the opposite case we could just drop $m-k$ columns from $\mathbf{W}$ and proceed as above. While this makes sense, the reader will probably feel uneasy at the thought of dumping information deliberately. And besides, how do we choose which columns to drop from $\mathbf{W}$?

Fortunately, there is a solution: assume, for simplicity, that the covariance matrix of the structural disturbances is a multiple of the identity matrix:[5]

$$\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{W}\right]=\sigma^2\mathbf{I}.$$

By hypothesis, $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{W}]=0$; therefore,

$$\mathrm{E}\left[\mathbf{W}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{W}|\mathbf{W}\right]=\sigma^2\mathbf{W}'\mathbf{W}=\sigma^2\Omega.$$

[5] In fact, this assumption is not strictly necessary, but makes for a cleaner exposition.

Now define $\underset{m\times 1}{v}$, $\underset{m\times k}{C}$ and $\underset{m\times k}{\mathbf{e}}$ as

$$\underset{m\times 1}{v} = \mathbf{W}'\mathbf{y} \qquad \underset{m\times k}{C} = \mathbf{W}'\mathbf{X} \qquad \underset{m\times k}{\mathbf{e}} = \mathbf{W}'\varepsilon;$$

so the following equality holds:

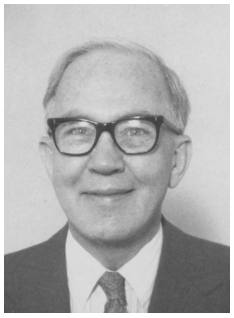
$$v = C\beta + \mathbf{e}, \tag{6.13}$$

Equation (6.13) may be seen as a linear model where the disturbance term has zero mean and covariance matrix $\sigma^2\Omega$. The number of explanatory variables is k (the column size of $\mathbf{X}$) but the peculiar feature of this model is that the number of "observations" is m (the column size of $\mathbf{W}$).

Since Ω is observable (up to a constant), we may apply the GLS estimator (see 4.2.1) to (6.13) and write

$$\begin{aligned}\tilde{\beta} &= \left[C'\Omega^{-1}C\right]^{-1}C'\Omega^{-1}v = \left[\mathbf{X}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y} = \\ &= (\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{y}. \end{aligned} \tag{6.14}$$

This clever idea is due to the English econometrician Denis Sargan,[6] whose name will also crop up later, in section 6.7.1.



DENIS SARGAN

The estimator $\tilde{\beta}$ in the equation above is technically called the **Generalised IV Estimator**, or **GIVE** for short. However, proving that (6.12) is just a special case of (6.14) when $m = k$ is a simple exercise in matrix algebra, left to the reader as an exercise, so when I speak of the IV estimator, what I mean is (6.14).

When $m = k$, the model is said to be **exactly identified**, as the estimator is based on solving (6.11), which is a system of m equations in m unknowns; if $\mathbf{W}'\mathbf{X}$ is invertible, it has one solution.

On the contrary, if $m > k$, (6.11) becomes a system with more equations than unknowns, so a solution does not ordinarily exist. The statistic we use is not a solution of (6.11), but is rather defined by re-casting the problem as a *sui generis* OLS model as in (6.13).

In this case, we say the model is **over-identified** and the difference $(m-k)$ is referred to as **over-identification rank**. The opposite case, when $m < k$, is a textbook case of under-identification, which I described in section 2.5. In short, one may say that a necessary condition for the existence of the IV estimator is that $m \geq k$; this is known as the **order condition**.

[6] For historical accuracy, it must be said that the idea of IV estimation was first put forward as early as 1953 by the Dutch genius Henri Theil. But it was Sargan who created the modern approach, in an article appeared in 1958.

As a by-product from estimating $\boldsymbol{\beta}$, you also get a residual vector $\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$, so that an estimator of σ^2 is readily available:

$$\tilde{\sigma}^2 = \frac{\tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}}{n}.$$

As shown in section 6.A.1, it can be proven that, under the set of assumptions I just made, the statistics $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ are CAN estimators. Therefore, the whole testing apparatus we developed in Chapter 3 can be applied without modifications since

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} \mathcal{N}(0, V).$$

The precise form of the asymptotic covariance matrix V is not important here; see 6.A.1. What is important in practice is that, under homoskedasticity, we have an asymptotically valid matrix we can use for hypothesis testing, which is

$$\hat{V} = \tilde{\sigma}^2(\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}.$$

In more general cases, robust alternatives (see section 4.2.2) are available.

Just like OLS, the IV estimator may be defined as the solution of an optimisation problem:

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\text{Argmin}}\, \boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{P}_{\mathbf{W}}\boldsymbol{\varepsilon}(\boldsymbol{\beta})$$

(compare the above expression with equation (1.14)).
In this book, we will not make much use of this property. However, defining $\tilde{\boldsymbol{\beta}}$ in this way would be the first step towards seeing it as a member of a very general category of estimators known as **GMM** (*Generalised Method of Moments*) estimators, which includes practically all estimators used in modern econometrics. The theory of GMM is beautiful: as a starting point, I heartily recommend Hayashi (2000).

6.2.2 The instruments

I will not distract the reader here with technicalities on the asymptotics of the IV estimator; you'll find those in Section 6.A.1. Here, I'm going to focus on two necessary conditions for consistency of $\tilde{\boldsymbol{\beta}}$ and explore what requisites they imply for the variables we choose as instruments. The two conditions are:

1. $\frac{1}{n}\sum \mathbf{w}_i\varepsilon_i \xrightarrow{\text{p}} \mathbf{0}$
2. $\frac{1}{n}\sum \mathbf{x}_i\mathbf{w}_i' \xrightarrow{\text{p}} A$, where A is a $k \times m$ matrix with rank k.

Condition 1 is more or less guaranteed by (6.10), that is by $\mathbf{w}_i$ being **exogenous**, which basically means "uncorrelated with the structural disturbance" ε_i; if this requirement isn't met, the limit in probability of $\tilde{\boldsymbol{\beta}}$ is not $\boldsymbol{\beta}$. End of story.

The implications of condition 2 are more subtle. The first one is: since the rank of A cannot be k if $m < k$, you need to have at least as many instruments as regressors. The good news is, this condition is not as stringent as it may seem at first sight: the fact that $\mathrm{E}[\mathbf{x}_i \cdot \varepsilon_i]$ is not a vector of zeros does not necessarily

mean that *all* its elements are nonzero. Some of the explanatory variables may be exogenous; in fact, in empirical models the subset of explanatory variables that may be suspected of endogeneity is typically rather small. Therefore, the exogenous subset of $\mathbf{x}_i$ is perfectly adequate to serve as an instrument, obvious examples being deterministic variables such as the constant. What the order condition really means is that, for each endogenous explanatory variable, we need at least one instrument *not also used as a regressor*.

Clearly, $m \geq k$ is not a sufficient condition for A to have rank k. For example, A may be square but have a column full of zeros. This would happen, for example, if the corresponding instrument was independent of all regressors $\mathbf{x}$. The generalisation of this idea leads to the concept of **relevance**.[7] The instruments must not only be exogenous, but must also be related to the explanatory variables.[8]

Note a fundamental difference between the order condition and the relevance condition: the order condition can be checked quite easily (all you have to do is count the variables); and even if you can't be bothered with checking, if the order condition fails the IV estimator $\tilde{\boldsymbol{\beta}}$ is not computable, since $(\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})$ is singular and your software will complain about this.

The relevance condition, instead, is much trickier to spot, since (with probability 1) $\frac{1}{n}\sum \mathbf{x}_i \mathbf{w}_i'$ will have rank k even if A doesn't. Hence, if $\text{rk}(A) < k$, you will be able to compute $\tilde{\boldsymbol{\beta}}$, but unfortunately it will be completely useless as an estimator. It can be proven that, in such an unfortunate case, the limit in probability of $\tilde{\boldsymbol{\beta}}$ is not a constant, but rather a random variable, so there's no way it can be a consistent estimator for any parameter.

In order to make this point clearer, let me give you an example. Suppose that the three random variables y_i, x_i and w_i were continuous and scalar and imagine that that w_i is not relevant. The IV estimator would simply be

$$\tilde{\boldsymbol{\beta}} = \frac{n^{-1}\sum_{i=1}^{n} x_i y_i}{n^{-1}\sum_{i=1}^{n} w_i x_i};$$

now focus on the denominator of the expression above: clearly, the probability that $n^{-1}\sum_{i=1}^{n} w_i x_i = 0$ is 0, so the probability that $\tilde{\boldsymbol{\beta}}$ exists is 1 for any finite n. However, if you compute its probability limit, you see the problem very quickly: if w_i is not relevant, then $n^{-1}\sum_{i=1}^{n} w_i x_i \xrightarrow{\text{p}} A = 0$ (which has, of course, rank 0 instead of 1, as we would require). Therefore the denominator will be a nonzero number which becomes smaller and smaller as $n \to \infty$. The reader should easily see that in this case we can't expect the asymptotic distribution of $\tilde{\boldsymbol{\beta}}$ to collapse to a point. In this case the estimator is not inconsistent because it converges to the wrong value, but rather because it doesn't converge at all.

[7] In other contexts, what I call the relevance condition is known as the **rank condition**.

[8] It should be noted that these properties are, to some extent, contradictory: if $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are correlated, any variable perfectly correlated to $\mathbf{X}$ could not be orthogonal to $\boldsymbol{\varepsilon}$. The trick here is that $\mathbf{W}$ is not perfectly correlated to $\mathbf{X}$.

Finally, instruments should be as "strong" as possible. The precise meaning of this phrase is the object of Section 6.7.2: here, I'll just mention the fact that inference with IV models could be quite problematic in finite samples, since the asymptotic approximations that we ordinarily use may work quite poorly. A common source of problems is the case of **weak** instruments: variables that are relevant, but whose connection with the regressors is so feeble that you need an inordinately large data set to use them for your purposes. This point will (hopefully) become clearer later, in the context of "two-stage" estimation (Section 6.5).

6.3 An example with real data

For this example, we are going to use a great classic from applied labour economics: the "Mincer wage equation"; the idea is roughly to have a model like the following:

$$y_i = \mathbf{z}_i'\boldsymbol{\beta}_0 + e_i\beta_1 + \varepsilon_i \tag{6.15}$$

where y_i is the log wage for an individual, e_i is their education level and the vector $\mathbf{z}_i$ contains other characteristics we want to control for (gender, work experience, etc). The parameter of interest is β_1, which measures the returns to education and that we would expect to be positive.

The reader, at this point, may dimly recall that we already estimated an equation like this, in section 1.5. Back then, we did not have the tools yet for interpreting the results from an inferential perspective, but the results were in agreement with commons sense. Why would we want to go back to a wage equation here?

The literature has long recognised that education may be endogenous, because the amount of education individuals receive is (ordinarily) decided by the individuals themselves. In practice, if the only reason to get an education is to have access to more lucrative jobs, individuals solve an optimisation problem where they decide, among other things, their own education level. This gives rise to a an endogeneity problem.[9]

For the reader's convenience, I'll reproduce here OLS estimates in Table 6.1. If education is endogenous, as economic theory suggests may be, then the "returns to education" parameter we find in the OLS output (about 5.3%) is a valid estimate of the marginal effect of education on the conditional expectation of wage, but is not a valid measure of the *causal* effect of education on wages, that is the increment in wage that an individual would have had if they had received an extra year of education.

I will now estimate the same equation via IV: the instruments I chose for this purpose are (apart from the three regressors other than education, which I take as exogenous) two variables that I will assume, for the moment, as valid instruments:

[9] The literature on this topic is truly massive. A good starting point is Card (1999).

```
OLS, using observations 1-1917
Dependent variable: lw

             coefficient   std. error    t-ratio    p-value
  ------------------------------------------------------------
  const      1.32891       0.0355309     37.40     2.86e-230 ***
  male       0.175656      0.0143224     12.26     2.42e-33  ***
  wexp       0.00608615    0.000671303    9.066    2.97e-19  ***
  educ       0.0526218     0.00202539    25.98     1.02e-127 ***

Mean dependent var   2.218765   S.D. dependent var   0.363661
Sum squared resid    177.9738   S.E. of regression   0.305015
R-squared            0.297629   Adjusted R-squared   0.296528
F(3, 1913)           270.2107   P-value(F)           3.3e-146
Log-likelihood      -441.8652   Akaike criterion     891.7304
Schwarz criterion    913.9645   Hannan-Quinn         899.9118
```

Table 6.1: Wage equation on the SHIW dataset — OLS estimates

- the individual's age: the motivation for this is that you don't choose when you're born, and therefore age can be safely considered exogenous; at the same time, regulations on compulsory education have changed over time, so it is legitimate to think that older people may have spent less time in education, so there are good chances age may be relevant.

- Parents' education level (measured as the higher between mother's and father's): it is a known fact that family environment is a powerful factor in educational choice. Yet, individuals can't decide on the educational level of their parents, so we may conjecture that this variable is both exogenous and relevant.

Table 6.2 shows the output from IV estimation; in fact, gretl (that is what I used) gives you richer output than this, but I'll focus on the part of immediate interest.

As you can see from the output, you get substantially different coefficients: not only you get that the returns from education appear to be quite stronger (7.9% versus 5.3%), but the other coefficients become larger too. Clearly, the question at this point becomes: OK, the two methods give different numbers. But are they *significantly* different? The tool we use to answer this question is the so-called **Hausman test**, which is the object of the next section.

6.4 The Hausman test

So far, we have taken as given that some of the variables in the regressor matrix $\mathbf{X}$ were endogenous, and that, as a consequence, OLS wouldn't yield consistent estimates of the parameters of interest. But of course, we don't know with cer-

```
TSLS, using observations 1-1917
Dependent variable: lw
Instrumented: educ
Instruments: const male wexp age peduc

             coefficient   std. error    t-ratio   p-value
  ------------------------------------------------------------
  const      0.943553      0.0684539      13.78    2.80e-41 ***
  male       0.182926      0.0149929      12.20    5.02e-33 ***
  wexp       0.00860475    0.000795375    10.82    1.62e-26 ***
  educ       0.0792106     0.00449758     17.61    1.85e-64 ***

Mean dependent var   2.218765   S.D. dependent var   0.363661
Sum squared resid    194.0070   S.E. of regression   0.318457
R-squared            0.292476   Adjusted R-squared   0.291366
F(3, 1913)           144.8624   P-value(F)           1.38e-84
```

Table 6.2: Wage equation on the SHIW dataset — IV estimates

tainty. One may argue that, if we have instruments whose quality we're confident about, we might as well stay on the safe side and use IV anyway. If we do, however, we may be using an inefficient estimator: it can be proven that, if $\mathbf{X}$ is exogenous, OLS is more efficient that IV under standard conditions.[10]

JERRY HAUSMAN

The Hausman test is based on the idea of comparing the two estimators and checking if their difference is statistically significant.[11] If it is, we conclude that OLS and IV have different probability limits, and therefore OLS can't be consistent, so our estimator of choice has to be IV. Otherwise, there is no ground for considering $\mathbf{X}$ endogenous, and we may well opt for OLS, which is more efficient.[12]

This idea can be generalised: if you have two estimators, one of which ($\tilde{\theta}$) is robust to some problem and the other one isn't ($\hat{\theta}$), the difference $\delta = \hat{\theta} - \tilde{\theta}$ should converge to a non-zero value if the problem is there, and to 0 otherwise. Therefore, we could set up a Wald-like statistic

$$H = \delta' \left[\widehat{V(\delta)} \right]^{-1} \delta; \tag{6.16}$$

where $\widehat{V(\delta)}$ is a consistent estimator of $\text{AV}[\delta]$, and, under some standard regularity conditions, it can be proven that H is asymptotically χ^2 under the null

[10] If you're curious, the proof is in section 6.A.2.

[11] As always, there is some paternity debate: some people call this the Wu-Hausman test; some others, the Durbin-Wu-Hausman test. While it is technically true that the same test statistic had been independently derived before (by Durbin in 1954 and by Wu in 1973), the idea became mainstream only after the publication of Hausman (1978).

[12] I know what you're thinking: this is the same logic we used in section 4.2.3 for the White test for heteroskedasticity. You're right.

($H_0 : \text{plim}(\delta) = 0$).[13]

The problem is, how do you compute $\widehat{V(\delta)}$? For the special case when the non-robust estimator is also efficient, we have a very nice result: the variance of the difference is the difference of the variances. A general proof is quite involved, but a for a sketch in the scalar case you can jump to section 6.A.3. In practice, if you have two estimators, and the situation is the one described in Table 6.3, an asymptotically valid procedure is just to compute H as

$$H = \left[\hat{\theta} - \tilde{\theta}\right]' \left(\text{AV}\left[\tilde{\theta}\right] - \text{AV}\left[\hat{\theta}\right]\right)^{-1} \left[\hat{\theta} - \tilde{\theta}\right].$$

Table 6.3: Hausman Test — special case

	if H_0 is true	if H_0 is false
$\hat{\theta}$	CAN and efficient	Inconsistent
$\tilde{\theta}$	CAN but not efficient	

In our case, the two estimators to compare are $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, so $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$. If we assume that OLS is efficient (which would be under homoskedasticity), then

$$\text{V}[\boldsymbol{\delta}] = \text{V}\left[\tilde{\boldsymbol{\beta}}\right] - \text{V}\left[\hat{\boldsymbol{\beta}}\right].$$

Since under H_0 OLS is consistent, then $\hat{\sigma}^2$ is a consistent estimator of σ^2 and we can use the matrix

$$\hat{\sigma}^2 \left[(\mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\right];$$

therefore,

$$H = \frac{\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)' \left[(\mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\right]^{-1} \left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}^2}. \tag{6.17}$$

In practice, actual computation of the test is performed even more simply, via an auxiliary regression: consider the model

$$\mathbf{y} = \mathbf{X}\beta + \hat{\mathbf{X}}\boldsymbol{\gamma} + \text{residuals}. \tag{6.18}$$

where $\hat{\mathbf{X}} \equiv \mathbf{P}_\mathbf{W}\mathbf{X}$. By the Frisch–Waugh theorem (see section 1.4.4)

$$\hat{\boldsymbol{\gamma}} = \left[\hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\hat{\mathbf{X}}\right]^{-1} \hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\mathbf{y};$$

now rewrite the two matrices on the right-hand side of the equation above as

$$\begin{aligned} \hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\hat{\mathbf{X}} &= \hat{\mathbf{X}}'\hat{\mathbf{X}} - \hat{\mathbf{X}}'\mathbf{P}_\mathbf{X}\hat{\mathbf{X}} = \hat{\mathbf{X}}'\hat{\mathbf{X}} - \hat{\mathbf{X}}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{X}} = \\ &= \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)\left[\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right) \qquad (6.19) \\ \hat{\mathbf{X}}'\mathbf{M}_\mathbf{X}\mathbf{y} &= \hat{\mathbf{X}}'\mathbf{y} - \hat{\mathbf{X}}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right) \qquad (6.20) \end{aligned}$$

[13] The number of degrees of freedom for the test is not as straightforward to figure out as it may seem. See below.

where we repeatedly used the equality $\hat{\mathbf{X}}'\mathbf{X} = \mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$; therefore,

$$\hat{\gamma} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\left[\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]^{-1}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right). \tag{6.21}$$

A Wald test for $\gamma = \mathbf{0}$ is

$$W = \frac{\hat{\gamma}'\left[\hat{\mathbf{X}}'\mathbf{M}_{\mathbf{X}}\hat{\mathbf{X}}\right]\hat{\gamma}}{\hat{\sigma}^2},$$

so, after performing a few substitutions, you get

$$W = \frac{\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)'\left[\left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]^{-1}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}^2} = H.$$

Of course, the possibility of $\mathbf{X}$ and $\mathbf{W}$ having some columns in common complicates slightly the setup above, and $\hat{\mathbf{X}}$ should only contain the projection on $\mathbf{W}$ of the endogenous regressors, because if a regressor is also contained in $\mathbf{W}$ its projection is obviously identical to the original. This means that the degrees of freedom for the Hausman test is equal to the number of explanatory variables that are effectively treated as endogenous (that is, are not present in the instrument matrix $\mathbf{W}$).

Example 6.1

Let us go back to the wage equation example illustrated in Section 6.3. While commenting Table 6.2, I mentioned the fact that my software of choice (gretl) offers richer output than what I reported. Part of it is the outcome of the Hausman test, that compares IV vs OLS:

```
Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 50.2987
  with p-value = 1.32036e-12
```

As you can see, our original impression that the two sets of coefficients were substantially different was definitely right. The p-value for the test leads to rejecting the null hypothesis very strongly. Therefore, IV and OLS have different limits in probability, which we take as a sign that education is, in fact, endogenous.

Note that the test statistic is matched against a χ^2 distribution with 1 degree of freedom, because there is one endogenous variable in the regressors list (that is, education).

6.5 Two-stage estimation

The IV estimator is also called **two-stage** estimator (hence the acronyms **TSLS**, for *Two-Stage Least Squares* or **2SLS**, for *2-Stage Least Squares*).

The reason is that the $\tilde{\boldsymbol{\beta}}$ statistic may be computed by two successive applications of OLS, called the two "stages".[14] In the era when computation was expensive, this was a nice trick to calculate $\tilde{\boldsymbol{\beta}}$ without the need for other software than OLS, but seeing IV as the product of a two-stage procedure has other advantages too.

In order to see what the two stages are, define $\hat{\mathbf{X}} = \mathbf{P}_{\mathbf{W}}\mathbf{X}$ and rewrite (6.14) as follows:

$$\tilde{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \tag{6.22}$$

The matrix $\hat{\mathbf{X}}$ contains, in the j-th column, the fitted value of a regression of the j-th column of $\mathbf{X}$ on $\mathbf{W}$; the regression

$$\mathbf{x}_i = \boldsymbol{\Pi}\mathbf{w}_i + \mathbf{u}_i \tag{6.23}$$

is what we call the **first stage** regression. In the **second stage**, you just regress $\mathbf{y}$ on $\hat{\mathbf{X}}$: the OLS coefficient equals $\tilde{\boldsymbol{\beta}}$. Note: this is a numerically valid procedure for computing $\tilde{\boldsymbol{\beta}}$, but the standard errors you get are *not* valid for inference. This is because second stage residuals are $\mathbf{e} = \mathbf{y} - \hat{\mathbf{X}}\tilde{\boldsymbol{\beta}}$, which is a different vector from the IV residuals $\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$. Consequently, the statistic $\frac{\mathbf{e}'\mathbf{e}}{n}$ does not provide a valid estimator of σ^2, which in turn makes the estimated covariance matrix invalid.

Readers who liked the geometrical interpretation of OLS as a projection might like to consider a different way of writing equation (6.22), that is

$$\tilde{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y},$$

from which you have that the fitted values from the GIVE estimator can be written as

$$\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}(\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} = \mathbf{Q}_{\hat{\mathbf{X}},\mathbf{X}}\mathbf{y}$$

The matrix $\mathbf{Q}_{\hat{\mathbf{X}},\mathbf{X}}$ is square and idempotent (but not necessarily symmetric) and performs what is called an *oblique* projection.

So, if the only computing facility you have is OLS (which was often the case in the 1950s and 1960s), you can compute the IV estimator via a repeated application of OLS. Moreover, you don't really have to run as many first-stage regressions as the number of regressors. You just have to run one first-stage regression for each endogenous element of $\mathbf{X}$ (recall the discussion in subsection 6.4 on the degrees of freedom for the Hausman test).

Example 6.2

Let's estimate the same model we used in section 6.3 via the two-stage method: the output from the first stage regression is in Table 6.4: as you can see, the dependent variable here is education, while the explanatory variables are the full instrument matrix $\mathbf{W}$*. There is not much to say about the first stage regression, except noting that the two "real" instruments (age and parents' education) are both highly significant. This will be important in the context of weak instruments (see section 6.7.2).*

[14] In fact, the word Henri Theil used when he invented this method was "rounds", but subsequent literature has settled on "stages".

```
OLS, using observations 1-1917
Dependent variable: educ

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        5.43051     0.491957      11.04     1.65e-27 ***
  male        -0.137838    0.142828      -0.9651   0.3346
  wexp        -0.205149    0.0129572    -15.83     3.78e-53 ***
  age          0.192950    0.0146614     13.16     6.26e-38 ***
  peduc        0.386233    0.0207287     18.63     2.47e-71 ***

Mean dependent var   11.73344   S.D. dependent var   3.598508
Sum squared resid    17665.37   S.E. of regression   3.039607
R-squared            0.287996   Adjusted R-squared   0.286507
F(4, 1912)           193.3448   P-value(F)           2.6e-139
Log-likelihood      -4848.785   Akaike criterion     9707.570
Schwarz criterion    9735.363   Hannan-Quinn         9717.797
```

Table 6.4: Wage equation on the SHIW dataset — first stage

```
OLS, using observations 1-1917
Dependent variable: lw

             coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      0.943553      0.0711035      13.27    1.64e-38 ***
  male       0.182926      0.0155733      11.75    8.22e-31 ***
  wexp       0.00860475    0.000826161    10.42    9.55e-25 ***
  hat_educ   0.0792106     0.00467166     16.96    3.48e-60 ***

SSR = 209.316, R-squared = 0.173936
```

Table 6.5: Wage equation on the SHIW dataset — second stage

Once the first stage regression is computed, we save the fitted values into a new variable called `hat_educ`*. Then, we replace the original education* `educ` *variable with* `hat_educ` *in the list of regressors and perform the second stage. Results are in Table 6.5; a comparison of these results with those reported in Table 6.2 reveals that:*

1. *the coefficients are identical;*
2. *the standard error are not, because the statistic you obtain by dividing the SSR from the second-stage regression (209.316) by the number of observations (1917) is not a consistent estimator for σ^2; therefore, the standard errors reported in table 6.5 differ from the correct ones by a constant scale factor (1.0387 in this case).*

6.5.1 The control function approach

The method I described in the previous subsection to compute the IV estimator via two successive stages is the traditional one. A slight variation on that procedure gives rise to what is sometimes called the **control function** approach.

The main idea can be grasped by considering a minimal model such as

$$y_i = x_i \cdot \beta + \varepsilon_i \tag{6.24}$$
$$x_i = w_i \cdot \pi + u_i, \tag{6.25}$$

where equation (6.25) is a "proper" linear model, and $\text{Cov}[u_i, \varepsilon_i]$ is some real number, not necessarily 0. If w_i is exogenous, the only way x_i can be correlated with ε_i is if $\text{Cov}[u_i, \varepsilon_i] \neq 0$. It is easy to show that OLS will overestimate β if $\text{Cov}[u_i, \varepsilon_i] > 0$ and underestimate it if $\text{Cov}[u_i, \varepsilon_i] < 0$. If, however, we defined

$$\nu_i = \varepsilon_i - \text{E}[\varepsilon_i | u_i]$$

and assumed linearity of $\text{E}[\varepsilon_i | u_i]$, we could write $\varepsilon_i = u_i \cdot \theta + \nu_i$ and recast (6.24) as

$$y_i = x_i \cdot \beta + u_i \cdot \theta + \nu_i \tag{6.26}$$

If we could observe u_i, we wouldn't need the IV estimator at all, because ν_i is orthogonal to both x_i and u_i,[15] so OLS would be our tool of choice to estimate β and θ at the same time.

Unfortunately, we don't observe u_i directly, but once we've run the first stage regression (6.25), we have the first-stage residuals $\hat{u}_i$, which hopefully shouldn't be too different: after all, (6.25) is a perfectly valid regression model, and the OLS estimate of π is consistent. Therefore, the difference

$$u_i - \hat{u}_i = (x_i - w_i \pi) - (x_i - w_i \hat{\pi}) = (\hat{\pi} - \pi)\, w_i$$

[15] Lazy writers like myself love the sentence: "the proof is left to the reader as an exercise".

should go to 0 asymptotically, since $\hat{\pi} \xrightarrow{p} \pi$; on these premises, the possibility of using $\hat{u}_i$ in place of the "true" u_i is tempting.

From a computational point of view, hence, the control function approach differs from the traditional two-stage method only in the second stage. Once you have performed the first stage, you use, in the second stage, the residuals from the first stage and add them as extra explanatory variables to the main equation. Call $\mathbf{E}$ the first-stage residuals and perform an OLS regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{E}$ together:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}\boldsymbol{\theta} + \boldsymbol{\nu}; \tag{6.27}$$

the estimates we get have a very nice interpretation.

Let's begin with $\boldsymbol{\beta}$: by the Frisch-Waugh theorem (see section 1.4.4), the OLS estimate of $\boldsymbol{\beta}$ is

$$(\mathbf{X}'\mathbf{M}_\mathbf{E}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{E}\mathbf{y};$$

now focus on the matrix $\mathbf{X}'\mathbf{M}_\mathbf{E}$:

$$\mathbf{X}'\mathbf{M}_\mathbf{E} = \mathbf{X}' - \mathbf{X}'\mathbf{P}_\mathbf{E} = \mathbf{X}' - \mathbf{X}'\mathbf{E}(\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}'.$$

Since, by definition, $\mathbf{E} = \mathbf{M}_\mathbf{W}\mathbf{X}$, by substitution we have

$$\mathbf{X}'\mathbf{M}_\mathbf{E} = \mathbf{X}' - \mathbf{X}'\mathbf{M}_\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{M}_\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{W} = \mathbf{X}' - \mathbf{X}'\mathbf{M}_\mathbf{W} = \mathbf{X}'\mathbf{P}_\mathbf{W};$$

therefore, the coefficient vector becomes

$$(\mathbf{X}'\mathbf{M}_\mathbf{E}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{E}\mathbf{y} = (\mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{W}\mathbf{y} = \tilde{\boldsymbol{\beta}}$$

So the OLS estimate of the coefficients associated with $\mathbf{X}$ in equation (6.27) is exactly equal to the IV estimator $\tilde{\boldsymbol{\beta}}$. Warning: just like the two-stage procedure, the control function approach does *not* yield correct standard errors for $\tilde{\boldsymbol{\beta}}$; explaining precisely why is far beyond the scope of this book, and readers will have to content themselves with knowing that this is a consequence of using the first-stage residuals $\hat{u}_i$ in place of the true u_i series.[16]

Moreover, the control function regression gives you, as a nice by-product, the Hausman test, as an exclusion test for the first-stage residuals: by applying the Frisch-Waugh theorem again, the OLS estimate of $\boldsymbol{\theta}$ in (6.27) is

$$\hat{\boldsymbol{\theta}} = \left[\mathbf{E}'\mathbf{M}_\mathbf{X}\mathbf{E}\right]^{-1}\mathbf{E}'\mathbf{M}_\mathbf{X}\mathbf{y};$$

of course $\hat{\mathbf{e}} = \mathbf{M}_\mathbf{X}\mathbf{y}$ are the OLS residuals. Now use the definition of $\mathbf{E}$ as $\mathbf{E} = \mathbf{M}_\mathbf{W}\mathbf{X}$ again:

$$\mathbf{E}'\mathbf{M}_\mathbf{X}\mathbf{y} = \mathbf{E}'\hat{\mathbf{e}} = \mathbf{X}'\mathbf{M}_\mathbf{W}\hat{\mathbf{e}} = \mathbf{X}'\hat{\mathbf{e}} - \mathbf{X}'\mathbf{P}_\mathbf{W}\hat{\mathbf{e}} = -\mathbf{X}'\mathbf{P}_\mathbf{W}\hat{\mathbf{e}};$$

[16] If you really want to know, the problem arises because we are using $\hat{\pi}$ to compute $\hat{u}_i$ by treating it as if it was the true π, and hence ignoring the fact that $\hat{\pi}$ is an estimate with a non-zero variance. This is a case of **generated regressors**; section 6.1 of Wooldridge (2010) is where you want to start from. Besides, Section 6.2 of the same book contains a much more accurate and thorough treatment of the control function approach than what I'm giving you here.

where the last equality comes from the OLS residuals $\hat{\mathbf{e}}$ being orthogonal to $\mathbf{X}$; therefore

$$\mathbf{E}'\mathbf{M}_{\mathbf{X}}\mathbf{y} = -\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{y} + \mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X}\cdot\hat{\beta} = \left[\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X}\right](\hat{\beta} - \tilde{\beta}),$$

and as a consequence $\hat{\boldsymbol{\theta}}$ is just

$$\hat{\boldsymbol{\theta}} = \left[\mathbf{E}'\mathbf{M}_{\mathbf{X}}\mathbf{E}\right]^{-1}\left[\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X}\right](\hat{\beta} - \tilde{\beta});$$

since the matrix $\left[\mathbf{E}'\mathbf{M}_{\mathbf{X}}\mathbf{E}\right]^{-1}\left[\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X}\right]$ is invertible, $\hat{\boldsymbol{\theta}}$ it can be zero if and only if $\hat{\beta} = \tilde{\beta}$. Therefore, the hypothesis $h_0 : \boldsymbol{\theta} = \mathbf{0}$ is logically equivalent to the null hypothesis of the Hausman test, and the test can be performed by a simple zero restriction on $\hat{\boldsymbol{\theta}}$. If (as often happens) $\boldsymbol{\theta}$ is a scalar, the result of the Hausman test is immediately visible as the significance t-test associated with that coefficient.

A very nice feature of the control function approach is that this approach is very natural to generalise to settings where our estimator of choice is not a least squares estimator, which happens quite often in applied work. But this book is entitled "basic econometrics", and I think I'll just stop here.

```
OLS, using observations 1-1917
Dependent variable: lw

             coefficient   std. error    t-ratio   p-value
  ------------------------------------------------------------
  const       0.943553     0.0647377     14.58     1.04e-45 ***
  male        0.182926     0.0141790     12.90     1.42e-36 ***
  wexp        0.00860475   0.000752195   11.44     2.34e-29 ***
  educ        0.0792106    0.00425341    18.62     2.89e-71 ***
  resid      -0.0341349    0.00481934    -7.083    1.98e-12 ***

SSR = 173.423, R-squared = 0.315587
```

Table 6.6: Wage equation on the SHIW dataset — control function

Example 6.3

Using the SHIW data again, after storing the residuals from the first-stage regression (see Table 6.4) under the name `resid`*, you can run an OLS regression like the one in Table 6.1, with* `resid` *added to the list of regressors. The results are in Table 6.6. Again, the coefficients for the original regressors are* $\tilde{\beta}$ *and again, the standard errors are not to be trusted (they're a bit smaller than the correct ones, listed in Table 6.2). The* t*-test for the* `resid` *variable, instead, is interpretable as a perfectly valid Hausman test, and the fact that we strongly reject the null again is no coincidence.*

6.6 The examples, revisited

6.6.1 Measurement error

A typical way to use IV techniques to overcome measurement error is the usage of a second measurement of the latent variable, whose contamination error is independent of the first one. In formulae: together with the two equations (6.1) and (6.2)

$$\begin{aligned} y_i &= x_i^* \beta + \varepsilon_i \\ x_i &= x_i^* + \eta_i \end{aligned}$$

suppose you have a third observable variable w_i such that

$$w_i = x_i^* + \nu_i$$

If η_i and ν_i are uncorrelated, then w_i is a valid instrument and the statistic

$$\tilde{\beta} = \frac{\sum w_i y_i}{\sum w_i x_i}$$

is a consistent estimator of β.

One famous application of this principle is provided in Griliches (1976), a landmark article in labour economics, where the author has two measurements of individual ability and uses one for instrumenting the other.

6.6.2 Simultaneous equation systems

Consider again equations (6.4)–(6.5):

$$\begin{aligned} q_t &= \alpha_0 - \alpha_1 p_t + u_t \\ p_t &= \beta_0 + \beta_1 q_t + v_t. \end{aligned}$$

As we proved in section 6.1.2, the systematic part of these two equations are not conditional means, so there's no way we can estimate their parameters consistently via OLS.

On the other hand, we can use (6.7) to deduce that $\mathrm{E}[p_t] = \pi_1$; clearly, the parameter π_1 can be estimated consistently very simply by taking the average of p_t (or regressing p_t on a constant, if you will). The same holds for $\mathrm{E}[q_t] = \pi_0$.

Can we use these two parameters to estimate the structural ones? The answer is no: the relationship between the (π_0, π_1) pair and the structural parameters is

$$\pi_0 = \frac{\alpha_0 - \beta_0 \alpha_1}{1 + \beta_1 \alpha_1} \qquad \pi_1 = \frac{\beta_0 + \beta_1 \alpha_0}{1 + \beta_1 \alpha_1},$$

which is a system of 2 equations in 4 unknowns; as such, it has infinitely many solutions. This is exactly the under-identification scenario we analysed in section 2.5.

Let us now consider this example in greater generality: a system of linear equations can be written as

$$\Gamma \mathbf{y}_t = B\mathbf{x}_t + \varepsilon_t; \tag{6.28}$$

the $\mathbf{y}_t$ vector contains the q endogenous variables, while $\mathbf{x}_t$ is an m-vector holding the exogenous variables. The matrix Γ (assumed non-singular) is $q \times q$ and B is a $q \times m$ matrix. In the demand-supply example,

$$\Gamma = \begin{bmatrix} 1 & \alpha_1 \\ -\beta_1 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}.$$

Equation (6.28) is known as the **structural form** of the system, because its parameters have a behavioural interpretation and are our parameters of interest.

By pre-multiplying (6.28) by Γ^{-1}, you get the so-called **reduced form**:

$$\mathbf{y}_t = \Pi\mathbf{x}_t + \mathbf{u}_t, \tag{6.29}$$

where $\Pi = \Gamma^{-1}B$ and $\mathbf{u}_t = \Gamma^{-1}\varepsilon_t$. In our example, the matrix Π is a column vector, containing π_0 and π_1.

If $\mathbf{x}_t$ is exogenous, then $\text{Cov}[\mathbf{x}_t, \varepsilon_t]$, so the correlation between $\mathbf{x}_t$ and $\mathbf{u}_t$ is zero; hence, OLS is a consistent estimator for the parameters of the reduced form. However, by postmultiplying (6.29) by $\mathbf{x}_t'$ you get:

$$\mathbf{y}_t\mathbf{x}_t' = \Pi\mathbf{x}_t\mathbf{x}_t' + \mathbf{u}_t\mathbf{x}_t',$$

which implies $\text{E}\left[\mathbf{y}_t\mathbf{x}_t'\right] = \Pi\text{E}\left[\mathbf{x}_t\mathbf{x}_t'\right]$. Ordinarily, this matrix should not contain zeros; if variables were centred in mean, it would be the covariance matrix between the vector $\mathbf{y}_t$ and the vector $\mathbf{x}_t$. Therefore, each element of $\mathbf{x}_t$ is correlated with each $\mathbf{y}_t$ despite being uncorrelated with ε_t. In other words, $\mathbf{x}_t$ is both *relevant* and *exogenous* and, as such, is a perfect instrument.

In the example above, $\mathbf{x}_t$ contains only the constant term, and the reduced form looks like:

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} \cdot 1 + \mathbf{u}_t;$$

it should be clear where under-identification comes from: in the demand equation you have 2 regressors (the constant and p_t) but only one instrument (the constant), and the same goes, *mutatis mutandis*, for the supply equation.

Consider now a different formulation, where:

$$q_t = \alpha_0 - \alpha_1 p_t + \alpha_2 y_t + u_t \tag{6.30}$$

$$p_t = \beta_0 + \beta_1 q_t + \beta_2 m_t + v_t, \tag{6.31}$$

where we use the two new variables y_t, the per-capita income at time t and m_t, the price of raw materials at time t; assume both are exogenous.

In this case, both equations are estimable via IV, because we have three regressors and three instruments for each (the same for both: constant, y_t and

m_t). In the context of simultaneous systems, the order condition I stated in section 6.2.2 is often translated as: for each equation in the system the number of included endogenous variables cannot be greater than the number of excluded exogenous variables. I'll leave it to the reader to work out the equivalence.

6.7 Are my instruments OK?

A fundamental requirement for the whole IV strategy is that the variables that we choose to use as instruments are (i) exogenous and (ii) relevant. Of course, we cannot assume that they are just because we say so, and it'd be nice to have some way of testing these assumptions. It turns out that neither property can be verified directly, but we do have statistics that we can interpret in a useful way.

6.7.1 The Sargan test

The Sargan test is often interpreted as a test for exogeneity of the instruments, but in fact things are a bit more subtle.

Exogeneity, in our context, means uncorrelatedness between the instruments $\mathbf{w}_i$ and the structural disturbances ε_i. Assume we are in the simplest case, where $\mathrm{V}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$ and $\frac{1}{n}\mathbf{W}'\mathbf{W} \xrightarrow{\mathrm{p}} B$. If the structural disturbances ε_i were observable, a test would be straightforward to construct: under H_0,

$$\frac{1}{\sqrt{n}}\mathbf{W}'\boldsymbol{\varepsilon} \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mathbf{0},\, \sigma^2 B\right)$$

and it can be proven that, under H_0,

$$\frac{\boldsymbol{\varepsilon}'\mathbf{P}_{\mathbf{W}}\boldsymbol{\varepsilon}}{\sigma^2} \xrightarrow{\mathrm{d}} \chi^2_m, \tag{6.32}$$

where m is the size of $\mathbf{w}_i$. Unfortunately, $\boldsymbol{\varepsilon}$ is unobservable, and therefore the quantity above is not a statistic and cannot be used as a test.

The idea of substituting disturbances with residuals like we did in section 6.5.1 takes us to the **Sargan test**. Its most important feature is that this test has a different asymptotic distribution than (6.32), since the degrees of freedom of the limit χ^2 is not m (the number of instruments), but rather $m-k$, where k is the number of elements in $\boldsymbol{\beta}$: in other terms, the over-identification rank (see section 6.2.1). In formulae,

$$S = \frac{\tilde{\boldsymbol{\varepsilon}}'\mathbf{P}_{\mathbf{W}}\tilde{\boldsymbol{\varepsilon}}}{\hat{\sigma}^2} \xrightarrow{\mathrm{d}} \chi^2_{m-k}. \tag{6.33}$$

This result may appear surprising at first. Consider, however, that under exact identification the numerator of S is identically zero,[17] so, in turn, the S statistic is identically 0, not a χ^2_m variable. Can this result be generalised?

[17] Blitz proof: if $m = k$, then $\tilde{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{y}$. Therefore, $\mathbf{P}_{\mathbf{W}}\tilde{\boldsymbol{\varepsilon}} = \mathbf{P}_{\mathbf{W}}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$; however, observe that $\mathbf{P}_{\mathbf{W}}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{W}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{X}(\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{y} = \mathbf{P}_{\mathbf{W}}\mathbf{y}$. As a consequence, $\mathbf{P}_{\mathbf{W}}\tilde{\boldsymbol{\varepsilon}} = \mathbf{P}_{\mathbf{W}}\mathbf{y} - \mathbf{P}_{\mathbf{W}}\mathbf{y} = \mathbf{0}$.

Take the conditional expectation $\mathrm{E}\left[y_i|\mathbf{w}_i\right]$; assuming linearity, if $\mathbf{w}_i$ is exogenous we could estimate its parameters via OLS in a model like

$$y_i = \mathbf{w}_i'\boldsymbol{\pi} + u_i; \tag{6.34}$$

in the simultaneous system jargon, equation (6.34) would be the reduced-form equation for y_i. To see how the parameters $\boldsymbol{\pi}$ relate to $\boldsymbol{\beta}$, write the first-stage equation (6.25) in matrix form as

$$\mathbf{X} = \mathbf{W}\boldsymbol{\Pi} + \mathbf{E}$$

so that the i-th row of $\mathbf{X}$ can be written as $\mathbf{x}_i' = \mathbf{w}_i'\boldsymbol{\Pi} + \mathbf{e}_i'$; now substitute in the structural equation:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i = \mathbf{w}_i'\boldsymbol{\Pi}\boldsymbol{\beta} + (\varepsilon_i + \mathbf{e}_i'\boldsymbol{\beta}); \tag{6.35}$$

clearly, the two models (6.34) and (6.35) become equivalent only if $\boldsymbol{\pi} = \boldsymbol{\Pi}\boldsymbol{\beta}$; in fact, the expression $\boldsymbol{\Pi}\boldsymbol{\beta}$ can be seen as a restricted version of $\boldsymbol{\pi}$, where the constraint is that $\boldsymbol{\pi}$ must be a linear combination of the columns of $\boldsymbol{\Pi}$ (or, more concisely, that $\boldsymbol{\pi} \in \mathrm{Sp}(\boldsymbol{\Pi})$).

The Sargan test is precisely a test for those restrictions: this begs three questions:

1. how do we compute the test statistic?
2. What is its limit distribution?
3. Which interpretation must be given to a rejection?

Number 1 is quite easy: the IV residuals $\widetilde{\varepsilon}$ are the residuals from the restricted model. All we have to do is apply the LM principle (see Section 3.5.1) and regress those on the explanatory variables from the unrestricted model. Compute nR^2, and your job is done. If you do, you end up exactly with the statistic I called S in equation (6.33).

As for its limit distribution, the fact that the number of degrees of freedom of the χ^2 limit distribution is $m-k$ can be intuitively traced back to the fact that the number of parameters of the unrestricted model (6.34) is m, while the number of the restricted parameters is k. Therefore, the number of constraints is $m-k$.[18]

Now we can tackle point 3: the null hypothesis in the Sargan test is that the m relationships implicit in the equation $\boldsymbol{\pi} = \boldsymbol{\Pi}\boldsymbol{\beta}$ are *non-contradictory*; if they were, it would mean that at least one element of the vector $\mathrm{E}[\mathbf{w}_i\varepsilon_i]$ is non-zero,

[18] A more rigorous argument goes as follows: if the restriction is true, then $\mathbf{P}_{\boldsymbol{\Pi}}\boldsymbol{\pi} = \boldsymbol{\pi}$, or, equivalently, $\mathbf{M}_{\boldsymbol{\Pi}}\boldsymbol{\pi} = \mathbf{0}$. Since the rank of $\mathbf{M}_{\boldsymbol{\Pi}}$ is $m-k$, we can write

$$\mathbf{M}_{\boldsymbol{\Pi}}\boldsymbol{\pi} = UV'\boldsymbol{\pi} = \mathbf{0}$$

where V and U are matrices with k rows and $m-k$ columns (see section 1.A.3). So, the null hypothesis implicit in the restriction is $H_0 : V'\boldsymbol{\pi} = \mathbf{0}$, which is a system of $m-r$ constraints.

and therefore at least one instrument in not exogenous. Unfortunately, the test cannot identify the culprit.

To clarify the matter, take for example a simple DGP where $y_i = x_i\beta + \varepsilon_i$ and we have two potential instruments, $w_{1,i}$ and $w_{2,i}$. We can choose among three possible IV estimators for β:

1. a simple IV estimator using $w_{1,i}$ only (call it β_1);
2. a simple IV estimator using $w_{2,i}$ only (call it β_2);
3. the GIVE estimator using both $w_{1,i}$ and $w_{2,i}$ (call it β_{12}).

Suppose β_1 turns out to be positive (and very significant), and β_2 turns out to be negative (and very significant); clearly, there must be something wrong. At least one between β_1 and β_2 cannot be consistent. The Sargan test, applied to the third model, would reject the null hypothesis and inform us that at least one of our two instruments is probably not exogenous. Hence, β_{12} would be certainly inconsistent, and we'd have to decide which one to keep between β_1 and β_2 (usually, on the basis of some economic reasoning). If, on the other hand, we were unable to reject the null, then we would probably want to use β_{12} for efficiency reasons, since it's the one that incorporates all the information available from the data.

In view of these features, the Sargan test is often labelled **overidentification test**, since what it can do is, at most, finding whether there is a contradiction between the m assumptions we make when we say "I believe instrument i is exogenous".

```
Sargan over-identification test -
  Null hypothesis: all instruments are valid
  Test statistic: LM = 0.138522
  with p-value = P(Chi-square(1) > 0.138522) = 0.709755
```

Table 6.7: Wage equation — Sargan test

Example 6.4

Table 6.7 is, again, an excerpt from the full output that gretl gives you after IV estimation (the main table is 6.2) and shows the Sargan test for the wage equation we've been using as an example; in this case, we have 1 endogenous variable (education) and two instruments (age and parents' education), so the overidentification rank is $2-1=1$. *The p-value for the test is over 70%, so the null hypothesis cannot be rejected. Hence, we conclude that our instruments form a coherent set and the estimates that we would have obtained by using age alone or parents' education alone would not have been statistically different from one another. Either our instruments are all exogenous or they are all endogenous, but in the latter case they would all be wrong in exactly the same way, which seems quite unlikely.*

6.7.2 Weak instruments

The relevance condition for instruments looks deceptively simple to check: instruments are relevant if the matrix $A = \mathrm{E}\left[\mathbf{x}_i \mathbf{w}_i'\right]$ is full-rank. Of course, this matrix is quite easy to estimate consistently, so in principle testing for relevance is straightforward.

In fact, the point above is subtler than it looks: estimating the rank of a matrix is not exactly trivial, because the rank is an integer, so most of our intuitive ideas on the relationship between an estimator and the unknown parameter (that make perfect sense when the latter is a real number) break down. Constructing a test for the hypothesis $\mathrm{rk}(A) = m$ is possible, but requires some mathematical tools I chose not to include in this book. The interested reader may want to google for "canonical correlations".

The practical problem we are often confronted with is that, although an instrument is technically relevant, its correlation with the regressors could be so small that finite-sample effects may become important. In this case, that instrument is said to be **weak**.[19]

The problem is best exemplified by a little simulation study: consider the same model we used in section 6.5.1:

$$y_i = x_i \cdot \beta + \varepsilon_i \tag{6.36}$$

$$x_i = w_i \cdot \pi + u_i \tag{6.37}$$

$$\begin{bmatrix} \varepsilon_i \\ u_i \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}\right)$$

with the added stipulation that $w_i \sim U(0,1)$. Of course, equation (6.36) is our equation of interest, while (6.37) is the "first stage" equation. Since ε_i and u_i are correlated, then x_i is itself correlated with ε_i, and therefore endogenous; however, we have the variable w_i, which meets all the requirements for being a perfectly valid instrument: it is exogenous (uncorrelated with ε_i) and relevant (correlated with x_i), as long as the parameter π is nonzero.

However, if π is a small number the correlation between x_i and w_i is very faint, so w_i is weak, despite being relevant. In this experiment, the IV estimator is simply

$$\tilde{\beta} = (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{y} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i x_i} :$$

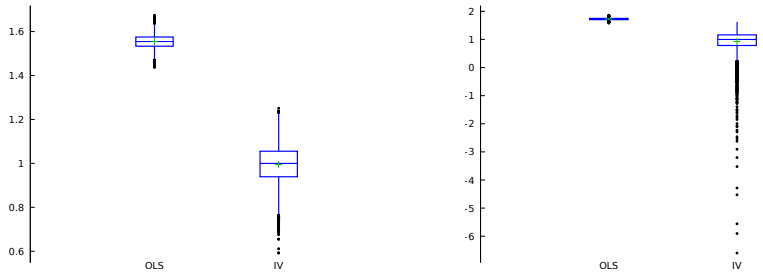
if you scale the denominator by $\frac{1}{n}$, it's easy to see that its probability limit is non-zero; however, its finite-sample distribution could well be spread out over a wide interval of the real axis, so you can end up dividing the numerator by

[19] Compared to the rest of the material contained in this book, inference under weak instruments is a fairly recent strand in econometric research. A recent review article I heartily recommend is Andrews et al. (2019), but a fairly accessible introductory treatment can also be found in Hill et al. (2018), Chapter 10. Chapter 12 of Hansen (2019) is considerably more technical, but highly recommended.

an infinitesimally small number and have an inordinately large statistic (not to mention the possibility of getting a wrong sign). Let me stress that this is a finite-sample issue: asymptotically, there are no problems at all; but since all we ever have are finite samples, the problem deserves consideration.

In order to show you what the consequences are, I generated 10000 artificial samples with 400 observations each and ran OLS and IV on equation (6.36), setting $\beta = 1$ and $\pi = 1$.

The results of the experiment are plotted in the left picture in Figure 6.1. If you want to repeat the experiment on your PC, the gretl code is at subsection 6.A.4. As you can see, everything works as expected: OLS has a smaller variance, but is inconsistent (none of the simulated $\hat{\beta}$ gets anywhere near the true value $\beta = 1$); IV, on the other hand, shows larger dispersion, but its distribution is nicely centred around the true value.



$\pi = 1$ $\pi = 1/3$

Figure 6.1: Weak instruments: simulation study

If instead you set $\pi = 1/3$, the simulation gives you the results plotted in the right-hand panel. Asymptotically, nothing changes: however, the finite-sample distribution of $\tilde{\beta}$ is worrying. Not only its dispersion is rather large (and there are quite a few cases when the estimated value for β is negative): its distribution is very far from being symmetric, which makes it questionable to use asymptotic normality for hypothesis testing. I'll leave it to the reader to figure out what happens if the instrumental variable w_i becomes *very* weak, which is what you would get by setting π to a very small value, such as $\pi = 0.1$.

More generally, the most troublesome finite-sample consequences of weak instruments are:

- the IV estimator is severely biased; that is, the expected value of its finite sample distribution may be very far from the true value β;[20]
- even more worryingly, the asymptotic approximations we use for our test statistics may be very misleading.

[20] I should add "provided it exists"; there are cases when the distribution of the IV estimator has no finite moments.

How do we spot the problem? Since this is a small-sample problem, it is not easy to construct a test for weak instruments: what should its null hypothesis be, precisely? What we can do, at most, is using some kind of descriptive statistic telling us if the *potential* defects of the IV estimator are likely to be an *actual* problem for the data we have.

For the simplest case, where you only have one endogenous variable in your model, the tool everybody uses is the so-called "first-stage F test", also labelled **partial F statistic**. You compute it as follows: take the first-stage regression (6.23) and perform an F-test (see Section 3.5.1) for the exclusion of the "true" instruments (that is, the elements of $\mathbf{w}_i$ not contained in $\mathbf{x}_i$). The suggestion contained in Staiger and Stock (1997) was that a value less than 10 could be taken as an indication of problems related to weak instruments.

Example 6.5

The weak instrument test for the example on the SHIW data we've been using in this chapter gives:

```
Weak instrument test -
  First-stage F-statistic (2, 1912) = 271.319
  Critical values for desired TSLS maximal size, when running
  tests at a nominal 5% significance level:

     size      10%      15%      20%      25%
    value    19.93    11.59     8.75     7.25

  Maximal size is probably less than 10%
```

The first-stage F statistic for the wage equation, as reported by gretl is 271.319, which is way above 10, so we don't have to worry.

The generalisation of this statistic is the so-called **Cragg-Donald** statistic, whose description and interpretation is somewhat more involved, and I'll just point you to the bibliographic references I made at the start of this section.

Finally, a warning: problems similar to weak instruments may also arise when the overidentification rank becomes large: the over-identification range is usually a rather small number, but in some contexts it could happen that we have an abundance of instruments. Common sense dictates that we should use as much information as we have available, but in finite samples things may not be so straightforward. A thorough analysis of the problem quickly becomes very technical, so I'll just quote Hall (2005), which contains an excellent treatment of the issue.

> There is a far more complex relationship between the behaviour of the [IV] estimator and the properties of the instrument vector in finite samples than is predicted by asymptotic theory.

6.A Assorted results

6.A.1 Asymptotic properties of the IV estimator

The limit in probability of the IV estimator (see 6.2.1 for its derivation)

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{y},$$

can be calculated by rewriting the equation above as a function of statistics for which the probability limits can be computed easily. Clearly, some regularity conditions (such as the observations being iid, for example) are assumed to hold so that convergence occurs; we'll take these as given and assume that sample moments converge in probability to the relevant moments.

Given the linear model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, we assume that:

1. $\frac{1}{n}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{w}_t' = \frac{\mathbf{X}'\mathbf{W}}{n} \xrightarrow{\text{p}} A$, where $\text{rk}(A) = k$;
2. $\frac{1}{n}\sum_{t=1}^{T}\mathbf{w}_t\mathbf{w}_t' = \frac{\mathbf{W}'\mathbf{W}}{n} \xrightarrow{\text{p}} B$, where B is invertible;
3. $\frac{1}{n}\sum_{t=1}^{T}\mathbf{w}_t u_t = \frac{\mathbf{W}'\boldsymbol{\varepsilon}}{n} \xrightarrow{\text{p}} \mathbf{0}$;

then $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{y} \xrightarrow{\text{p}} \boldsymbol{\beta}$. The proof is a simple application of Slutsky's theorem:

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{W}}\boldsymbol{\varepsilon} = \\ &= \boldsymbol{\beta} + \left[\left(\frac{\mathbf{X}'\mathbf{W}}{n}\right)\left(\frac{\mathbf{W}'\mathbf{W}}{n}\right)^{-1}\left(\frac{\mathbf{W}'\mathbf{X}}{n}\right)\right]^{-1}\left(\frac{\mathbf{X}'\mathbf{W}}{n}\right)\left(\frac{\mathbf{W}'\mathbf{W}}{n}\right)^{-1}\left(\frac{\mathbf{W}'\boldsymbol{\varepsilon}}{n}\right)\end{aligned}$$

so that

$$\tilde{\boldsymbol{\beta}} \xrightarrow{\text{p}} \boldsymbol{\beta} + \left[AB^{-1}A'\right]^{-1}AB^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}. \tag{6.38}$$

It is instructive to consider the role played by the ranks of A and B; the matrix B must be invertible, because otherwise $AB^{-1}A'$ wouldn't exist. Since B is the probability limit of the second moments of the instruments, this requirement is equivalent to saying that all instruments must carry separate information, and cannot be collinear.

Note that the requisite is only that the instruments shouldn't be collinear: the stronger requisite of independence is not needed. As a consequence, it is perfectly OK to use nonlinear transformations of one instrument to create additional ones.
For example, if you have a variable w_i that you assume independent of ε_i, you can use as instruments $w_i, w_i^2, \log(w_i), \ldots$ (provided of course that the transformed variables have finite moments).
This strategy is a special case of something called **identification through nonlinearity**; although it feels a bit like cheating (and is frowned upon by some), it is perfectly legitimate, at least asymptotically, as long as each transformation carries some extra amount of information.

The rank of A, instead, must be k for $\left[AB^{-1}A'\right]$ to be invertible. This means that instruments $\mathbf{w}_i$ must be relevant (see Section 6.2.2) for all the regressors $\mathbf{x}_i$. If $\left[AB^{-1}A'\right]$ is not invertible, then the probability limit in (6.38) does not exist. If, instead, it's invertible, but very close to being singular (as in the case of weak instruments — see Section 6.7.2), then its inverse will be a matrix with inordinately large values. This is mainly a problem for the distribution of $\tilde{\boldsymbol{\beta}}$: if we also assume

4. $\frac{1}{\sqrt{n}}\sum_{t=1}^{T}\mathbf{w}_t u_t = \frac{\mathbf{W}'\boldsymbol{\varepsilon}}{\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0,Q)$;

then $\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\text{d}} \mathcal{N}(0,\Sigma)$, where

$$\Sigma = \left[AB^{-1}A'\right]^{-1} AB^{-1}QB^{-1}A'\left[AB^{-1}A'\right]^{-1}.$$

In the standard case $Q=\sigma^2 B$ (from the homoskedasticity assumption $\text{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{W}\right] = \sigma^2\mathbf{I}$), and therefore

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\text{d}} \mathcal{N}\left(0,\sigma^2\left[AB^{-1}A'\right]^{-1}\right). \tag{6.39}$$

So the precision of the IV estimator is severely impaired any time the matrix $\left[AB^{-1}A'\right]$ is close to being singular.

The last thing is proving that $\tilde{\sigma}^2$ is consistent: from

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{X}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)+\boldsymbol{\varepsilon},$$

you get

$$\tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}} = \left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)+2\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)'\mathbf{X}'\boldsymbol{\varepsilon}+\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$$

and therefore

$$\frac{1}{n}\tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}} = \left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)+2\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)'\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)+\left(\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n}\right).$$

By taking probability limits,

$$\tilde{\sigma}^2 = \frac{1}{n}\tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}} \xrightarrow{\text{p}} \mathbf{0}'(Q)\,\mathbf{0}+2\cdot\mathbf{0}'\boldsymbol{\lambda}+\sigma^2 = \sigma^2,$$

where $Q=\text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)$ and $\boldsymbol{\lambda}=\text{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right)$. Note that $\tilde{\sigma}^2$ is consistent even though $\boldsymbol{\lambda}\neq\mathbf{0}$. Consistency of $\tilde{\sigma}^2$ is important because it implies that we can use the empirical counterparts of the asymptotic covariance matrix in equation (6.39) and use $\tilde{\sigma}^2(\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-1}$ as a valid covariance matrix for Wald tests.

6.A.2 Proof that OLS is more efficient than IV

In the OLS vs IV case, the proof that OLS is more efficient than IV if $\mathbf{X}$ is exogenous can be given as follows: given the model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, define the following quantities:

$$Q = \text{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) \qquad A = \text{plim}\left(\frac{\mathbf{W}'\mathbf{X}}{n}\right) \qquad B = \text{plim}\left(\frac{\mathbf{W}'\mathbf{W}}{n}\right);$$

under homoskedasticity, we have that $\text{AV}\left[\hat{\boldsymbol{\beta}}\right] = \sigma^2 Q^{-1}$ (see section 3.2.2) and $\text{AV}\left[\tilde{\boldsymbol{\beta}}\right] = \sigma^2\left[A'B^{-1}A\right]^{-1}$ (see 6.A.1), where $\sigma^2 = \text{V}[\varepsilon_i]$. In order to prove that $\text{AV}\left[\tilde{\boldsymbol{\beta}}\right] - \text{AV}\left[\hat{\boldsymbol{\beta}}\right]$ is positive semi-definite, we re-use two of the results on positive definite matrices that we employed in section 4.A.2:

1. if Q and P are invertible and $Q - P$ is psd, then $P^{-1} - Q^{-1}$ is also psd;
2. if Q is psd, then $P'QP$ is also psd for any matrix P.

Begin by applying property 1 above and define

$$\Delta \equiv \sigma^2 \cdot \left[\text{AV}\left[\tilde{\boldsymbol{\beta}}\right]^{-1} - \text{AV}\left[\hat{\boldsymbol{\beta}}\right]^{-1}\right] = Q - A'B^{-1}A;$$

Since $\sigma^2 > 0$, it is sufficient to prove that Δ is psd. Now define the vector $\mathbf{z}_i' = [\mathbf{x}_i' \quad \mathbf{w}_i']$ and consider the probability limit of its second moments:

$$C = \text{plim}\left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right) = \begin{bmatrix} Q & A' \\ A & B \end{bmatrix}$$

where C is clearly psd; now define H as

$$H = \begin{bmatrix} I & -A'B^{-1} \end{bmatrix}$$

so, by property 2, the product HCH' is also psd; but

$$HCH' = \begin{bmatrix} I & -A'B^{-1} \end{bmatrix} \begin{bmatrix} Q & A' \\ A & B \end{bmatrix} \begin{bmatrix} I \\ -B^{-1}A \end{bmatrix} = Q - A'B^{-1}A = \Delta,$$

and the proof is complete.

6.A.3 Covariance matrix for the Hausman test (scalar case)

Suppose we have two consistent estimators of a scalar parameter θ; call them $\hat{\theta}$ and $\tilde{\theta}$; assume also that the joint asymptotic distribution of $\hat{\theta}$ and $\tilde{\theta}$ is normal. Then,

$$\text{AV}\begin{bmatrix} \hat{\theta} \\ \tilde{\theta} \end{bmatrix} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}.$$

Consider now the statistic $\grave{\theta}(\lambda) = \lambda\hat{\theta} + (1-\lambda)\tilde{\theta}$, where $\lambda \in \mathbb{R}$. Obviously, $\grave{\theta}(\lambda)$ is also a consistent estimator for any λ:

$$\grave{\theta}(\lambda) \xrightarrow{\text{p}} \lambda\theta + (1-\lambda)\theta = \theta.$$

Its asymptotic variance is

$$\text{AV}\left[\grave{\theta}(\lambda)\right] = \begin{pmatrix} \lambda & 1-\lambda \end{pmatrix} \cdot \begin{bmatrix} a & b \\ b & c \end{bmatrix} \cdot \begin{pmatrix} \lambda \\ 1-\lambda \end{pmatrix} = \lambda^2 a + 2\lambda(1-\lambda)b + (1-\lambda)^2 c.$$

If you choose λ so that $\text{AV}\left[\grave{\theta}(\lambda)\right]$ is minimised, you get

$$\lambda^* = \frac{c-b}{a-2b+c};$$

Now, if $\hat{\theta}$ is efficient, the optimal value of λ^* must be 1, because $\grave{\theta}(\lambda^*)$ cannot be more efficient than $\hat{\theta}$, so the two statistics must coincide. But if $\lambda^* = 1$, then $a = b$. Therefore, if $\hat{\theta}$ is efficient, the joint asymptotic covariance matrix is

$$\text{AV}\begin{bmatrix} \hat{\theta} \\ \tilde{\theta} \end{bmatrix} = \begin{bmatrix} a & a \\ a & c \end{bmatrix}.$$

so

$$\text{AV}\left[\hat{\theta} - \tilde{\theta}\right] = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{bmatrix} a & a \\ a & c \end{bmatrix} \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = c - a = \text{AV}\left[\tilde{\theta}\right] - \text{AV}\left[\hat{\theta}\right].$$

6.A.4 Hansl script for the weak instrument simulation study

```
function matrix run_experiment(scalar pi)
    # function to generate the simulations

    scalar rho = 0.75       # correlation betweeen u and eps
    series w = uniform()    # generate the instrument
    scalar H = 10000        # number of replications
    matrix b = zeros(H, 2)  # allocate space for the statistics

    loop h = 1 .. H --quiet
        # the structural disturbances (unit variance)
        series eps = normal()
        # the reduced form disturbances (unit variance,
        # correlated with eps by construction)
        series u = rho * eps + sqrt(1-rho^2) * normal()
        # generate x via the first-stage equation
        series x = pi*w + u
        # generate y via the structural equation
        series y = x + eps
        # estimate beta by OLS
        ols y x --quiet
        # store OLS estimate into the 1st column of b
```

```
        b[h,1] = $coeff
        # estimate beta by IV
        tsls y x ; w --quiet
        # store IV estimate into the 2nd column of b
        b[h,2] = $coeff
    endloop

    cnameset(b, strsplit("OLS IV")) # set column names for matrix b
    return b
end function

###
### main
###

set verbose off
nulldata 400
set seed 1234     # set random seed for replicability

# run experiment with pi = 1.0 and plot results
b10 = run_experiment(1.0)
boxplot --matrix=b10 --output=display
# run experiment with pi = 0.333 and plot results
b03 = run_experiment(1/3)
boxplot --matrix=b03 --output=display
```

Chapter 7

Panel data

7.1 Introduction

So far, we have made a sharp distinction between cross-sectional and time-series datasets. In a cross-section, you observe a "screenshot" of many individuals at a certain time; a time series, instead, observes one thing through time.

In **panel** datasets, you observe *multiple* individuals (that we will generally call **units**) through time. Therefore, the typical element of a variable y will bear a double subscript: $y_{i,t}$ is the value for unit i at time t. In a parallel fashion, the explanatory variables will be indexed similarly, as $\mathbf{x}_{i,t}$. As a consequence, we merge the two conventions used earlier and assume that $i = 1 \dots n$ and $t = 1 \dots T$ so, for example, a typical excerpt of a panel dataset looks more or less like this:

id	year	y	x	z
	⋮			
451	2015	12	1	1
451	2016	14	1	0
451	2017	11	3	0
452	2010	12	5	0
452	2011	12	2	1
	⋮			

In this example, the "id" column identifies the different units, and the "year" column identifies time, so the first row shown says that the value of y for unit 451 in the year 2015 is 12, or, in formulae, $y_{451,2015} = 12$, $x_{452,2010} = 5$, and so on.

In this chapter I will use the symbol N for the total number of observations. From a practical point of view, a panel dataset may be **balanced** or unbalanced: in the former case, you observe data over a common time range for each unit, so you get valid data for each (i, t) combination and $N = n \cdot T$. Otherwise, some rows may be missing, and not all time periods are available for all units, so $N < nT$. This is the most common case in practice.

Typically, most panel datasets will contain data for many units for short time periods: this situation is normally referred to as the "large n, small T" case, but other cases are possible. For example, macroeconomists regularly deal with datasets where units are countries and the amount of data can be considerable in the time dimension. In most microeconomic applications, however, you have many individuals observed for short time spans. As we will see, this aspect becomes important for the asymptotic analysis of the estimators we have for panel datasets.

In some cases, it makes sense to consider the factors that provoke the appearance or disappearance of a certain unit in the dataset. A classic example is firms going bankrupt. Of course, these random factors may interact with the Data Generating Process is very subtle ways. This phenomenon is known as **sample attrition** and in some cases may be very relevant to the empirical analysis.
In the elementary treatment we give here, however, we assume that this issue is moot, as the factors that determine whether a unit is observable or not are completely independent from the DGP we want to study.

The importance of panel datasets has grown exponentially since the IT revolution of the 1980s-1990s: more and more datasets of this type are available, simply as a consequence of the mechanisation of databases. For example: I have been doing my weekly shopping for more than thirty years always at the same supermarket chain, and I regularly swipe my customer card each time I go. Those guys, potentially, know *everything* about my habits: what I like, what I dislike, how much I spend each week, what I buy only during a discount promotion, and so on. And they have the same information about millions of customers. Just imagine the kind of datasets giants like Amazon possess. It should be no surprise that econometricians have devoted a lot of energy into methods for panel datasets and, as always, this book will only scratch the surface. If you want to go deeper, Wooldridge (2010) is what everybody considers the ultimate reference, but in my opinion Hsiao (2022) is also a must-have.

A mechanical application of the line of thought we followed in chapter 3 would disregard the panel nature of the dataset entirely and just focus on the regression function $\mathrm{E}\left[y_{i,t}|\mathbf{x}_{i,t}\right]$. Of course this approach is possible, and leads to our usual OLS statistic, which in this context is often called the **pooled** estimator of the conditional mean parameters. While this is a technically valid procedure, it is almost never a good idea, because we can do something smarter with the information contained in the panel structure of the dataset and redefine the object of our interest (from $\mathrm{E}\left[y_{i,t}|\mathbf{x}_{i,t}\right]$ to something else), like we did in chapter 5, so as to give a much more meaningful description of the DGP.

7.2 Individual effects

Consider the balanced panel dataset displayed in Table 7.1, where you have $N = 18$ observations, pertaining to $n = 3$ different units, with $T = 6$. The application

Table 7.1: Small example panel dataset

id	time	y	x
1	1	1.6	1.6
1	2	1.0	1.8
1	3	2.2	1.0
1	4	2.0	1.0
1	5	1.8	1.0
1	6	2.2	0.8
2	1	3.2	4.2
2	2	3.4	3.2
2	3	3.0	4.2
2	4	3.6	2.4
2	5	3.8	3.2
2	6	3.2	3.6
3	1	3.8	6.8
3	2	5.0	4.8
3	3	5.2	5.4
3	4	4.6	5.8
3	5	4.4	6.0
3	6	3.6	7.0

of OLS to these data gives the "pooled" estimate of $\mathrm{E}\left[y|x\right]$, which is

$$\hat{y} = \underset{(4.41)}{1.62} + \underset{(4.97)}{0.45}\, x,$$

where you have t-ratios bewteen parentheses and an R^2 index of 60.7%. The slope parameter, our customary indicator of the relationship between x and y, equals 0.45 and is very significant (its t-ratio is 4.97). What you see is a strong, significant positive link between x and y.

Figure 7.1: Example data with OLS line

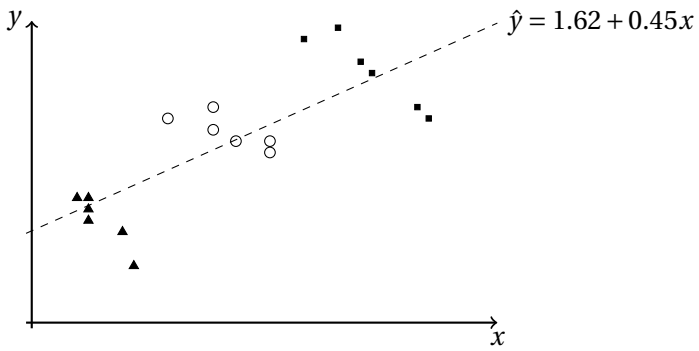


Figure 7.1 displays the data together with the fitted line, using different sym-

bols to identify different units. In this context, the model we're fitting is

$$y_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\beta} + \varepsilon_{i,t} \tag{7.1}$$

and we're not using the information we have on the different units at all.

How can we improve on the above? The idea is to introduce *heterogeneity* between units into the picture, and generalise the DGP by allowing for the possibility of each units having its own set of parameters. A fully general application of this principle would entail considering an object like

$$m_i(\mathbf{x}) = \mathrm{E}\left[y_{i,t}|\mathbf{x}_{i,t}\right]$$

(note that the regression function m has a subscript i). In principle, this approach would lead us to estimating a different regression function for each unit, which is undesirable for various reasons: first, in the typical "large n, small T" scenario, it is quite possible that T, the number of observations you have for one unit, is smaller than k, the number of parameters in your model, which would make estimation impossible. Moreover, that level of generality is not even needed. In most contexts, it is perfectly reasonable to assume that heterogeneity between units does not affect the marginal effects of $\mathbf{x}$ on y. In other words, even if individuals are different, it's often likely that the way they respond to variations in the observables is the same. If this is the case, then $\boldsymbol{\beta}$ is the same for all units and we may settle for

$$y_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\beta} + \alpha_i + \varepsilon_{i,t}, \tag{7.2}$$

where the α_i term is commonly known as the **individual effect**. We can use vectors and matrices for writing (7.2) more compactly, expressing all the observations for unit i as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha_i\boldsymbol{\iota} + \boldsymbol{\varepsilon}_i \tag{7.3}$$

where of course $\mathbf{y}$ is a $T \times 1$ vector, $\mathbf{X}$ is a $T \times k$ matrix, and $\boldsymbol{\iota}$ is, as usual, a conformable vector of ones.[1] The presence of the individual effects in equation (7.2) means that each unit is potentially different from all the others because there is a term α_i, constant through time, that shifts the level of $y_{i,t}$ by some amount. The $\boldsymbol{\beta}$ vector, instead, is homogeneous across units.

In the simplest cases, it is customary to assume that, once heterogeneity is taken into account, the disturbances are well-behaved, so the covariance matrix for the whole $\boldsymbol{\varepsilon}$ vector is $\sigma^2_\varepsilon I$, where σ^2_ε is a positive scalar and I is a $N \times N$ identity matrix. More general scenarios will be considered in section 7.3.4.

There are two main points to note about individual effects:

1. individual effects are unobservable (anything observable can be part of the set of explanatory variables);

[1] If the panel were unbalanced, each unit would have its own time span, and we should say T_i rather than T. But we'll avoid this complication.

2. individual effects are time invariant.

For example, imagine that $y_{i,t}$ is the percentage of malnourished population in country i at time t. There could be many factors that explain differences between observations, the most obvious one being GDP per capita; this is observable, so it goes into $\mathbf{x}_{i,t}$. Another one could be the fertility of soil; for the sake of the example, assume that fertility is unobservable. Clearly, the soil of each country is typical of that country. If we also assume that its characteristics don't change through time, soil fertility is one of the many possible factors that may contribute to α_i.

At this stage, we're making no assumptions on the relationship between observables and individual effects. For all we know, soil quality and per capita GDP could be related or not. Another example, close to the one I used in section 6.3, is the Mincer wage equation: if you have a panel dataset with individuals' wage and education, the individual effect could be rightfully interpreted as "unobservable ability". Is it independent of education? Maybe it is, maybe not. In the toy dataset depicted in Figure 7.1, the average value of x seems to be different across units, so one could think that observable and unobservable factors are unlikely to be independent of one another.

For the estimation of $\boldsymbol{\beta}$ in equation (7.2), there are two ways to take individual effects into account:

fixed-effects approach: treat the individual effects as parameters to estimate and make no assumptions about them.[2] This approach is described in Section 7.3.

random-effects approach: make some assumptions on the individual effects and treat them as random variables. This leads to more efficient estimators, provided certain conditions are met (but if they aren't the consequences could be catastrophic). Section 7.4 is about this.

7.3 Fixed effects

7.3.1 Using dummy variables

In this section, we treat individual effects as parameters. Therefore, a very crude way to estimate equation (7.2) is to add individual dummies, that is

$$y_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\beta} + \alpha_1 d^1_{i,t} + \alpha_2 d^2_{i,t} + \cdots + \alpha_n d^n_{i,t} + \varepsilon_{i,t} \tag{7.4}$$

[2] This approach, in principle, may lead to some complications because, apart from $\boldsymbol{\beta}$, you have n different α_i parameters to estimate. In general, when the number of parameters to estimate is not fixed, but is a function of the sample size, we may not be able to estimate consistently any of them. In the statistical literature, this is known as the **incidental parameters problem**, but fortunately in linear models (the only ones we consider here) we don't have to worry about this issue: more on this at the end of section 7.3.3.

where $d^1_{i,t}, d^2_{i,t}$ etc are a set of dummies for unit 1, 2 and so on, respectively (so no, the number near the letter d is not an exponent). Therefore, the model for unit k would simply reduce to (7.2), since for that unit $d^k_{i,t} = 1$ and all the other dummies are 0. If units differ on account of some unobserved factor that shifts the level of y for each one of them but keeps the marginal effects β equal across units, then we have an ordinary linear model in which each unit has its own intercept.

In matrix notation, eq. (7.4) would read

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{7.5}$$

where, with the vector notation used in equation (7.3), the relevant matrices look like this:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \boldsymbol{\iota} & 0 & \dots & 0 \\ 0 & \boldsymbol{\iota} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\iota} \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix};$$

note that $\mathbf{X}$ is an $N \times k$ matrix, whereas $\mathbf{D}$ is $N \times n$.[3]

As usual, the parameters we're interested in are the β vector, and the estimate you get by applying OLS to (7.4) is known as the **LSDV** (Least Squares with Dummy Variables) estimator. In principle, one could also consider the estimates of the individual effects α_i, but this is less interesting and is not done very often.

What is the interpretation of β in this context, and how different is it from its "pooled" counterpart? The estimate we get of β from equation (7.4) is the marginal effect of $\mathbf{x}$ on y *once heterogeneity between units has been taken into account.*

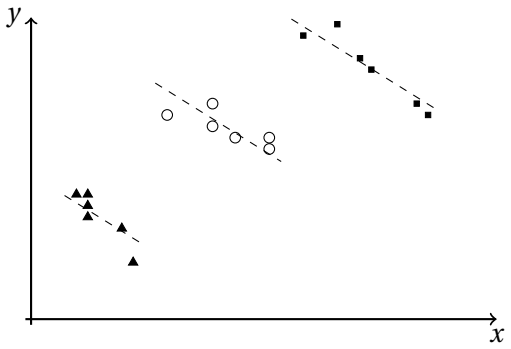
For example, the estimates you'd get by applying model (7.4) to the example dataset in Table 7.1 are

$$\hat{y}_{i,t} = \underset{(-6.24)}{-0.62}\, x_{i,t} + \underset{(16.0)}{2.54}\, d^1_{i,t} + \underset{(15.3)}{5.53}\, d^2_{i,t} + \underset{(13.5)}{8.15}\, d^3_{i,t}$$

where, again, the numbers in round brackets are t-ratios and the fitted value lines are displayed in Figure 7.2. Not only R^2 jumps to 92.8% here, but the slope coefficient changes sign (also: it's even more significant)! Do we have a contradiction here? Not really: if we look at what happens if we follow each unit through time, we have a negative association between y and x. In each individual's experience, when x goes up, y goes down (on average). However, in our example units with larger values of x generally have larger values of y, so the overall conditional mean of y on x has a positive slope, because it doesn't take into account unobservable differences between individuals. By explicitly

[3] For you linear algebra addicts: if the panel is balanced, the structure of the $\mathbf{D}$ matrix could be handled in a very elegant and effective way using a cool tool called **Kronecker product**: those interested may jump to Section 7.A.1.

Figure 7.2: Example data with FE estimate



considering individual effects, we eliminate heterogeneity and shed light on the negative relationship that each individual observes.

From a practical point of view, the insertion of the unit dummies creates a few issues. First, the regressor matrix would have $k+n$ columns, so if n is large OLS estimation involves the inversion of a disproportionately large matrix, but even that wouldn't be a serious problem for modern computers unless n is in the thousands or so. In addition, to carry out estimation you can't have a constant in your model unless you drop one of the unit dummies to avoid the collinearity problem known as the "dummy trap" (see Section 1.3.3), but apart from this, estimation is a straightforward application of OLS. In the example above, inserting a constant and dropping the dummy for unit 1 would give

$$\hat{y}_{i,t} = 2.54 - 0.62x_{i,t} + 2.98d^2_{i,t} + 5.60d^3_{i,t},$$

which is clearly equivalent (apart from rounding errors).

Finally, the possibility of time-invariant regressors raises a somewhat more delicate point: these cannot coexist with the unit dummies for collinearity reasons. The classic example is gender: if units are persons, the possibility of observing an individual changing their gender in our sample is usually very low. Therefore, one of the columns of the **X** matrix will contain, for each unit, the same value repeated from $t = 1$ to T. It is a simple exercise to prove that such a column would be a linear combination of the columns of **D**, and therefore OLS would be unfeasible. However, this issue can be circumvented by using a slightly different estimation technique, that I'll illustrate in Section 7.4.2.

One nice thing about this setup is that it makes heterogeneity testable rather easily. Assuming (without loss of generality) that we drop the individual dummy for the first unit, the the null hypothesis for the test is $H_0 : \alpha_2 = \alpha_3 = \ldots = \alpha_n = 0$, which would be equivalent to homogeneity across units. Under H_0 the preferred model would be the pooled one, so this kind of test is often termed a **poolability** test. The details of the test are unimportant: this is just a linear test on param-

eters if the $R\beta = \mathbf{d}$ form, so you have the choice of using any of the procedures described in section 3.5; of course, this is a joint test where the number of hypotheses is $n-1$.

In the simple example above the F-form of the test would yield a p-value of 7.55615e-08, so the visual impression that units 1, 2 and 3 are indeed different from each other would be strongly confirmed.

7.3.2 The "within" transformation

The LSDV approach is computationally very inefficient if you're in the typical "large n, small T" case, because of the column size of the regressor matrix.[4] Fortunately, there is a very handy approach for obtaining the same statistic in a different way. This approach also has the virtue of highlighting a few features of the estimator, and is based on the so-called **within** transformation.

The within transformation for a variable essentially amounts to subtracting the per-unit averages. For example, the within transformation for y_i is

$$\widetilde{y}_{i,t} \equiv y_{i,t} - \bar{y}_i,$$

where $\bar{y}_i$ is the average of the observations for unit i. Following most of the literature, I will use the tilde as a decoration for within-transformed variables.

The reason why this is called the "within" transfromation is motivated by a traditional decomposition of the variance of a variable. A precise definition of the decomposition of variance in "within" and "between" components is one of those pedantic things that make descriptive statistics one of the most boring things on Earth. Let's just say that the transformation above annihilates all the differences between units (the per-unit average of $\widetilde{y}_{i,t}$ is 0 by construction) and all the information that is left comes from variability within units through time (hence the name). Therefore, the within tranformation of a time-invariant variable, such as gender in the example above, gives you a vector of zeros.

The matrix representation of the within transformation is very useful: at the very beginning of this book (see Section 1.2) I showed that the average of $\mathbf{y}_i$ can be written as

$$\bar{y}_i = (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\mathbf{y}_i.$$

Therefore, we can easily compute the vector of the deviations of $\mathbf{y}_i$ from its own mean as

$$\widetilde{\mathbf{y}}_i = \mathbf{y}_i - \boldsymbol{\iota}\bar{y}_i = \mathbf{y}_i - \boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\mathbf{y}_i = \mathbf{y}_i - P\mathbf{y}_i = Q\mathbf{y}_i;$$

where I used P and Q as synonyms for $\mathbf{P}_\iota$ and $\mathbf{M}_\iota$, respectively (we'll use these matrices quite often in Section 7.4, so it's good to have a quick alternative notation; besides, I'm trying to stay consistent with the notation traditionally used in

[4] On the other hand, nothing prevents you from also adding "time dummies" for $t = 1$, $t = 2$ etc. if T really is a small number. This is actually quite common practice. Section 7.A.4 shows how this works in practice.

most textbooks).[5] The reader is invited to check that applying the within transformation to the toy example in Table 7.1 gives the data shown in Table 7.2.

Table 7.2: Within tranformation

y	x	$\bar{y}$	$\bar{x}$	$\tilde{y}$	$\tilde{x}$
1.6	1.6	1.8	1.2	-0.2	0.4
1	1.8	1.8	1.2	-0.8	0.6
2.2	1	1.8	1.2	0.4	-0.2
2	1	1.8	1.2	0.2	-0.2
1.8	1	1.8	1.2	0	-0.2
2.2	0.8	1.8	1.2	0.4	-0.4
3.2	4.2	3.367	3.467	-0.167	0.733
3.4	3.2	3.367	3.467	0.033	-0.267
3	4.2	3.367	3.467	-0.367	0.733
3.6	2.4	3.367	3.467	0.233	-1.067
3.8	3.2	3.367	3.467	0.433	-0.267
3.2	3.6	3.367	3.467	-0.167	0.133
3.8	6.8	4.433	5.967	-0.633	0.833
5	4.8	4.433	5.967	0.567	-1.167
5.2	5.4	4.433	5.967	0.767	-0.567
4.6	5.8	4.433	5.967	0.167	-0.167
4.4	6	4.433	5.967	-0.033	0.033
3.6	7	4.433	5.967	-0.833	1.033

With the help of the within transformation, we'll rewrite equation (7.2) so as to eliminate the individual effects.[6] If you average observations for unit i through time, you get

$$\bar{y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} y_{i,t} = \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{x}'_{i,t}\boldsymbol{\beta} + \alpha_i + \varepsilon_{i,t}\right] = \bar{\mathbf{x}}'_{i,t}\boldsymbol{\beta} + \alpha_i + \bar{\varepsilon}_{i,t}, \tag{7.6}$$

in obvious notation. Now subtract equation (7.6) from (7.2):

$$\tilde{y}_{i,t} = \tilde{\mathbf{x}}'_{i,t}\boldsymbol{\beta} + \tilde{\varepsilon}_{i,t} \tag{7.7}$$

and the α_i terms have disappeared. In vector form, the above would read

$$\tilde{\mathbf{y}}_i = Q\mathbf{y}_i = Q\mathbf{X}_i\boldsymbol{\beta} + Q\boldsymbol{\iota}\alpha_i + Q\boldsymbol{\varepsilon}_i = \tilde{\mathbf{X}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_i, \tag{7.8}$$

where the main simplification trivially comes from $Q\boldsymbol{\iota} = \mathbf{0}$. Naturally, we are assuming that $\mathbf{X}$ constains no time-invariant regressors, which would become columns of zeros, for the reasons given above.

[5] Here we're assuming the panel is balanced to minimise the fuss, but the extension to unbalanced panels is straighforward, as long as you admit that the P and Q matrices could have different size for different individuals.

[6] The within transofrmation is a convenient way to sweep out the α_i terms, but it's by no means the only one: considering first differences, that is $\Delta y_{i,t}$, would work just the same, with a few slight adjustments.

Intuition suggests that, having removed the individual effect by means of the within transformation, you can estimate β by applying OLS to (7.7). This is indeed the case, and the result is known, unsurprisingly, as the "within" estimator.

The amazing result is that this statistic is exactly the same as you'd get from using OLS on (7.4). The proof is quite simple if we consider the within transformation as a matrix operation: the within transformation can be expressed in matrix terms as the premultiplication of the original data by an $N \times N$ square and singular matrix that we call $\mathbf{Q}$:

$$\widetilde{\mathbf{y}} = \mathbf{Q}\mathbf{y} \qquad \widetilde{\mathbf{X}} = \mathbf{Q}\mathbf{X}.$$

The $\mathbf{Q}$ matrix is a block-diagonal matrix, where all elements on the diagonal are the Q matrices defined above, so it looks like this:

$$\mathbf{Q} \equiv \begin{bmatrix} Q & 0 & \dots & 0 \\ 0 & Q & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q \end{bmatrix}. \tag{7.9}$$

Clearly, it is also possible to define $\mathbf{P} = \mathbf{P_D}$ analogously:

$$\mathbf{P} \equiv \begin{bmatrix} P & 0 & \dots & 0 \\ 0 & P & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P \end{bmatrix}.$$

We won't need $\mathbf{P}$ now, but we'll use it later in Section 7.4. Therefore,

$$\mathbf{Q}\mathbf{y} = \begin{bmatrix} Q & 0 & \dots & 0 \\ 0 & Q & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} Q\mathbf{y}_1 \\ Q\mathbf{y}_2 \\ \vdots \\ Q\mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{y}}_1 \\ \widetilde{\mathbf{y}}_2 \\ \vdots \\ \widetilde{\mathbf{y}}_n \end{bmatrix}$$

and the algebra for $\widetilde{\mathbf{X}}$ is just the same. As a consequence, the within estimator, which is just OLS on (7.7), can be written in matrix notation as[7]

$$\hat{\beta} = (\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}\mathbf{y} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{y}}, \tag{7.10}$$

and the corresponding model is called the **within regression**.

To prove that (7.10) is just the LSDV estimator, note that $\mathbf{Q} = \mathbf{M_D}$, where $\mathbf{D}$ is the $N \times n$ matrix with all the unit dummies I used in equation (7.5) (the proof is in section 7.A.5). Therefore, equivalence between the LSDV and within estimators follows from the Frisch-Waugh theorem: the OLS estimate for equation (7.5) satisfies

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{D}\hat{\alpha} + \mathbf{e}, \tag{7.11}$$

[7] Quite evidently, $\mathbf{Q}$ is idempotent: I'll leave the proof to the reader.

so

$$\mathbf{M_D y} = \mathbf{M_D X}\hat{\boldsymbol{\beta}} + \mathbf{e} \Longrightarrow \mathbf{X}'\mathbf{M_D y} = \mathbf{X}'\mathbf{M_D X}\hat{\boldsymbol{\beta}},$$

which implies (7.10) (you may need to go back to Section 1.4.4 for the different passages).

In practice, then, the LSDV and within estimators are exactly the same thing, so they have the same interpretation. For example, if you regress $\widetilde{y}$ on $\widetilde{x}$ in Table 7.2, the OLS coefficient you get is -0.62, exactly equal to the one we found for model (7.4). You may use either term for them, or even a third alternative, possibly even more popular: the **fixed-effects** estimator, or **FE** for short, which I'll indicate by $\hat{\beta}_{FE}$.

Some readers may be troubled by the fact that the disturbances in equation (7.7) are correlated. This is easily seen by considering the vector representation (7.8): since $\widetilde{\varepsilon}_i = Q\varepsilon_i$, it follows that

$$\mathrm{V}\left[\widetilde{\varepsilon}_i\right] = Q\mathrm{V}\left[\varepsilon_i\right]Q'.$$

Even in the ideal case, where $\mathrm{V}\left[\varepsilon_i\right] = \sigma_\varepsilon^2 I$, the covariance matrix of $\widetilde{\varepsilon}_i$ would be $\mathrm{V}\left[\widetilde{\varepsilon}_i\right] = \sigma_\varepsilon^2 Q$, which is obviously non-diagonal (keep in mind that Q is symmetric and idempotent). This, however, is not a problem, since it can be proven that in this case OLS coincides with GLS, so OLS takes care of the problem quite effectively.

I'm not proving this because we'd need a slightly more sophisticated definition of GLS than I gave in chapter 4.2.1, on account of the fact that Q is singular and I'd have to use the "Moore-Penrose" inverse I hinted at in Section 1.A.4. Just trust me, OK?

With the LSDV approach, the estimates for the individual effects $\hat{\alpha}_i$ are obtained directly. However, calculating them via the within estimator is also rather easy: rewrite equation (7.11) as

$$\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{FE} = \mathbf{D}\hat{\boldsymbol{\alpha}} + \mathbf{e}.$$

If you pick a single unit, this implies

$$\mathbf{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{FE} = \boldsymbol{\iota}\hat{\alpha}_i + \mathbf{e}_i;$$

now premultiply by $\frac{1}{T}\boldsymbol{\iota}'$ and use the fact $\boldsymbol{\iota}'\mathbf{e}_i = 0$: the result is

$$\hat{\alpha}_i = \frac{1}{T}\boldsymbol{\iota}'(\mathbf{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{FE}) = \frac{1}{T}\sum_{t=1}^{T} u_{i,t}, \tag{7.12}$$

where

$$u_{i,t} = y_{i,t} - \mathbf{x}_{i,t}'\hat{\boldsymbol{\beta}}_{FE}. \tag{7.13}$$

So, all you have to do is compute the residuals you'd get by using the within estimate on the untransformed data and take means by unit.

7.3.3 Asymptotics for the FE estimator

While the meaning of the word "asymptotic" is straightforward in cross-sectional or time-series datasets, it is not so for panel data. The number of rows in our dataset N can go to infinity if either n or T do so, or both. In this book, we'll concentrate on the case when T is fixed and $n \to \infty$; the reader, however, should be aware that in more sophisticated scenarios the case $T \to \infty$ may have be considered too.

The best starting point to analyse the asymptotic behaviour of the fixed-effect estimator is to consider its LSDV representation: under the hypothesis that no heteroskedasticity or serial correlation issues arise, standard OLS inference applies to equation (7.5), and therefore $\hat{\boldsymbol{\beta}}_{FE}$ is consistent and asymptotically normal, with a limit covariance matrix given by

$$V = \sigma_\varepsilon^2 \mathrm{E}\left[\widetilde{\mathbf{x}}_{i,t} \cdot \widetilde{\mathbf{x}}'_{i,t}\right]^{-1}$$

Assuming we have a consistent estimator for σ_ε^2, then a consistent estimate of $\mathrm{V}\left[\hat{\boldsymbol{\beta}}_{FE}\right]$ is

$$\hat{V} = \hat{\sigma}_\varepsilon^2 (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1},$$

where we're keeping T fixed here, as usual. The questions of interest are:

1. Do we have a consistent estimator for σ_ε^2?
2. Is $\mathrm{E}\left[\widetilde{\mathbf{x}} \cdot \widetilde{\mathbf{x}}'\right]$ nonsingular, for $n \to \infty$?

For devising an estimator of the variance, the customary ingredient we've been using all along is the sum of squared residuals. So far, the SSR divided by the number of observations has always done the trick. In the context of FE estimation, however, things are not so simple, and the appropriate estimator to use is

$$\hat{\sigma}_\varepsilon^2 = \frac{SSR}{N-n}. \tag{7.14}$$

The reason why the denominator is different from the total number of observations is very interesting, but a bit too distracting at this point, so the interested reader should jump to Section 7.A.6.[8]

As for the asymptotic behaviour of $\frac{1}{N}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, we must assume that there is sufficient time-variability in the regressors not only to compute the estimator, but also to allow $\mathrm{E}\left[\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}'\right]$ to have full rank. Of course this excludes time-invariant regressors, since the within transformation turns them into columns of zeros, but also explanatory variables for which the within variation cannot be assumed to increase for $n \to \infty$. This is in fact a rather general point: consistency of $\hat{\boldsymbol{\beta}}_{FE}$ depends on its variance going to 0: this happens only if the *within* variation in regressors grows without bounds (the $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ matrix goes to infinity). Therefore,

[8] Most textbooks, and all the software I'm aware of, use in fact a slightly different formula, where SSN is divided by $N-n-k$ rather than $N-n$; asymptotically, it makes no difference.

if we have one or more regressors whose variation through time is limited, we shouldn't expect our estimates to have nice properties in terms of precision.

For example: suppose you have a dataset with 1000 individuals and you observe two of them changing their gender. In this case, the gender dummy becomes technically time-varying, so you can use it for fixed-effect estimation. However, you can't expect your estimates to be particularly precise, as your $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ matrix will be near-singular. Moreover, in order to consider your estimator as consistent, you must be willing to assume that, in principle, the number of transgender people in your sample would increase as n grows (which could be reasonable or not).

Therefore, as long as our object of interest is inference on $\boldsymbol{\beta}$, we can just happily use the within regression, provided we make the necessary adjustment to our estimator of σ^2_ε. If we wanted to make inference on the individual effects, instead, things are not so simple, since in the usual "large n, small T" scenario, the estimates of α_i you get from LSDV are *not* consistent for $n \to \infty$. This is easy to see by considering equation (7.12): the statistic $\hat{\alpha}_i$, is calculated on the T observations we have the i-th individual, so its variance is not a function of n at all, and n going to infinity has no effect on the distribution of $\hat{\alpha}_i$. Fortunately, this is not a problem for estimating $\boldsymbol{\beta}$, because our estimator $\hat{\boldsymbol{\beta}}_{FE}$ doesn't depend on the estimated individual effects. Nevertheless, the reader should be aware that in statistical models where the object of interest is not a linear regression function, it may be impossible to estimate the parameters of interest separately from the individual effects, and inconsistency may be a very serious problems. This is the so-called "incidental parameters" problem I hinted at a few pages back.

7.3.4 Heteroskedasticity and dependence between observations

In fact, we could allow for greater generality by considering several extensions: the first one that comes to mind is heteroskedasticity, with unit-specific variances for $\varepsilon_{i,t}$. This is not a particularly serious problem, since appropriate adaptations of the robust estimators à la White (see Section 4.2.2) are quite simple and effective.

More worryingly: if we consider equation (7.2), it is clear that the hypothesis $V(\boldsymbol{\varepsilon}) = \sigma^2_\varepsilon I$ implies that, apart from the individual effects, all observations are incorrelated with each other. This includes observations pertaining to the same unit at different times. In many cases, this could be unrealistic.

In fact, time persistence can be a very likely possibility, since model (7.2) is almost certain to neglect some time-varying unobservable factors that evolve gradually through time. By applying the sample logic as in chapter 5, we could allow for some kind of ADL structure in equation (7.2). The kind of models you'd get are normally called **dynamic panel** models, and have become increasingly popular since the late 1980s. However, inference is considerably more complex than in the static models we consider here: the tool that is almost invariably used is the Generalised Method of Moments (GMM), which you can think of

as a generalisation of the IV technique that chapter 6 is about. The obligatory reference for these cases is again Wooldridge (2010), but Biørn (2017) is also very good.

A different possibility for dealing with the persistence issue is to stick with the static formulation (7.2) and assume that any kind of time-dependence between observations can be accommodated via correlation between disturbances. It turns out that, if this is the case, the fixed-effects estimator $\hat{\boldsymbol{\beta}}_{FE}$ is consistent under a fairly large spectrum of conditions. The only problem is, like in static models with heteroskedasticity, that in order to perform inference correctly, the covariance matrix for $\hat{\boldsymbol{\beta}}_{FE}$ needs an appropriate adjustment. This leads us to the idea of **clustered** covariance matrix.

A full description of cluster-robust inference has no place in this book; suffice it to say that you divide your observations in observable groups called **clusters**, and you allow the $\varepsilon_{i,t}$ random variable to be arbitrarily correlated inside the group. The variable which tells you which group an observation belongs to is called the "clustering" variable.

Of course, the most obvious choice for clustering is the variable indexing units (the "id" variable in Table 7.1). In this case, equation (7.3) would be generalised so as to allow the covariance matrix to be pretty much anything, instead of a scalar matrix:

$$\mathrm{V}[\boldsymbol{\varepsilon}_i] = \mathrm{E}\left[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i'\right] = \boldsymbol{\Sigma}_i.$$

Note that the covariance matrix bears the subscript i, so we're also implicitly allowing for arbitrary forms of heteroskedasticity. Other choices, however, are possible. For example, it may be not unrealistic to imagine that some correlation may exist across different units: the classic example is a panel where units are geographical entities, where units are regions and clusters are countries, but one could also think of individuals belonging to the same household, firms in the same sector, etc. Let me just say that the literature on this topic has exploded in the past 15 years, and that Cameron and Miller (2010) or Cameron and Miller (2015) provide excellent surveys.

7.4 Random effects

The basic idea that gives rise to the **random effects** estimator (abbreviated as **RE**) is that in some cases we may be willing to put some structure on the individual effects, rather than being completely agnostic about them as we do in fixed-effects estimation.

Since individual effects are taken to represent a heap of time invariant, unobserved, and possibly very diverse factors that describe how units differ from one another, it'd be natural to describe the α_i terms as random variables. If we assume the existence of moments, then the assumption $\mathrm{E}[\alpha_i] = 0$ implies no loss of generality, and $\mathrm{V}[\alpha_i] = \sigma_\alpha^2$ is nothing more than a mild regularity condi-

tion. With these assumptions, then equation (7.2) can be rewritten as

$$y_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\beta} + \omega_{i,t} \tag{7.15}$$

where $\omega_{i,t} = \alpha_i + \varepsilon_{i,t}$. The same equation for unit i can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\omega}_i \tag{7.16}$$

where $\boldsymbol{\omega}_i \equiv \alpha_i\boldsymbol{\iota} + \boldsymbol{\varepsilon}_i$. By making the harmless assumption that $\mathrm{E}\left[\alpha_i\varepsilon_{i,t}\right] = 0$ for all i and t, the covariance matrix of $\boldsymbol{\omega}_i$ equals

$$\Sigma = \mathrm{V}[\boldsymbol{\omega}_i] = \mathrm{E}\left[\boldsymbol{\omega}_i\boldsymbol{\omega}'_i\right] = \mathrm{V}[\boldsymbol{\varepsilon}_i] + \sigma^2_\alpha\boldsymbol{\iota}\boldsymbol{\iota}' = \sigma^2_\varepsilon I + \sigma^2_\alpha\boldsymbol{\iota}\boldsymbol{\iota}', \tag{7.17}$$

where the last equality comes from the assumption that the disturbances are well-behaved. If we also assume independence between units, equation (7.15) for the whole sample would therefore become

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\omega} \tag{7.18}$$

where

$$\mathrm{V}[\boldsymbol{\omega}] = \Omega = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma \end{bmatrix} \tag{7.19}$$

is a block-diagonal matrix.

We are now in the position to substantiate the claim I made at the end of Section 7.1, when I said that using the pooled OLS estimator is almost never a good idea with a panel dataset: for a start, the covariance matrix of the disturbance term is not scalar, which suggests that even though OLS on (7.15) was consistent, valid inference requires at least with some form of robust covariance matrix estimation (see Section 4.2.2). Besides, consistency itself may be at risk: even if $\mathrm{E}[\alpha_i] = 0$, there is no guarantee that α_i and $\mathbf{x}_i$ should be independent, or at least incorrelated (see the discussion at the end of Section 7.2). If $\mathrm{E}[\alpha_i|\mathbf{x}_i] \neq 0$, if follows that $\mathrm{E}\left[\omega_{i,t}|\mathbf{x}_i\right] \neq 0$ and therefore $\mathrm{E}\left[y_{i,t}|\mathbf{x}_{i,t}\right] \neq \mathbf{x}'_{i,t}\boldsymbol{\beta}$: the classic endogeneity problem we analysed in Chapter 6, that renders the pooled estimator inconsistent.

In the light of these two possible problems, what could an effective strategy be? Let's put the endogeneity issue aside for the moment (we'll come back to it in section 7.4.2). If $\mathrm{E}[\alpha_i|\mathbf{x}_i] = 0$, one may conjecture that OLS should be more efficient than the FE estimator, since the FE estimator uses only the "within" variation in the data, but we could use the "between" information (that is, differences between units) to gain some efficiency.

In fact, we can do even better than OLS: from equation (7.17), the structure of the covariance matrix of Σ is known, bar two scalars, σ^2_ε and σ^2_α. Therefore,

if these two scalars were known, we could use the GLS estimator, described in Section 4.2.1), which I'm reproducing here for your convenience:

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X}'\Omega^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}.$$

This solution would take care of two problems at once: we'd be using the most efficient estimator possible, and we wouldn't have to worry about robust inference. In practice, the two variances we need to get the job done are unknown, but asymptotically consistent estimators would be just as good, so a FGLS estimator would be available. This is what we call the RE estimator.

It turns out that, as often happens, once the original data are suitably modified, the RE estimator can be rewritten as OLS on the transformed data. The transformation we need is known as "quasi-differencing": for each observation, we subtract a fraction of the per-unit average from the original data:

$$\check{y}_{i,t} = y_{i,t} - \theta\bar{y}_i,$$

where $0 \leq \theta \leq 1$. In vector form,

$$\check{\mathbf{y}}_i = \mathbf{y}_i - \theta\boldsymbol{\iota}\bar{y}_i = (I - \theta P)\,\mathbf{y}_i,$$

where, again, P is an alias for $\mathbf{P}_{\boldsymbol{\iota}}$. Quasi-differencing for the whole sample can be written as

$$\check{\mathbf{y}} = (I - \theta\mathbf{P})\,\mathbf{y},$$

where $\mathbf{P}$ was defined in section 7.3.2 or, equivalently, as

$$\check{\mathbf{y}} = [\mathbf{Q} + (1-\theta)\mathbf{P}]\,\mathbf{y}, \tag{7.20}$$

given that $\mathbf{Q} = I - \mathbf{P}$

For a given value of θ, the RE estimator is just OLS on the quasi-differenced data, that is

$$\grave{\boldsymbol{\beta}}(\theta) = \left[\check{\mathbf{X}}'\check{\mathbf{X}}\right]^{-1}\check{\mathbf{X}}'\check{\mathbf{y}}. \tag{7.21}$$

As is easy to check, quasi-differencing with $\theta = 1$ is just the within transformation, so $\grave{\boldsymbol{\beta}}(1) = \hat{\boldsymbol{\beta}}_{FE}$. At the other end of the spectrum, where $\theta = 0$, the original data are unmodified, so $\grave{\boldsymbol{\beta}}(0)$ is the just pooled OLS estimator. Note that, for $\theta < 1$, time-invariant variables do *not* become zero, and so they are perfectly useable.

Derivation of the optimal choice of θ for GLS estimation is a bit techincal, and is in Section 7.A.7 for those interested. Here, I'm just giving you the solution straight away, which is

$$\theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}. \tag{7.22}$$

Note that, when σ_ε^2 is large compared to σ_α^2, θ will be near 0: heterogeneity between units is negliglible and the optimal estimator is practically OLS. Conversely, if σ_ε^2 is very small compared to σ_α^2, then θ is close to 1 and the within

estimator is optimal, since all the variance in $\omega_{i,t}$ comes from individual effects, which are eliminated by the within transformation.

As I said earlier, the two variances σ_ε^2 and σ_α^2 are unknown in practice, so they must be estimated. The almost universal solution is to use FE for σ_ε^2; as for σ_α^2, there are various alternatives and it is not clear if the "best" one even exists. Shortly after the RE estimator was invented, in the late 1960s, quite a lot of work was devoted to this issue, and the method most software uses is the one by Swamy and Arora (1972), but you should be aware that you may get different results form different programs because different (equally defensible) methods are adopted.

Anyway: once we have consistent estimates of σ_ε^2 and σ_α^2, to compute FGLS we just plug them into equation (7.22) and obtain

$$\hat{\theta} = 1 - \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2 + T\hat{\sigma}_\alpha^2}}, \tag{7.23}$$

a consistent estimate of θ. By using $\hat{\theta}$, we can compute the quasi-differenced data $\check{\mathbf{y}}$ and $\check{\mathbf{X}}$ and, finally, compute $\hat{\boldsymbol{\beta}}_{RE}$ as OLS on the quasi-differenced data. And that is what we call the RE estimator:[9]

$$\hat{\boldsymbol{\beta}}_{RE} = \check{\boldsymbol{\beta}}(\hat{\theta}). \tag{7.24}$$

7.4.1 The Hausman test

Having dealt with the particular covariance structure of $\omega_{i,t}$, we now turn to the other issue I mentioned above, that is the possibility of the observables $\mathbf{x}_{i,t}$ being correlated with the individual effect α_i. In many cases, this is a very real possibility: think about the example I used on page 211, where GDP per capita is one of the regressors $\mathbf{x}_{i,t}$ and soil fertility is one of the things that go into the individual effect α_i: who says that GDP per capita and soil fertility are independent? More generally, it's easy to imagine other examples, such as unobserved ability and schooling in a Mincer wage equation.

As I argued above, this is not a problem for the FE estimator, since the within transformation just sweeps the individual effect away, but it would make OLS and the RE estimator inconsistent. Therefore, we could compare the FE and RE estimators to see if they are similar, much in the same way as we did in Section 6.4 when we compared the OLS and IV estimators. This comparison gave rise to the "Hausman test", and this case is just the same. In fact, the original article (Hausman, 1978) uses exactly the two examples we have in this book, that is OLS vs IV and RE vs FE.

[9] As the reader might imagine, robust versions of the RE estimator exists, but I'll refrain from illustrating the details, and I'll just say that there is no additional worry compared to the FE case, and they work as one would expect.

What should we expect from the comparison? $\hat{\boldsymbol{\beta}}_{FE}$ is robust but inefficient; $\hat{\boldsymbol{\beta}}_{RE}$ is efficient but potentially inconsistent.[10] Under the null hypothesis of no correlation between $\mathbf{x}_{i,t}$ and α_i, the difference

$$\boldsymbol{\delta} = \hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE}$$

should converge to 0 in probability, because both statistics share the same limit. Conversely, large values of $\boldsymbol{\delta}$ should be taken as an indicator of endogeneity of $\mathbf{x}_{i,t}$.

The Hausman test can be carried out in a variety of ways, some numerically equivalent, some only asymptotically. A choice that is used by several software packages is to perform an auxiliary regression of the form

$$\check{\mathbf{y}}_i = \check{\mathbf{X}}_i\boldsymbol{\beta} + \widetilde{\mathbf{X}}_i\boldsymbol{\gamma} + \mathbf{u}_i, \tag{7.25}$$

and then a Wald test for the hypothesis $H_0 : \boldsymbol{\gamma} = \mathbf{0}$. With a bit of algebra, it can be proven that this is equivalent to $\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE} \xrightarrow{\text{p}} \mathbf{0}$.

Therefore, the course of action to take is straightforward: after RE estimation, look at the Hausman test. If the null is rejected, $\hat{\boldsymbol{\beta}}_{RE}$ is probably inconsistent, and $\hat{\boldsymbol{\beta}}_{FE}$ is preferable. Otherwise, we may happily use $\hat{\boldsymbol{\beta}}_{RE}$, which is better than $\hat{\boldsymbol{\beta}}_{FE}$ because it's more efficient. As simple as that.

7.4.2 Correlated Random Effects, aka "the Mundlak trick"

An alternative strategy for dealing with the possible correlation between the regressors $\mathbf{x}_{i,t}$ and the individual effect α_i comes from modelling explicitly the correlation between them. This gives rise to an estimator sometimes called the **correlated random effects** estimator, or **CRE** for short, proposed first by Mundlak (1978). As we will see shortly, however, the result will be less exciting than one may hope, but side benefits will be substantial.

The key intuition is to consider the conditional expectation of α_i to $\bar{\mathbf{x}}_i$ and assume it is a linear function,

$$\mathrm{E}[\alpha_i|\bar{\mathbf{x}}_i] = \bar{\mathbf{x}}_i'\boldsymbol{\gamma}; \tag{7.26}$$

note that the conditioning variable we're using here is not $\mathbf{x}_{i,t}$, but rather its average through time. Since α_i is time-invariant, it is quite natural to assume that a time average of the $\mathbf{x}_{i,t}$ should capture the effect we're after.

Therefore, if you define $u_i = \alpha_i - \bar{\mathbf{x}}_i'\boldsymbol{\gamma}$ you can re-write (7.2) as

$$y_{i,t} = \mathbf{x}_{i,t}'\boldsymbol{\beta} + \bar{\mathbf{x}}_i'\boldsymbol{\gamma} + u_i + \varepsilon_{i,t} = y_{i,t} = \mathbf{x}_{i,t}'\boldsymbol{\beta} + \bar{\mathbf{x}}_i'\boldsymbol{\gamma} + \eta_{i,t}, \tag{7.27}$$

where, by construction, none of the two error terms u_i and $\varepsilon_{i,t}$ is correlated with the explanatory variables. In vector form,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + P\mathbf{X}_i\boldsymbol{\gamma} + \boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} + \bar{\mathbf{X}}_i\boldsymbol{\gamma} + \boldsymbol{\eta}_i.$$

[10] Naturally, we have both parameters only for time-varying regressors, so the comparison is limited to the subset of $\hat{\boldsymbol{\beta}}_{RE}$ that matches $\hat{\boldsymbol{\beta}}_{FE}$.

If you substitute α_i with u_i, and therefore $\omega_{i,t}$ in equation (7.15) with $\eta_{i,t}$ in equation (7.27), you see that the structure of the covariance matrix of $\boldsymbol{\eta}_i$ is absolutely identical, so nothing stops you from using FGLS on equation (7.27). Therefore, in practice, Mundlak's CRE estimator is just the RE estimator with the time averages of $\mathbf{X}_i$ as additional regressors.

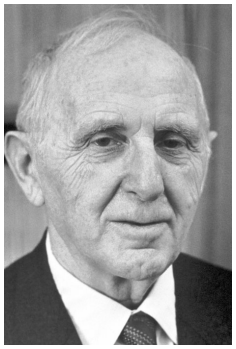
The fist thing to say is that the estimate of $\boldsymbol{\beta}$ you get is nothing new. With a little bit of matrix algebra, it can be proven (I do it in Section 7.A.8) that the estimated $\boldsymbol{\beta}$ vector is numerically equal to the within estimator $\hat{\boldsymbol{\beta}}_{FE}$. Therefore, one may think that the Mundlak procedure is just a tortuous avenue to get something we already had. Not quite: one nice thing of the CRE estimator is that it provides us with a nice way to use time-invariant explanatory variables, which is impossible with the LSDV or the within approaches.

Moreover, testing the hypothesis $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ is very interesting: under the null, the endogeneity problem just goes away. Therefore, rejection of the null would imply we have to stick with FE, but otherwise we could gain efficiency and go with RE. It should come as no surprise that testing this hypothesis is equivalent to the Hausman test I described in the previous subsection.

7.5 An example with real data

7.5.1 The Kuznets curve

The American economist Simon Kuznets (Nobel prize winner in 1971) is credited with an idea that has become known as the "Kuznets curve". In short, the basic intuition is that developing economies go through several structural changes that provoke an increase in inequality in the early stages, and a decrease later. Clearly, this idea is too mechanical and simplistic to paint an accurate picture, but if there is something to it, we should observe that inequality is highest in middle-income economies.



SIMON KUZNETS

I collected some data from the World Bank's WDI database: per capita income and the Gini index (the standard measure of income inequality) for the years between 2008 and 2022.[11]

As often happens in these cases, the panel is heavily unbalanced. We have lots of data for some economies, but for some countries we only have one or two datapoints. Having said this, our dataset comprises 1044 observations for 157 countries. A scatterplot of the Gini index versus log GDP per capita is shown in Figure 7.3.

The curve you see in the figure is the fitted line from a pooled OLS regression of the Gini index versus GDP per capita (in logs) and its square. I added to

[11] In the interest of replicability: the measure of GDP per capita I used is GDP per capita in constant 2015 US$ (WDI code: `NY.GDP.PCAP.KD`.) The WDI code for the Gini index is `SI.POV.GINI`.

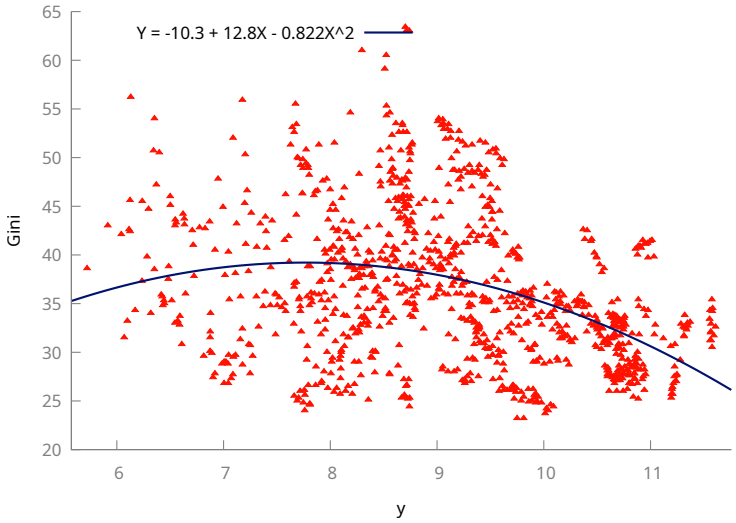


Figure 7.3: The Kuznets curve

this model a dummy for European countries, since these countries have had a historical and cultural preference for social equality that some people consider a dangerous socialist drift. The results are shown in Table 7.3.

```
Pooled OLS, using 1044 observations
Included 157 cross-sectional units
Time-series length: minimum 1, maximum 15
Dependent variable: Gini

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const      -15.2335      7.78943       -1.956   0.0508   *
  y           12.7708      1.74755        7.308   5.40e-13 ***
  y2          -0.712896    0.0969748     -7.351   3.97e-13 ***
  Europe      -9.35281     0.426650     -21.92    7.23e-88 ***

Mean dependent var   36.39588   S.D. dependent var   7.629774
Sum squared resid    35233.57   S.E. of regression   5.820518
R-squared            0.419705   Adjusted R-squared   0.418031
F(3, 1040)           250.7304   P-value(F)           2.1e-122
```

Table 7.3: The Kuznets curve: OLS estimates

Here we seem to have a confirmation of Kuznets' hypothesis: the curvature is negative (the coefficient for y2 is negative and significant) and the distribution of income for European countries is confirmed to be more even than other countries with similar levels of GDP per capita.

However, this is a pooled estimate, conceptually similar to the plot I showed you earlier, in Figure 7.1. Is it possible that the results we are seeing neglect the effect of unobserved heterogeneity between countries. Therefore, we turn to FE

estimates.

7.5.2 Fixed-effects estimates

A word of warning on the presence of a constant in the FE estimate. Strictly speaking, the intercept is a time-invariant regressor, so it should not appear in the FE output. However, most econometric software (including gretl, which is what I'm using) adopt a slightly different convention on the definition of the matrix **D** in (7.5), so that an intercept is in fact calculated.[12]

```
Fixed-effects, using 1044 observations
Included 157 cross-sectional units
Time-series length: minimum 1, maximum 15
Dependent variable: Gini
Omitted due to exact collinearity: Europe

            coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       69.4905      20.2572       3.430    0.0006  ***
  y           -0.546958     4.60603     -0.1187   0.9055
  y2          -0.331499     0.260585    -1.272    0.2037

Mean dependent var   36.39588   S.D. dependent var   7.629774
Sum squared resid    2687.375   S.E. of regression   1.742579
LSDV R-squared       0.955739   Within R-squared     0.129623
LSDV F(158, 885)     120.9498   P-value(F)           0.000000
Log-likelihood      -1974.926   Akaike criterion     4267.851
Schwarz criterion    5055.031   Hannan-Quinn         4566.408
rho                  0.615823   Durbin-Watson        0.582907

Test for differing group intercepts -
  Null hypothesis: The groups have a common intercept
  Test statistic: F(155, 885) = 69.1486
  with p-value = P(F(155, 885) > 69.1486) = 0
```

Table 7.4: The Kuznets curve: fixed-effects estimates

Having said this, Table 7.4 is relatively straightforward to comment: the "Europe" dummy drops out of the equation on account of it being time-invariant, as explained in Section 7.3.1. Moreover, the the poolability test rejects the null very strongly (the p-value is so small that the software just prints 0). This means that heterogeneity between units (countries in this case) is substantial and a pooled model may yield misleading results, as long as we're interested in the effect of GDP on inequality. In fact, the Kuznets curve simply disappears: the coefficients on per capita GDP and its square are not significant.

Nevertheless, it can be verified that the joint hypothesis of both coefficients being zero delivers a very small p-value (2.09218e-27): dropping the quadratic

[12] The difference amounts to modifying the withing transformation by adding back, for each observation, the overall mean: $\tilde{y}_{i,t} = y_{i,t} - \bar{y}_i + \bar{y}$.

term gives a marginal effect of -6.36334, with a t-statistic of -11.41:[13] it seems we do have a uniformly inverse relationship, instead of a concave curve.

Therefore, having eliminated the variation between countries, what we observe is the relationship between GDP and inequality *through time*: if we concentrate on the individual history of each country, we observe that on average inequality decreases with economic growth, instead of the "inverted-U" relationship described by Kuznets.

One last thing to note is that the estimated value for the first-order autocorrelation of residuals $\hat{\rho}$ is 0.6158, so we have a substantial autocorrelation problem. In principle, we should go for a dynamic model, but here we're following the easier route of just using cluster-robust standard errors, by unit. That is, we employ a different estimator for the variance of $\hat{\beta}_{FE}$, that permits (a) arbitrary correlation through time between observations for the same country and (b) heteroskedasticity between countries.

	coefficient	std. error	t-ratio	p-value	
const	69.4905	40.3405	1.723	0.0869	*
y	-0.546958	8.90973	-0.06139	0.9511	
y2	-0.331499	0.491751	-0.6741	0.5012	

Table 7.5: The Kuznets curve: fixed-effects estimates with robust standard errors

As can be seen in table 7.5, the estimated standard errors are quite different from Table 7.4. This is in fact a very common phenomenon: while it is very rare in cross-sectional models that robust inference delivers substantially divergent results from plain estimation, it panel dataset clustering by unit almost always inflates standard errors by a great deal, and the interpretation of results may have be adjusted, even radically.

In this case, however, the meaning conveyed by the model stays the same: the Kuznets curve vanishes, although the joint test still rejects the null (the p-value is 1.14766e-06) and the conclusions are the same.

7.5.3 Random-effects estimates

Having established that heterogeneity between countries is something we cannot ignore, maybe we could gain efficiency by using the RE estimator (Section 7.4); to be on the safe side, I'll use cluster-robust inference.

Note that in this case the quasi-differencing operation I described in section 7.4 is a little bit more complicated, because the panel is heavily unbalanced and you have different numbers of observations for different countries. Equation (7.22) contains the symbol T, so what should we use here? The solution is to adopt a different θ for countries with different numbers of observations, so data

[13] I'm not reporting the whole restricted regression for the sake of brevity. This chapter is already long enough. You can try it yourself if you want.

for each unit are quasi-differenced using

$$\hat{\theta}_i = 1 - \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2 + T_i \hat{\sigma}_\alpha^2}};$$

(note the "i" subscript). Clearly, (7.22) is just a special case of the equation above, that applies to balanced panels where $T_i = T$ for all units. Gretl reports the average value of θ used, which is 0.85298.

```
Random-effects (GLS), using 1044 observations
Included 157 cross-sectional units
Time-series length: minimum 1, maximum 15
Dependent variable: Gini
Standard errors clustered by unit

            coefficient   std. error     z      p-value
  ------------------------------------------------------
  const       18.7662      17.9769       1.044   0.2965
  y            7.38777      4.16427      1.774   0.0760  *
  y2          -0.579803     0.237245    -2.444   0.0145  **
  Europe      -3.53635      1.32306     -2.673   0.0075  ***

Mean dependent var   36.39588   S.D. dependent var   7.629774
Sum squared resid    48078.17   S.E. of regression   6.795925
Log-likelihood      -3480.511   Akaike criterion     6969.022
Schwarz criterion    6988.825   Hannan-Quinn         6976.533
rho                  0.615823   Durbin-Watson        0.582907

'Between' variance = 39.864
'Within' variance = 3.03658
mean theta = 0.85298
corr(y,yhat)^2 = 0.256021

Breusch-Pagan test -
  Null hypothesis: Variance of the unit-specific error = 0
  Asymptotic test statistic: Chi-square(1) = 3065.99
  with p-value = 0

Hausman test -
  Null hypothesis: GLS estimates are consistent
  Asymptotic test statistic: Chi-square(2) = 22.3276
  with p-value = 1.4178e-05
```

Table 7.6: The Kuznets curve: random-effects estimates

Comparing $\hat{\beta}_{FE}$ with $\hat{\beta}_{RE}$, we observe a striking difference:

variable	FE	RE
y	-0.547	7.388
y2	-0.331	-0.588

It looks as if the two estimates should come out as significantly unlike one another, and this is indeed the case: the Hausman test rejects quite strongly (p-value = 1.4178e-05), so it's unlikely that the two estimators converge to the same

probability limit. This is what happens when one or more of the explanatory variables (presumably GDP per capita, in our case) is correlated with the individual effect α_i, on account of the endogeneity problem that this provokes. In cases like these, the RE estimator is inconsistent, so we'd better stay with FE.

Finally, note that gretl (like all other software packages do) reports a test as the **Breusch-Pagan test**. This is a test for the hypothesis $H_0 : \sigma_\alpha^2 = 0$: under the null, the individual effects are in fact not even random variables at all, because they have zero mean and zero variance, so $\alpha_i = 0$ for all units. Therefore, it can be seen as the random-effects equivalent to the poolability test I described earlier. In this case, the null is strongly rejected (the p-value is so small that the software just says 0), so we see that heterogeneity is substantial, again. Note that this is an entirely different test from the BP test for heteroskedasticity I mentioned earlier in Section 4.2.3. The two tests share the same authors, but the similarity stops there.

7.5.4 Correlated random effects

```
Random-effects (GLS), using 1044 observations
Included 157 cross-sectional units
Time-series length: minimum 1, maximum 15
Dependent variable: Gini
Standard errors clustered by unit

            coefficient   std. error      z       p-value
  ------------------------------------------------------------
  const        8.96061     16.6681       0.5376    0.5909
  y           -0.546958     8.92260     -0.06130   0.9511
  y2          -0.331499     0.492462    -0.6731    0.5009
  Europe      -7.90436      1.20454     -6.562     5.30e-11 ***
  Py           8.10053     10.4543       0.7748    0.4384
  Py2         -0.118126     0.575871    -0.2051    0.8375
```

Table 7.7: The Kuznets curve: CRE estimates

The final estimate we see is the CRE estimate (see Section 7.4.2). There's hardly anything to see here: the coefficients for the time-varying variables $y_{i,t}$ and $y_{i,t}^2$ are absolutely identical to those in Table 7.5, as they should; their standard errors are not exactly the same, but that's a consequence of using robust SEs. If we had used plain GLS standard errors, they would have been identical too; the difference is minor anyway. So, for the time-varying variables we have nothing more than the FE estimate, and the interpretation is obviously the same. On the contrary, the CRE technique allows us to keep the time-invariant dummy for Europe in the model, which is (unsurprisingly) negative and significant.

Finally, note the insertion of the two "Mundlak" extra regressors, labelled `Py` and `Py2` in the table, which contain the per-unit averages of $y_{i,t}$ and $y_{i,t}^2$, respectively. Although they are not significant individually, an F test for joint significance of the two "Mundlak" extra regressors yields 11.1638, with a p-value

of 1.59597e-05, which is (unsurprisingly) nearly identical to the Hausman test shown in table 7.6.

7.A Assorted results

In this chapter, I used several matrix algebra concepts and results that had not been necessary before. Therefore, this section starts with a quick and rudimentary treatment of a few linear algebra topics. For more details, see Lütkepohl (1996), Abadir and Magnus (2005) or Horn and Johnson (2012).

7.A.1 The Kronecker product

The usual way of multiplying two matrices, where $C = AB$ comes from taking all possible inner products of the rows of A and the columns of B is not the only way to define a way of multiplying two matrices.

An alternative is provided by the so-called **Kronecker product**, also known as **tensor product**, which is defined as follows. Take two matrices A and B, and sat that A is $r \times c$ and B is $m \times n$. Then their Kronecker product $A \otimes B$ is a matrix with $r \cdot m$ rows and $c \cdot n$ columns, in which each element of A is multiplied by the whole matrix B.

$$A \otimes B = \begin{bmatrix} a_{1,1}B & a_{1,2}B & \dots & a_{1,c}B \\ a_{2,1}B & a_{2,2}B & \dots & a_{2,c}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{r,1}B & a_{r,2}B & \dots & a_{r,c}B. \end{bmatrix}$$

Note that, as a consequence of its definition, with the Kronecker product no conformability issues arise. On the other hand, like with ordinary matrix product, Kronecker product is not commutative: $A \otimes B \neq B \otimes A$.

The Kronecker product has many nice properties, but the only ones we will need concern their combination with transposition, inversion and the ordinary matrix product. It can be proven that

$$(A \otimes B)' = A' \otimes B' \qquad (A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \qquad (A \otimes B)(C \otimes D) = (AC \otimes BD)$$

Note: the last equality assumes that the matrices are conformable.

In many cases, the Kronecker product makes it much easier to work with "large matrices with a structure". For example, if the panel is balanced the $\mathbf{D}$ matrix defined in equation (7.5) can be written as $\mathbf{D} = I \otimes \iota$ and the variance of ω in equation (7.19) is $\mathbb{V}[\omega] = I \otimes \Sigma$, where I is $n \times n$; unfortunately, with unbalanced panels such elegance is unattainable.

Finally: the "vec" operator I illustrated in Section 4.A.3 and the Kronecker product play together very nicely. The basic property you need to know is that

$$\text{vec}(ABC) = \left(C' \otimes A\right) \text{vec}(B),$$

so for example if

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad B = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad C = \begin{bmatrix} 3 & 6 & 9 \end{bmatrix}$$

you may verify that

$$ABC = \begin{bmatrix} -3 & -6 & -9 \\ -3 & -6 & -9 \end{bmatrix}$$

so

$$\text{vec}(ABC) = \begin{bmatrix} -3 \\ -3 \\ -6 \\ -6 \\ -3 \\ -9 \\ -9 \end{bmatrix}$$

which is equal to

$$\left(C' \otimes A\right) \text{vec}(B) = \begin{bmatrix} 3 & 6 \\ 9 & 12 \\ 6 & 12 \\ 18 & 24 \\ 9 & 18 \\ 27 & 36 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

7.A.2 The trace operator

Given a square matrix C with n rows and columns, the trace operator is defined simply as

$$\text{tr}(C) = \sum_{i=1}^{n} C_{i,i},$$

that is, the sum of all the elements on the diagonal. Clearly, the trace of a scalar is the scalar itself.

This operator is useful in many contexts, mostly related to the fact that, for any given $r \times c$ matrix A (possibly, with $r \neq c$),

$$\text{tr}\left(A'A\right) = \sum_{i=1}^{r} \sum_{j=1}^{c} A_{i,j}^2.$$

The two notable properties of the trace operator we use in our context are:

Linearity : $\text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)$, and it is also true that $\text{tr}(\lambda C) = \lambda \cdot \text{tr}(C)$, where λ is a scalar. Note that linearity implies that the trace and expectation operators can be interchanged: if C is a random matrix,

$$\text{E}\left[\text{tr}(C)\right] = \text{tr}\left(\text{E}\left[C\right]\right).$$

Commutation : $\text{tr}(AB) = \text{tr}(BA)$, which implies the amusing property I like to call the "train" property:

$$\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA);$$

that is: the argument of the trace operator is like a train, where you can detach a wagon from one end and stick it to the other end. For example, of $\mathbf{x}$ is a vector, the trace of the $\mathbf{x}\mathbf{x}'$ matrix can be computed very easily as

$$\text{tr}\left(\mathbf{x}\mathbf{x}'\right) = \text{tr}\left(\mathbf{x}'\mathbf{x}\right) = \mathbf{x}'\mathbf{x},$$

with the second equality coming from $\mathbf{x}'\mathbf{x}$ being a scalar.

7.A.3 A neat matrix inversion trick

Suppose P is idempotent and $Q = I - P$; therefore Q is idempotent too and $PQ = QP = [0]$. Assume that a matrix A can be written as

$$A = \alpha P + \beta Q,$$

where α and β are nonzero scalars. Then, there is an amazingly simple way to write the inverse of A:

$$A^{-1} = \frac{1}{\alpha}P + \frac{1}{\beta}Q.$$

The proof is by direct multiplication:

$$\left(\alpha P + \beta Q\right)\left(\frac{1}{\alpha}P + \frac{1}{\beta}Q\right) = \frac{\alpha}{\alpha}P + \frac{\beta}{\beta}Q = P + Q = I$$

because $PQ = QP = [0]$ by construction and the cross-products drop out.

Note that, by the same logic, it's also possible to compute the "inverse square root" of A, that is a matrix that gives A^{-1} when multiplied by itself:

$$A^{-1/2} = \frac{1}{\sqrt{\alpha}}P + \frac{1}{\sqrt{\beta}}Q,$$

and again, the proof is by direct multiplication. In fact, the result could be generalised to any exponent k:

$$A^k = \alpha^k P + \beta^k Q.$$

7.A.4 Time dummies

The addition of time dummies to a fixed-effect model is straightforward, and amounts to adding to the dataset a set of T dummies identifying time periods; actually, you normally add $T-1$ to avoid the dummy trap.

Therefore, equation (7.4) would become, after dropping the dummies for unit 1 and time 1,

$$y_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\beta} + \alpha_2 d^2_{i,t} + \cdots + \alpha_n d^n_{i,t} + \gamma_2 t^2_{i,t} + \cdots + \gamma_T t^T_{i,t} + \varepsilon_{i,t};$$

this model is often called the **two-way** fixed-effects model. In the toy dataset in Table 7.1, this would give:

id	time	y	x	t^2	t^3	...	t^6
1	1	1.6	1.6	0	0	...	0
1	2	1	1.8	1	0	...	0
1	3	2.2	1	0	1	...	0
1	4	2	1	0	0	...	0
1	5	1.8	1	0	0	...	0
1	6	2.2	0.8	0	0	...	1
2	1	3.2	4.2	0	0	...	0
2	2	3.4	3.2	1	0	...	0
2	3	3	4.2	0	1	...	0
2	4	3.6	2.4	0	0	...	0
2	5	3.8	3.2	0	0	...	0
2	6	3.2	3.6	0	0	...	1
3	1	3.8	6.8	0	0	...	0
3	2	5	4.8	1	0	...	0
3	3	5.2	5.4	0	1	...	0
3	4	4.6	5.8	0	0	...	0
3	5	4.4	6	0	0	...	0
3	6	3.6	7	0	0	...	1

Note that in the "large n, small T" scenario the number of dummies you use is in fact relatively small, and does not create any computational problem. From the viewpoint of the interpretation of results, the effect you have is that in your estimate you not only get rid of heterogeneity across units, but also across time periods. This is especially useful when some unobserved factor affects all units in a given period. For example, imagine your dataset describes turnover by firms and includes year 2020: surely you'll want to control for the COVID pandemic, since it's reasonable to assume that it affected most, if not all, the units you observe.

Alternatively, you may want to economise on the number of regressors used to clean unobservable time effects by using a time trend, and possibly its square. How advisable this is depends on the data you have.

7.A.5 Proof that $\mathbf{Q} = \mathbf{M_D}$

Here we assume that the panel is balanced for simplicity, although the unbalanced case would be completely analogous and the conclusion would be the same, but the algebra would be somewhat messier. The $\mathbf{Q}$ matrix, defined in equation (7.9) and repeated here for convenience, is:

$$\mathbf{Q} \equiv \begin{bmatrix} Q & 0 & \dots & 0 \\ 0 & Q & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q \end{bmatrix}.$$

Now we prove that $\mathbf{Q}$ is in fact $\mathbf{M_D}$: first, note that $\mathbf{D}'\mathbf{D} = T \cdot \mathbf{I}$:

$$\begin{bmatrix} \boldsymbol{\iota}' & 0 & \dots & 0 \\ 0 & \boldsymbol{\iota}' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\iota}' \end{bmatrix} \begin{bmatrix} \boldsymbol{\iota} & 0 & \dots & 0 \\ 0 & \boldsymbol{\iota} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\iota} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\iota}'\boldsymbol{\iota} & 0 & \dots & 0 \\ 0 & \boldsymbol{\iota}'\boldsymbol{\iota} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\iota}'\boldsymbol{\iota} \end{bmatrix} = T \cdot \mathbf{I}$$

Therefore, $(\mathbf{D}'\mathbf{D})^{-1} = \frac{1}{T}\mathbf{I}$. As a consequence,

$$\mathbf{P_D} = \frac{1}{T}\mathbf{D}\mathbf{D}' = \frac{1}{T} \begin{bmatrix} \boldsymbol{\iota} & 0 & \dots & 0 \\ 0 & \boldsymbol{\iota} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\iota} \end{bmatrix} \begin{bmatrix} \boldsymbol{\iota}' & 0 & \dots & 0 \\ 0 & \boldsymbol{\iota}' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\iota}' \end{bmatrix} = \begin{bmatrix} P & 0 & \dots & 0 \\ 0 & P & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P \end{bmatrix}.$$

and so

$$\mathbf{M_D} = \mathbf{I} - \mathbf{P_D} = \begin{bmatrix} I-P & 0 & \dots & 0 \\ 0 & I-P & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & I-P \end{bmatrix} = \begin{bmatrix} Q & 0 & \dots & 0 \\ 0 & Q & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q \end{bmatrix} = \mathbf{Q}.$$

A more compact proof can be given by using the Kronecker product, described in Section 7.A.1: with a balanced panel dataset one can write $\mathbf{D}$ as $I \otimes \boldsymbol{\iota}$, where I is $n \times n$ and $\boldsymbol{\iota}$ is $T \times 1$, and therefore

$$\mathbf{P_D} = (I \otimes \boldsymbol{\iota}) \left[(I \otimes \boldsymbol{\iota})'(I \otimes \boldsymbol{\iota}) \right]^{-1} (I \otimes \boldsymbol{\iota}') = (I \otimes \boldsymbol{\iota}) \left[T \cdot I \right]^{-1} I \otimes \boldsymbol{\iota}' = \frac{1}{T}(I \otimes \boldsymbol{\iota}\boldsymbol{\iota}') = I \otimes P;$$

as a consequence,

$$\mathbf{M_D} = I \otimes (I - P) = I \otimes Q = \mathbf{Q},$$

as claimed.

7.A.6 The estimator of the variance in the within regression

In order to derive equation (7.14), we need to proceed in steps. Again, I'll assume that our panel is balanced for simplicity, but this restriction could be easily dropped at the cost of more cumbersome notation.

First, let's define the residuals from the within regression as

$$u_{i,t} = \widetilde{y}_{i,t} - \widetilde{\mathbf{x}}_{i,t}\hat{\boldsymbol{\beta}}_{FE}.$$

Now note that the SSR from the within regression can be written as

$$SSR_W = \sum_{i=1}^{n}\sum_{t=1}^{T} u_{i,t}^2 = \sum_{i=1}^{n} \mathbf{u}_i'\mathbf{u}_i;$$

if we maintain independence between units, the rightmost expression is the sum of n independent rv, and the probability limit of

$$\frac{1}{n}SSR_W = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}_i'\mathbf{u}_i$$

should just equal $\mathrm{E}\left[\mathbf{u}_i'\mathbf{u}_i\right]$.

On the other hand, consistency of $\hat{\boldsymbol{\beta}}_{FE}$ implies that the within residuals $u_{i,t}$ converge to the centred disturbances $\ddot{\varepsilon}_{i,t}$ as $n \to \infty$, ans so, by extension, does the whole vector for a single unit

$$\mathbf{u}_i \xrightarrow{\mathrm{p}} \ddot{\boldsymbol{\varepsilon}}_i.$$

Therefore, one may say that, for $n \to \infty$, $\mathrm{E}\left[\mathbf{u}_i'\mathbf{u}_i\right]$ should converge to $\mathrm{E}\left[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i\right]$ and therefore

$$\frac{1}{n} SSR_W = \xrightarrow{\mathrm{p}} \mathrm{E}\left[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i\right]$$

This limit can be computed by noting that $\ddot{\boldsymbol{\varepsilon}}_i = Q\boldsymbol{\varepsilon}_i$, and so, by using the properties of the trace operator (if you're not 100% confident on the trace operator, section 7.A.2 is for you):

$$\ddot{\boldsymbol{\varepsilon}}_i'\ddot{\boldsymbol{\varepsilon}}_i = \mathrm{tr}\left(\ddot{\boldsymbol{\varepsilon}}_i'\ddot{\boldsymbol{\varepsilon}}_i\right) = \mathrm{tr}\left(\ddot{\boldsymbol{\varepsilon}}_i\ddot{\boldsymbol{\varepsilon}}_i'\right) = \mathrm{tr}\left(Q\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'Q\right).$$

The expected value of the above is

$$\mathrm{E}\left[\mathrm{tr}\left(Q\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'Q\right)\right] = \mathrm{tr}\left(\mathrm{E}\left[Q\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'Q\right]\right) = \mathrm{tr}\left(Q\mathrm{E}\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\right]Q\right) = \mathrm{tr}\left(\sigma_\varepsilon^2 QQ\right) = \sigma_\varepsilon^2 \mathrm{tr}\left(Q\right)$$

As for the trace of Q, note that $Q = \mathbf{M}_{\boldsymbol{\iota}}$, so

$$\mathrm{tr}\left(Q\right) = \mathrm{tr}\left(I\right) - \mathrm{tr}\left(\boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\right) = T - \mathrm{tr}\left((\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{\iota}\right) = T - 1$$

so, finally

$$\mathrm{E}\left[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i\right] = (T-1)\sigma_\varepsilon^2.$$

By combining results, it's easy to see that

$$\frac{SSR_W}{n} \xrightarrow{\mathrm{p}} (T-1)\sigma_\varepsilon^2$$

and therefore a consistent estimator of σ_ε^2 is provided by

$$\hat{\sigma}_\varepsilon^2 = \frac{SSR_W}{n(T-1)} = \frac{SSR_W}{N-n}.$$

7.A.7 The RE estimator as FGLS

Let's begin with a brief restatement of what a GLS estimator is: suppose we have a model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \mathrm{V}\left[\boldsymbol{\varepsilon}\right] = \Omega;$$

we need a matrix H such that

$$H\Omega H' = kI, \tag{7.28}$$

where k is some arbitrary positive scalar, then we could transform the model above by premultiplying everything by H:

$$H\mathbf{y} = H\mathbf{X}\boldsymbol{\beta} + H\boldsymbol{\varepsilon} = \check{\mathbf{y}} + \check{\mathbf{X}}\boldsymbol{\beta} + \check{\boldsymbol{\varepsilon}}.$$

It's easy to check that the covariance matrix of $\check{\varepsilon}$ is $HV[\varepsilon]H' = kI$, so the transformed model is homoskedastic and OLS on the transformed data is efficient and standard inference applies. The GLS estimator is therefore

$$\tilde{\boldsymbol{\beta}} = (\check{\mathbf{X}}'\check{\mathbf{X}})^{-1}\check{\mathbf{X}}'\check{\mathbf{y}} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y},$$

where the second equality comes from

$$\Omega^{-1} = (1/k)H'H,$$

which I'm not proving, but it's easy enough for the reader to demonstrate as an exercise.

The matrix Ω is in our case given in equation (7.19), but in fact the peculiar structure of the matrix implies that all we need to do is find a transformation for the model *for each individual*, that is equation (7.16) (reported here for convenience):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\omega}_i;$$

As argued above (see equation (7.17)), the covariance matrix of $\boldsymbol{\omega}_i$ is[14]

$$\Sigma = \sigma_\varepsilon^2 I + \sigma_\alpha^2 \boldsymbol{\iota}\boldsymbol{\iota}'$$

therefore, a simple solution to the GLS problem lies in finding a matrix H such that $H\Sigma H'$ is a scalar multiple of the identity matrix or, equivalently, a matrix H such that $H'H$ is a scalar multiple of Σ^{-1}.

In order to do so, it is useful to rewrite Σ in terms of the idempotent matrices P and Q:

$$\Sigma = \sigma_\varepsilon^2 I + \sigma_\alpha^2 \boldsymbol{\iota}\boldsymbol{\iota}' = \sigma_\varepsilon^2 Q + (\sigma_\varepsilon^2 + T\sigma_\alpha^2)P = \sigma_\varepsilon^2\left[Q + \frac{\sigma_\varepsilon^2 + T\sigma_\alpha^2}{\sigma_\varepsilon^2}P\right]$$

Therefore, via the result shown in Section 7.A.3, it's easy to see that the appropriate matrix H is the "inverse square root of Σ", $H = \Sigma^{-1/2}$, that can be written (apart from the σ_ε^2 scalar) as

$$\begin{aligned} H &= Q + \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}P = \\ &= (I - P) + \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}P = \\ &= I + \left(\sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}} - 1\right)P = I - \theta P \end{aligned}$$

where

$$\theta \equiv 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}.$$

[14] As usual, I'm using the convenient simplification of assuming that the dataset is balanced and you have T observations for each unit. Again, generalisation to unbalanced panels is possible but somewhat messier.

7.A.8 Proof that CRE yields FE

As noted in Section 7.4 (equation (7.20)), the quasi-differenced version of $\mathbf{y}$ can be written as

$$\check{\mathbf{y}} = [\mathbf{Q} + (1-\theta)\mathbf{P}]\,\mathbf{y} = \widetilde{\mathbf{y}} + (1-\theta)\bar{\mathbf{y}}.$$

It also follows that

$$\mathbf{Q}\check{\mathbf{y}} = \mathbf{Q}\,[\mathbf{Q} + (1-\theta)\mathbf{P}]\,\mathbf{y} = \mathbf{Q}\mathbf{y} = \widetilde{\mathbf{y}} \tag{7.29}$$

$$\mathbf{P}\check{\mathbf{y}} = \mathbf{P}\,[\mathbf{Q} + (1-\theta)\mathbf{P}]\,\mathbf{y} = (1-\theta)\mathbf{P}\mathbf{y} = (1-\theta)\bar{\mathbf{y}} \tag{7.30}$$

and analogous expressions trivially apply to $\mathbf{X}$. Now write the augmented model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \bar{\mathbf{X}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

and apply quasi-differencing so that GLS is just OLS on the transformed model

$$\check{\mathbf{y}} = \check{\mathbf{X}}\boldsymbol{\beta} + \bar{\mathbf{X}}[(1-\theta)\cdot\boldsymbol{\gamma}] + \check{\boldsymbol{\varepsilon}}.$$

To find the estimate of $\boldsymbol{\beta}$, use the Frisch-Waugh theorem:

$$\hat{\boldsymbol{\beta}} = \left[\check{\mathbf{X}}'\mathbf{M}_{\bar{\mathbf{X}}}\check{\mathbf{X}}\right]^{-1}\check{\mathbf{X}}'\mathbf{M}_{\bar{\mathbf{X}}}\check{\mathbf{y}}.$$

From equation (7.30), it follows that

$$\check{\mathbf{X}}'\bar{\mathbf{X}} = \mathbf{X}'\,[\mathbf{Q} + (1-\theta)\mathbf{P}]\,\mathbf{P}\mathbf{X} = (1-\theta)\bar{\mathbf{X}}'\bar{\mathbf{X}}.$$

and therefore

$$\check{\mathbf{X}}'\mathbf{M}_{\bar{\mathbf{X}}} = \check{\mathbf{X}}' - \check{\mathbf{X}}'\bar{\mathbf{X}}(\bar{\mathbf{X}}'\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}' = \left[\widetilde{\mathbf{X}}' + (1-\theta)\bar{\mathbf{X}}'\right] - (1-\theta)\bar{\mathbf{X}}' = \widetilde{\mathbf{X}}'$$

so

$$\hat{\boldsymbol{\beta}} = \left[\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right]^{-1}\widetilde{\mathbf{X}}'\check{\mathbf{y}} = \hat{\boldsymbol{\beta}}_{FE}$$

where the last equality comes from writing $\widetilde{\mathbf{X}}'\check{\mathbf{y}}$ as $\mathbf{X}'\mathbf{Q}\check{\mathbf{y}}$ and applying (7.29).

Bibliography

ABADIR, K. AND J. MAGNUS (2005): *Matrix Algebra*, Cambridge University Press.

ANDERSEN, H. AND B. HEPBURN (2016): "Scientific Method," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, Metaphysics Research Lab, Stanford University, summer 2016 ed.

ANDREWS, I., J. H. STOCK, AND L. SUN (2019): "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11.

ANGRIST, J. D. AND J.-S. PISCHKE (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press.

AXLER, S. (2015): *Linear algebra done right*, Springer, 2nd ed.

BIAU, D. J., B. M. JOLLES, AND R. PORCHER (2009): "P value and the theory of hypothesis testing: an explanation for new researchers," *Clinical orthopaedics and related research*, 468, 885–892.

BIERENS, H. J. (2011): *Introduction to the Mathematical and Statistical Foundations of Econometrics*, Cambridge University Press.

BILLINGSLEY, P. (1986): *Probability and Measure*, Wiley series in probability and mathematical statistics, John Wiley and Sons, 2nd ed.

BIØRN, E. (2017): *Econometrics of panel data: Methods and applications*, Oxford University Press.

BROCKWELL, P. AND R. DAVIS (1991): *Time Series: Theory and Methods*, Springer Series in Statistics, Springer.

CAMERON, A. C. AND D. L. MILLER (2010): "Robust inference with clustered data," Tech. rep., Working paper.

——— (2015): "A practitioner's guide to cluster-robust inference," *Journal of human resources*, 50, 317–372.

CAMERON, A. C. AND P. K. TRIVEDI (2005): *Microeconometrics*, Cambridge University Press.

CARD, D. (1999): "The causal effect of education on earnings," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 3, Part A, chap. 30, 1801–1863, 1 ed.

CASELLA, G. AND R. L. BERGER (2002): *Statistical inference*, Duxbury Pacific Grove, CA, 2nd ed.

DADKHAH, K. (2011): *Foundations of mathematical and computational economics. 2nd ed.*, Berlin: Springer, 2nd ed. ed.

DAVIDSON, J. (1994): *Stochastic limit theory: An introduction for econometricians*, Oxford University Press.

——— (2000): *Econometric Theory*, Wiley-Blackwell.

DAVIDSON, J., D. HENDRY, F. SRBA, AND S. YEO (1978): "Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom," *Economic Journal*, 88, 661–92.

DAVIDSON, R. AND J. G. MACKINNON (1993): *Estimation and inference in econometrics*, Oxford University Press.

——— (2004): *Econometric theory and methods*, Oxford University Press New York.

DIXIT, A. K. (1990): *Optimization in economic theory*, Oxford University Press.

DURLAUF, S. AND L. BLUME (2008): *The New Palgrave Dictionary of Economics*, Palgrave Macmillan UK.

EFRON, B. AND T. HASTIE (2016): *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press, 1st ed.

EPPERSON, J. F. (2013): *An Introduction to Numerical Methods and Analysis*, Wiley Publishing, 2nd ed.

FANAEE-T, H. AND J. GAMA (2014): "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence*, 2, 113–127.

FREEDMAN, D. AND P. STARK (2016): "What is the chance of an earthquake?" Tech. Rep. 611, Department of Statistics, University of California, Berkeley.

GALLANT, R. A. (1997): *An Introduction to Econometric Theory*, Princeton University Press.

GALTON, F. (1886): "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.

GOURIEROUX, C. AND A. MONFORT (1995): *Statistics and Econometric Models*, Cambridge University Press.

GRILICHES, Z. (1976): "Wages of Very Young Men," *Journal of Political Economy*, 84, 69–85.

HALL, A. (2005): *Generalized Method of Moments*, Advanced texts in econometrics, Oxford University Press.

HANSEN, B. E. (2019): "Econometrics," https://www.ssc.wisc.edu/~bhansen/econometrics/.

HANSEN, L. P. AND T. J. SARGENT (2013): *Recursive Models of Dynamic Linear Economies*, no. 10141 in Economics Books, Princeton University Press.

HAUSMAN, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

HAYASHI, F. (2000): *Econometrics*, Princeton: Princeton Univ. Press.

HILL, R., W. CARTER, E. GRIFFITHS, AND G. LIM (2018): *Principles of Econometrics*, John Wiley and Sons, 5th ed.

HORN, R. A. AND C. R. JOHNSON (2012): *Matrix Analysis*, Cambridge University Press, 2nd ed.

HSIAO, C. (2022): *Analysis of Panel Data*, Econometric Society Monographs, Cambridge University Press, 4 ed.

KING, G. AND M. E. ROBERTS (2015): "How robust standard errors expose methodological problems they do not fix, and what to do about it," *Political Analysis*, 23, 159–179.

LÜTKEPOHL, H. (1996): *Handbook of matrices*, John Wiley and Sons.

LÜTKEPOHL, H. (2005): *New introduction to multiple time series analysis*, Springer.

MACKINNON, J. G. (2006): "Bootstrap Methods in Econometrics," *Economic Record*, 82, S2–S18.

MACKINNON, J. G., M. Ø. NIELSEN, AND M. D. WEBB (2023): "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 272–299.

MARSAGLIA, G. (2004): "Evaluating the Normal Distribution," *Journal of Statistical Software*, 11, 1–11.

MULLAINATHAN, S. AND J. SPIESS (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

MUNDLAK, Y. (1978): "On the pooling of time series and cross section data," *Econometrica*, 69–85.

POPPER, K. R. (1968): *Conjectures and Refutations: The Growth of Scientific Knowledge.*, New York: Harper & Row.

RUUD, P. A. (2000): *An introduction to classical econometric theory*, Oxford University Press.

SIMS, C. A. (1972): "Money, Income, and Causality," *American Economic Review*, 62, 540–552.

SPANOS, A. (1999): *Probability theory and statistical inference: econometric modeling with observational data*, Cambridge University Press.

STAIGER, D. AND J. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

SWAMY, P. A. V. B. AND S. S. ARORA (1972): "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models," *Econometrica*, 40, 261–275.

THURMAN, W. N. AND M. E. FISHER (1988): "Chickens, Eggs, and Causality, or Which Came First?" *American Journal of Agricultural Economics*, 70, 237–238.

VERBEEK, M. (2017): *A Guide to Modern Econometrics*, John Wiley and Sons, 5th ed.

WASSERSTEIN, R. L. AND N. A. LAZAR (2016): "Editorial," *The American Statistician*, 70, 129–133.

WHITE, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 817–838.

——— (1994): *Estimation, Inference and Specification Analysis*, Cambridge University Press.

WILLIAMS, D. (1991): *Probability with Martingales*, Cambridge University Press.

WOOLDRIDGE, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, 2nd ed.