Basic Econometrics

Riccardo (Jack) Lucchetti

June 10, 2020
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. However, let me recommend appendix A in Hansen (2019).

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. For a more substantial read, it's hard not to recommend, again, the
I don't think I have to justify point number three. I am writing this in 2020, 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.

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 Stefano Fachin, Francesca Mariani, Giulio Palomba, Luca Pedini, Matteo Picchio and Claudia Pigini 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. Chapters 4 and 5 contain extensions of the basic model for cross-sectional and time-series datasets, respectively. Finally, chapter 6 deals with instrumental variable estimation. 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.

This work is licensed under a Creative Commons “Attribution-ShareAlike 3.0 Unported” license.
Contents

Foreword ............................................. i

1 OLS: algebraic and geometric properties 1
  1.1 Models ......................................... 1
  1.2 The average ..................................... 3
  1.3 OLS as a descriptive statistic .......... 6
    1.3.1 OLS on a dummy variable ....... 6
    1.3.2 The general case ....................... 10
    1.3.3 Collinearity and the dummy trap ...... 13
    1.3.4 Nonlinearity ......................... 15
  1.4 The geometry of OLS ..................... 16
    1.4.1 Projection matrices ................. 18
    1.4.2 Measures of fit .................... 20
  1.5 The Frisch-Waugh theorem ............ 23
  1.6 An example ............................... 25
  1.A Assorted results ...................... 27
    1.A.1 Matrix differentiation rules ...... 28
    1.A.2 Vector spaces ..................... 29
    1.A.3 Rank of a matrix ................. 30
    1.A.4 Rank and inversion ............. 31
    1.A.5 Step-by-step derivation of the sum of squares function ... 32
    1.A.6 Numerical collinearity .......... 33
    1.A.7 Definiteness of square matrices .... 34
    1.A.8 A few more results on projection matrices .... 34
    1.A.9 Proof of equation (1.21) ........ 36

2 Some statistical inference 39
  2.1 Why do we need statistical inference? .. 39
  2.2 A crash course in probability .......... 40
    2.2.1 Probability and random variables ... 41
    2.2.2 Independence and conditioning .... 43
    2.2.3 Expectation ......................... 44
    2.2.4 Conditional expectation .......... 46
  2.3 Estimation ................................ 47
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1 Consistency</td>
<td>48</td>
</tr>
<tr>
<td>2.3.2 Asymptotic normality</td>
<td>50</td>
</tr>
<tr>
<td>2.4 Hypothesis Testing</td>
<td>55</td>
</tr>
<tr>
<td>2.4.1 The $p$-value</td>
<td>59</td>
</tr>
<tr>
<td>2.5 Identification</td>
<td>61</td>
</tr>
<tr>
<td>2.A Assorted results</td>
<td>63</td>
</tr>
<tr>
<td>2.A.1 Jensen's lemma</td>
<td>63</td>
</tr>
<tr>
<td>2.A.2 The normal and $\chi^2$ distributions</td>
<td>64</td>
</tr>
<tr>
<td>2.A.3 Gretl script to reproduce example 2.5</td>
<td>68</td>
</tr>
<tr>
<td>3 Using OLS as an inferential tool</td>
<td>69</td>
</tr>
<tr>
<td>3.1 The regression function</td>
<td>69</td>
</tr>
<tr>
<td>3.2 Main statistical properties of OLS</td>
<td>72</td>
</tr>
<tr>
<td>3.2.1 Consistency</td>
<td>72</td>
</tr>
<tr>
<td>3.2.2 Asymptotic normality</td>
<td>73</td>
</tr>
<tr>
<td>3.3 Specification testing</td>
<td>75</td>
</tr>
<tr>
<td>3.3.1 Tests on a single coefficients</td>
<td>75</td>
</tr>
<tr>
<td>3.3.2 More general tests</td>
<td>77</td>
</tr>
<tr>
<td>3.4 Example: reading the output of a software package</td>
<td>79</td>
</tr>
<tr>
<td>3.4.1 The top table: the coefficients</td>
<td>80</td>
</tr>
<tr>
<td>3.4.2 The bottom table: other statistics</td>
<td>82</td>
</tr>
<tr>
<td>3.5 Exogeneity and causal effects</td>
<td>83</td>
</tr>
<tr>
<td>3.6 Restricted Least Squares and hypothesis testing</td>
<td>84</td>
</tr>
<tr>
<td>3.6.1 Two alternative test statistics</td>
<td>87</td>
</tr>
<tr>
<td>3.7 Prediction</td>
<td>90</td>
</tr>
<tr>
<td>3.8 The so-called “omitted-variable bias”</td>
<td>92</td>
</tr>
<tr>
<td>3.A Assorted results</td>
<td>95</td>
</tr>
<tr>
<td>3.A.1 Consistency of $\hat{\sigma}^2$</td>
<td>95</td>
</tr>
<tr>
<td>3.A.2 The Gauss-Markov theorem</td>
<td>95</td>
</tr>
<tr>
<td>3.A.3 Derivation of RLS</td>
<td>97</td>
</tr>
<tr>
<td>3.A.4 Asymptotic properties of the RLS estimator</td>
<td>99</td>
</tr>
<tr>
<td>4 Diagnostic testing in cross-sections</td>
<td>101</td>
</tr>
<tr>
<td>4.1 Diagnostics for the conditional mean</td>
<td>102</td>
</tr>
<tr>
<td>4.1.1 The RESET test</td>
<td>102</td>
</tr>
<tr>
<td>4.1.2 Interactions and the Chow test</td>
<td>104</td>
</tr>
<tr>
<td>4.2 Heteroskedasticity and its consequences</td>
<td>109</td>
</tr>
<tr>
<td>4.2.1 If $\Sigma$ were known</td>
<td>110</td>
</tr>
<tr>
<td>4.2.2 Robust estimation</td>
<td>112</td>
</tr>
<tr>
<td>4.2.3 White's test</td>
<td>114</td>
</tr>
<tr>
<td>4.2.4 So, in practice…</td>
<td>117</td>
</tr>
<tr>
<td>4.A Assorted results</td>
<td>119</td>
</tr>
<tr>
<td>4.A.1 Proof of equation (4.2)</td>
<td>119</td>
</tr>
<tr>
<td>Chapter</td>
<td>Section</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>4.A</td>
<td>2</td>
</tr>
<tr>
<td>4.A</td>
<td>3</td>
</tr>
<tr>
<td>4.A</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Dynamic Models</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>7</td>
</tr>
<tr>
<td>5.A</td>
<td>Assorted results</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Instrumental Variables</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>7.1</td>
</tr>
<tr>
<td></td>
<td>7.2</td>
</tr>
<tr>
<td>6.A</td>
<td>Assorted results</td>
</tr>
</tbody>
</table>
6.A.1  Asymptotic properties of the IV estimator  . . . . . . . . . . 186
6.A.2  Proof that OLS is more efficient than IV  . . . . . . . . . . 188
6.A.3  Covariance matrix for the Hausman test (scalar case)  . . . 188
6.A.4  Hansl script for the weak instrument simulation study  . . . 189

Bibliography 191
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 the same data points, we also have a vector of $k$ elements containing auxiliary data possibly helpful in better understanding the differences between the $y_i$s; we call these explanatory variables,² or $x_i$ in symbols.³ To continue the previous examples, $x_i$ 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).

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

$$y_i = m(x_i)$$

¹“All models are wrong, but some are useful” (G. E. P. Box).
²Terminology 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.
³I will almost always use boldface symbols to indicate vectors.
where we implicitly assume that if $x_i$ is not too different from $x_j$, 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.\footnote{The philosophically inclined reader may at this point google for “Bertrand Russel’s turkey”.} 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 $x_i$. In this case, the function becomes a constant:

$$y_i \approx m(x_i) = m$$

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 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 $y_i$s, 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
$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \mathbf{1}' \mathbf{y}, \quad (1.1)$$

where $\mathbf{1}$ 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 $\mathbf{1}$ 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.

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 $g(\mathbf{y}) = g(m \cdot \mathbf{1})$. 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 but parsimonious description of the whole data set, the question that naturally arises is: how much information is lost?

---

\(^5\)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”.
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, \( \mathbf{e} = \mathbf{y} - \mathbf{1} \cdot \mathbf{m} \).

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:

\[
\hat{m} = \text{Argmin}_{m \in \mathbb{R}} L(m) = \text{Argmin}_{m \in \mathbb{R}} C(\mathbf{y} - \mathbf{1} \cdot m) \quad (1.2)
\]

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

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 does it look like? I can’t just google ‘loss function’, can I?” True. Apart from extraordinary cases when the loss function is a natural consequence of the problem itself, 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.

---

\(^6\)Note for international readers: in the Italian academic systems, 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 \( \mathbf{0} \cdot \mathbf{1} = \mathbf{0} \).

\(^8\)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.
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

$$L(m) = \sum_{i=1}^{n} (y_i - m)^2 = (y - \mu \cdot m)'(y - \mu \cdot m) = e'e$$

(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 $m$;

$$\frac{dL(m)}{dm} = \sum_{i=1}^{n} \frac{d(y_i - m)^2}{dm} = -2 \sum_{i=1}^{n} (y_i - m)$$

The derivative must be 0 for a minimum, so that

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

which in turn implies

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

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

$$L(m) = (y - \mu \cdot m)'(y - \mu \cdot m) = y'y - 2m'\mu'y + m^2\mu\mu',$$

so the derivative is

$$\frac{dL(m)}{dm} = -2\mu'y + 2m'\mu' = -2\mu'(y - \mu \cdot m) = 0$$

whence

$$\mu'y = (\mu'\mu) \cdot \hat{m} \implies \hat{m} = (\mu'\mu)^{-1} \mu'y = \hat{Y}$$

because of course $\mu'\mu = n$. The value of $L(m)$ at the minimum, that is $L(\hat{m}) = e'e = \sum_{i=1}^{n} (y_i - \hat{Y})^2$ 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 is

\[ \hat{m} = \arg \min_{m \in \mathbb{R}} L(m) ; \]

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.

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 \( y \) are subject to some preliminary transformation.

In practice: the average of \( z \), 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\(^8\) 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 \( z_i = I(y_i \in A) \), where \( I(\cdot) \) is the so-called “indicator function”, that gives 1 when its argument is true and 0 when false. Evidently, the average of the \( z_i \), \( \bar{Z} \), is the relative frequency of \( E \). 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 \( x_i \) back in. For the moment, let’s consider the special case where \( x_i \) 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 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

---

8I’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.
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.

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

\[
\tilde{Y}_m = \frac{\sum_{x_i=1} y_i}{n_m} = \frac{S_m}{n_m} \quad \tilde{Y}_f = \frac{\sum_{x_i=0} y_i}{n_f} = \frac{S_f}{n_f}
\]

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

\[
m(x_i) = m_m x_i + m_f (1 - x_i)
\]

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

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

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, \( m(x_i) = m_m x_i + m_f (1 - x_i) \).

\(^{10}\) 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.
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
\[ y = X\beta + e, \quad (1.5) \]
where
\[ \beta = \begin{bmatrix} m_f \\ m_m - m_f \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \]
and \( X \) is a matrix with \( n \) rows and 2 columns; the first column is \( x \) and the second one is \( x \). The \( i \)-th row of \( X \) 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 \( \beta \) that minimises the loss function \( e'e \). The solution is not difficult: find the solutions to
\[
\frac{d}{d\beta} e'e = \frac{d}{d\beta} (y-X\beta)'(y-X\beta) = \frac{d}{d\beta} (y'y - 2\beta'X'y + \beta'X'X\beta) = 0
\]
By using the well-known rules for matrix differentiation, you have
\[ X'y = X'X\hat{\beta}, \quad (1.6) \]
What we have to do now is solve equation (1.6) for \( \hat{\beta} \). The solution is unique if \( X'X \) is invertible (if you need a refresher on matrix inversion, and related matters, subsection 1.A.3 is for you):
\[ \hat{\beta} = (X'X)^{-1}X'y. \quad (1.7) \]
Equation (1.7) is the single most important equation in this book, and this is why I framed it into a box. The vector \( \hat{\beta} \) is defined as the vector that minimises the sum of squared residuals:
\[ \hat{\beta} = \text{Argmin}_{\beta \in \mathbb{R}^2} e'e, \]
\[ 11 \text{ Need I remind the reader of the rule for transposing a matrix product, that is } (AB)' = B'A'? Obviously not. \]
\[ 12 \text{ Not so well-known, maybe? Jump to subsection 1.A.1.} \]
and the expression in equation (1.7) turns the implicit definition into an explicit formula that you can use to calculate \( \hat{\beta} \).

The coefficients \( \hat{\beta} \) obtained from (1.7) are known as **OLS coefficients**, or **OLS statistic**, from Ordinary Least Squares.\(^{13}\) A very common idiom that economists use when referring to the calculation of OLS is “regressing \( y \) on \( X \)”. 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 \( \beta \), we pick the one that makes eq. (1.6) true, and therefore minimises the associated loss function \( e'e \). The vector \( \hat{y} = X\hat{\beta} \) is our approximation to \( y \). The term we normally use for the elements of \( \hat{y} \) are the **fitted values**: the closer they are to \( y \), the better we say that the model fits the data.

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

\[
X'X = \begin{bmatrix} n & n_m \\ n_m & n_m \end{bmatrix}
\]

\[
X'y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{x_i=1} y_i \end{bmatrix} = \begin{bmatrix} S_m + S_f \\ S_m \end{bmatrix}
\]

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 \((2 \times 2)\) matrices, which I will also assume known,\(^{14}\)

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

so that

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

and finally

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

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 \( \bar{Y}_f \).

Once again, opting for a quadratic loss function \((e'e)\) delivers a solution consistent with common sense. In this case, however, our approximate description of the vector \( y \) uses a function whose parameters are the statistics we are interested in.

---

\(^{13}\)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.

\(^{14}\)If you’re in trouble, go to subsection 1.A.3
1.3.2 The general case

In reading the previous subsection, the discerning reader will have noticed that, in fact, the assumption that \( x \) is a dummy variable plays a very marginal rôle. 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 \( \hat{\beta} \) will not contain the averages by sub-samples, but the fact that the loss function is minimised by \( \hat{\beta} = (X'X)^{-1}X'y \) keeps being true.

Example 1.1

Suppose that

\[
y = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 3 \\ 0 \\ 1 \end{bmatrix}, \quad x = \begin{bmatrix} 4 \\ 3 \\ 2 \\ 5 \\ 1 \\ 1 \end{bmatrix}
\]

The reader is invited to check that

\[
X'X = \begin{bmatrix} 6 & 16 \\ 16 & 56 \end{bmatrix} \Rightarrow (X'X)^{-1} = \begin{bmatrix} 0.7 & -0.2 \\ -0.2 & 3/40 \end{bmatrix}, \quad X'y = \begin{bmatrix} 10 \\ 33 \end{bmatrix}
\]

and therefore

\[
\hat{\beta} = \begin{bmatrix} 0.4 \\ 0.475 \end{bmatrix}, \quad \hat{y} = \begin{bmatrix} 2.3 \\ 1.825 \\ 1.345 \\ 2.775 \\ 0.875 \\ 0.875 \end{bmatrix}, \quad e = \begin{bmatrix} -1.3 \\ 1.175 \\ 0.65 \\ 0.225 \\ -0.875 \\ 0.125 \end{bmatrix}
\]

Hence, the function \( m(x_i) \) minimising the sum of squared residuals is \( m(x_i) = 0.4 + 0.475x_i \) and \( e'e \) 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

\[\text{Before you triumphantly shout "It's wrong!", remember to stick } \iota \text{ and } x \text{ together.}\]
1.3. **OLS AS A DESCRIPTIVE STATISTIC**

Figure 1.1: OLS on six data points

![Figure 1.1: OLS on six data points](image)

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 and white dots).

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 $x_i$, that is $m(x_i) = x_i' \beta$, 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 $x'_i$, 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 $e_i(\beta) = y_i - x'_i \beta$, then the vector of residuals is

$$e(\beta) = y - X \beta$$

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

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

$$X'e(\hat{\beta}) = 0$$

(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**:

$$X'X \cdot \hat{\beta} = X'y$$

and therefore, if $X'X$ is invertible, $\hat{\beta} = (X'X)^{-1}X'y$. 
If you think that all this is very clever, well, you’re right. The inventor of this is arguably the greatest mathematicians of all time: the great Carl Friedrich Gauss, also known as the princeps mathematicorum.\(^{16}\)

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

1. the scalar \(y'y\);
2. the \(k\)-element vector \(X'y\) and
3. the \(k \times k\) matrix \((X'X)^{-1}\),

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

\[
e'e = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - y'X\hat{\beta} - \hat{\beta}'X'y + \hat{\beta}'(X'X)\hat{\beta}
\]

and using (1.10) you have

\[
e'e = y'y - \hat{\beta}'(X'y).
\]

Equation (1.11) expresses the SSR as the difference between a scalar and the inner product of two \(k\)-element vectors \(\hat{\beta}\) and \((X'y)\). The number of rows of \(y\), 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 \(X\) 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 \(\beta\) 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

\[
\frac{\partial m(x)}{\partial x_j} = \beta_j
\]

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

\(^{16}\)To be fair, the French mathematician Adrien-Marie Legendre rediscovered it independently a few years later.
1.3. **OLS AS A DESCRIPTIVE STATISTIC**

Magnitude (if so, by how much?). However, you should resist the temptation to give the coefficients a counterfactual interpretation (If A. S. Student 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 3.5).

Focusing on marginal effects is what we do most often in econometrics, because the question of interest is not really approximating \( y \) given \( x \), but rather understanding what the effect of \( x \) 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 \( \beta \), rather than \( m(x) \). 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 \( x \) 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), \( X'X \) must be invertible. Now, you may ask: what if it's singular? This is an interesting case. The solution \( \hat{y} \) can still be found, but there is more than one vector \( \hat{\beta} \) associated with it. In fact, there are infinitely many. Let me give you an example. Suppose that \( X \) contain only one non-zero column, \( x_1 \). The solution is easy to find:

\[
\hat{\beta}_1 = \frac{x_1'y}{x_1'x_1},
\]

so that \( \hat{y} = \beta_1 x_1 \). Now, add to \( X \) a second column, \( x_2 \), which happens to be a multiple of \( x_1 \), so \( x_2 = \alpha x_1 \). Evidently, \( x_2 \) adds no information to our model, because it contains exactly the same information as \( x_1 \) so \( \hat{y} \) remains the same. Now, however, we can write it in infinitely many ways:

\[
\hat{y} = \beta_1 x_1 = 0.5\beta_1 x_1 + 0.5\frac{\beta_1}{\alpha} x_2 = 0.01\beta_1 x_1 + 0.99\frac{\beta_1}{\alpha} x_2 = \ldots
\]

because obviously \( \frac{\beta_1}{\alpha} x_2 = \beta_1 x_1 \). In other words, there are infinitely many ways to combine \( x_1 \) and \( x_2 \) to obtain \( \hat{y} \), even though the latter is unique and the objective function has a well-defined minimum.

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 \( X \) has full rank.
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, the problem should not arise. 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

\[
x = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}
\]

would be substituted by the matrix

\[
Z = \begin{bmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{bmatrix}
\]

so the first column of \( Z \) 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 \( Z \) unmodified leads to a collinearity problem if the model contains a constant, since the sum of the columns of the matrix \( Z \) is by construction \( \iota \). Hence, the matrix

\[
X = [\iota \ Z]
\]

has not full rank, so \( XX \) doesn’t have full rank either, and consequently is not invertible.\(^{17}\)

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

\(^{17}\)If you have problems following this argument, sections 1.A.3 and 1.A.4 may be of help.
1.3. OLS AS A DESCRIPTIVE STATISTIC

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.

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 \( \beta_2 \) indicates the difference between a northern observation and a southern one, in the same way as \( \beta_3 \) 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

\[
L(\beta) = \sum_{i=1}^{n} [y_i - m(x_i, \beta)]^2,
\]

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

\[
m(x_i) = \beta_1 + \beta_2 x_i + \beta_3 x_i^2.
\]

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 \( X \) matrix would be, in this case,

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

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 \( \beta_3 \) 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 \( \beta \) 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 \( \beta \) and \( x_i \); 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.

More generally, what we can treat via OLS is the class of models where

\[
m(x_i) = \sum_{j=1}^{k} \beta_j g_j(x_i),
\]

where \( x_i \) 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 \( X \) matrix. Clearly, once you have computed the \( \hat{\beta} \) vector, the marginal effects are easy to calculate:

\[
\frac{\partial m(x_i)}{\partial x_i} = \sum_{j=1}^{k} \hat{\beta}_j \frac{\partial g_j(x_i)}{\partial x_i}.
\]

### 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.
1.4. THE GEOMETRY OF OLS

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. \( d(a, b) = 0 \iff a = b \)
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 \( \|x\| = d(x, 0) \).

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

\[
d(x, y) = \sqrt{(x - y)'(x - y)}
\]

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

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.

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

\[
\hat{y} = \text{Argmin}_{x \in \text{Sp}(X)} \| y - x \|
\]

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

\[
\hat{\beta} = \text{Argmin}_{\beta \in \mathbb{R}^k} \| y - X\beta \|. \quad (1.14)
\]

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 \( \| y - X\beta \| \) is the same as the minimum of \( (y - X\beta)'(y - X\beta) \), and therefore

\[
\text{Argmin}_{\beta \in \mathbb{R}^k} \| y - X\beta \| = \hat{\beta} = (X'X)^{-1}X'y
\]

from which

\[
\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y.
\]

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

1.4.1 Projection matrices

In the previous subsection I pointed out that \( \hat{y} \) is a linear transform of \( y \). The matrix that operates the transform is said to be a projection matrix.\(^{18}\) 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 \( y \), the space spanned by \( X \) is the screen and the shadow of the fly is \( \hat{y} \).

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

\[
P_X = X(X'X)^{-1}X'
\]

Figure 1.2: Example: projection of a vector on another one

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

The base property of \( P_X \) is that, by construction, \( P_XX = X \), as you can easily check. Moreover, it’s symmetric and idempotent.

\[
P_X = P_X' \quad P_XP_X = P_X.
\]

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 matrix \( P_X \) takes a vector from wherever it is and moves it onto the closest point of \( \text{Sp}(X) \); but if the starting point already belongs to \( \text{Sp}(X) \), obviously no movement.

\(^{18}\)If I were pedantic, I’d have to say orthogonal projection, because you also get a similar tool called oblique projection. But we’ll never use it.
1.4. THE GEOMETRY OF OLS

takes place at all, so applying $P_X$ to a vector more than once produces no extra
effects ($P_X y = P_X P_X y = P_X P_X \cdots P_X y$).

It can also be proven that $P_X$ is singular; 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
$\text{Sp}(X)$: 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 $P_X$ being singular is that no matrix $A$
exists such that $A \cdot P_X = I$, and therefore no matrix exists such that $A\hat{y} = y$, which means
that $y$ is impossible to reconstruct from its projection.

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

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

$$M_X = I - P_X.$$  

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

$$M_X P_X = P_X M_X = [0],$$

where I’m using the notation $[0]$ for “a matrix full of zeros”.

Some more noteworthy properties: $M_X$ is symmetric, idempotent and sin-
gular, just like $P_X$.20 As for is rank, it can be proven that its rank equals $n - k$,
where $n$ is the number of rows of $X$ and $r = \text{rk}(X)$.

A fundamental property this matrix enjoys is that every vector of the type
$M_X y$ is orthogonal to $\text{Sp}(X)$, so it forms a 90° angle with any vector that can be
written as $X\lambda$.21 These properties are very convenient in many cases; a notable
one is the possibility of rewriting the SSR as a quadratic form:22

$$e' e = (M_X y)' (M_X y) = y'M_X M_X y = y'M_X y$$

where the second equality comes from symmetry and the third one from idem-
potency. By the way, the above expression could be further manipulated to re-

---

19To be specific: it can be proven that $\text{rk}(P_X) = \text{rk}(X)$, so $P_X$ is a $n \times n$ matrix with rank $k$; evi-
dently, 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.

20In fact, $M_X$ is itself a projection matrix, but let’s not get into this, ok?

21Let me remind the reader that two vectors are said to be \textbf{orthogonal} if their inner product is
0. In formulae: $x \perp y \iff x'y = 0$. A vector is orthogonal to a space if it’s orthogonal to all the points
that belong to that space: $y \perp \text{Sp}(X) \iff y'X = 0$, so $y \perp X\lambda$ for any $\lambda$.

22A \textbf{quadratic form} is an expression like $x'Ax$, where $x$ is a vector and $A$ is a square matrix,
usually symmetric. I sometimes use the metaphor of a sandwich and call $x$ the “bread” and $A$ the
“cheese”. 
obtain equation (1.11):

\[ y'M_Xy = y'(I - P_X)y = y'y - y'P_Xy = y'y - y'X(X'X)^{-1}X'y = y'y - \hat{\beta}'(X'y). \]

**Example 1.2**

Readers are invited to check (by hand or using a computer program of their choice) that, with the matrices used in example 1.1, \( P_X \) equals

\[
P_X = \begin{bmatrix}
0.3 & 0.2 & 0.1 & 0.4 & 0 & 0 \\
0.2 & 0.175 & 0.15 & 0.225 & 0.125 & 0.125 \\
0.1 & 0.15 & 0.2 & 0.05 & 0.25 & 0.25 \\
0.4 & 0.225 & 0.05 & 0.575 & -0.125 & -0.125 \\
0 & 0.125 & 0.25 & -0.125 & 0.375 & 0.375 \\
0 & 0.125 & 0.25 & -0.125 & 0.375 & 0.375 
\end{bmatrix}
\]

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 \( x_i \) become easy to represent in a compact and intuitive way:

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Symbol</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS Coefficients</td>
<td>( \hat{\beta} )</td>
<td>( (X'X)^{-1}X'y )</td>
</tr>
<tr>
<td>Fitted values</td>
<td>( \hat{y} )</td>
<td>( P_Xy )</td>
</tr>
<tr>
<td>Residuals</td>
<td>( e )</td>
<td>( M_Xy )</td>
</tr>
<tr>
<td>Sum of squared residuals</td>
<td>SSR</td>
<td>( e'e = y'M_Xy )</td>
</tr>
</tbody>
</table>

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

### 1.4.2 Measures of fit

We are now ready to tackle a very important issue. How good is our model? We know that \( \hat{\beta} \) is the best we can choose if we want to approximate \( y_i \) via \( \hat{y}_i = x'_i \hat{\beta} \), 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:

\[ 0 \leq \hat{y}'\hat{y} = y'P_Xy \leq y'y, \]
1.4. THE GEOMETRY OF OLS

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

$$0 \leq \frac{\hat{y}'\hat{y}}{y'y} = 1 - \frac{e'e}{y'y} = R^2_u \leq 1.$$ (1.16)

This index bears the name $R^2_u$ (“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 0, $y$ and $\hat{y}$ form a right triangle (see also Figure 1.2), in which you get a “good” leg, that is $\hat{y}$, and a “bad” one, the segment linking $\hat{y}$ and $y$, which is congruent to $e$: we'd like the bad leg to be as short as possible. After Pythagoras’ theorem, the $R^2_u$ 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 1.3

With the matrices used in example 1.1, you get that $y'y = 29$ and $e'e = 1.5$; therefore,

$$R^2_u = 1 - \frac{1.5}{29} \approx 0.9483$$

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

$$y'y = (y - 0)'(y - 0),$$

that is, the SSR for a model in which $\hat{y} = 0$. 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 $X = \iota$. In this case, the SSR is just the deviance of $y$, that is the sum of squared deviations from the mean, which can be written as $y'M_y$. If $\iota \in \text{Sp}(X)$ (typically, when the model contains a constant term), then a decomposition similar to (1.15) is possible: since $y = \hat{y} + e$, then obviously

$$y'M_y = \hat{y}'M_\iota \hat{y} + e'M_\iota e = \hat{y}'M_\iota \hat{y} + e'e$$ (1.17)

because if $\iota \in \text{Sp}(X)$, then $M_\iota e = e$. Therefore,

$$0 \leq e'e \leq \hat{y}'M_\iota y,$$

Subsection 1.A.8 should help the readers who want this result proven.
where the second inequality comes from the fact that $\hat{y}'M\hat{y}$ is a sum of squares and therefore non-negative. The modified version of $R^2$ is known as **centred R-square**:

$$R^2 = 1 - \frac{e'e}{y'M_y y} \tag{1.18}$$

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 $X$ and another (smaller) one, which only uses $\iota$, because $y'M_y y$ is just the SSR of a model in which we regress $y$ on $\iota$. 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:

Model A $\quad y = X\beta$

Model B $\quad y = X\beta + Z\gamma = W\theta$

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

The matrix $Z$ contains additional regressors to model A. It is important to realise that the information contained in $Z$ 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 distance between A. S. Tudent’s bed and St. Peter’s basilica in Rome.

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

$$SSR_A = e_a'e_a \quad SSR_B = e_b'e_b$$

where $e_a = M_X y$ and $e_b = M_W y$. Since $X \in \text{Sp}(W)$, clearly $P_W X = X$ and therefore

$$M_W M_X = M_W,$$

which implies $M_W e_a = e_b$; as a consequence,

$$SSR_B = e_b'e_b = e_a'M_W e_a = e_a'e_a - e_a'^{'}P_W e_a \leq e_a'e_a = SSR_A.$$  

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

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 Z is*. The $R^2$ index isn’t any better: proving that

$$SSR_B \leq SSR_A \Rightarrow R_B^2 \geq R_A^2.$$
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 $R^2$:

$$\overline{R^2} = 1 - \frac{e'e}{y'M_y y n - k},$$ (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.

### 1.5 The Frisch-Waugh theorem

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

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

Applying equation (1.7) we get the following:

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

It would seem that finding an analytical closed form for $\beta_1$ and $\beta_2$ as functions of $Z$, $W$ and $y$ is quite difficult; fortunately, it isn’t so: start from

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

and premultiply the equation above by $M_W$:

$$M_W y = M_W Z\hat{\beta}_1 + e,$$

since $M_W W = 0$ (by construction) and $M_W e = e$ (because $e = M_X y$, but $\text{Sp}(W) \subset \text{Sp}(X)$, so $M_W M_X = M_X$).\(^{25}\) Now premultiply by $Z'$:

$$Z'M_W y = Z'M_W Z\hat{\beta}_1$$

since $Z'e = 0$. As a consequence,

\(^{24}\)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.

\(^{25}\)If you’re getting a bit confused, you may want to take a look at section 1.A.8.
\[ \hat{\beta}_1 = (Z'MWZ)^{-1}Z'MWy \] (1.20)

Since \( MW \) is idempotent, an alternative way to write (1.20) could be
\[ \hat{\beta}_1 = \left[(Z'MW)(MWZ)\right]^{-1}(Z'MW)(MWy); \]
therefore \( \hat{\beta}_1 \) is the vector of the coefficients for a model in which the dependent variable is the vector of the residuals of \( y \) with respect to \( W \) and the regressor matrix is the matrix of residuals of \( Z \) with respect to \( W \). For symmetry reasons, you also obviously get a corresponding expression for \( \hat{\beta}_2 \):
\[ \hat{\beta}_2 = (W'MZW)^{-1}W'MWy \]

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

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

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, consider the case when \( X \) can be written as \( X = [Z \ u] \), where \( u \) is a column vector containing all zeros except for the \( m \)-th element, which is 1.

In scalar form, our model would read
\[ y_i = z_i' \beta + u_i \gamma + e_i \]
where is a dummy variable such that \( u_i = 1 \) if \( i = m \) and 0 otherwise. Thanks to the Frisch-Waugh theorem, it's easy to prove that the OLS coefficient for \( Z \) can be written as
\[ \hat{\beta} = \left[ \sum_{i \neq m} z_i z_i' \right]^{-1} \sum_{i \neq m} z_i y_i; \] (1.21)
that is, the coefficient you would get if you regressed \( y \) on \( Z \) alone on a sample that contains all observations except for the \( m \)-th one. The proof, if you're interested, is in section 1.A.9.

\[ ^{26}\text{Some call a vector like } u \text{ a basis vector. Others simply say "the } m \text{-th column of the identity matrix". We'll re-use this concept in section 3.3.1 and call } u \text{ an 'extraction vector'.} \]
1.6. AN EXAMPLE

An interpretation that this theorem can be given is the following: the coefficients for a group of regressors measure the response of $\hat{y}$ 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 $y$ contains data on the wages for $n$ employees, that $Z$ is their education level and $W$ is a geographical dummy variable (North vs South). The vector $\tilde{y} = M_W y$ will contain the differences between the individual wages and the average wage of the region where they live, in the same way as $\tilde{Z} = M_W Z$ contains the data on education as deviation from the regional mean. Therefore, regressing $\tilde{y}$ on $\tilde{Z}$ is a way to implicitly take into account that differences in wages between regions may depend on different educational levels. Consequently, by regressing $y$ on both the “education” variable and the regional dummy variable, the coefficient for education will measure its effect on wages controlling for geographical effects.

1.6 An example

For this example, I got some data from the 2016 SHIW dataset; our dataset contains 1917 individuals, who are full-time employees. 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 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.

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

Table 1.1: Wage example

---

27SHIW 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/](https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/).

28I can send you the details on the construction of the dataset from the raw data, if you’re interested. Just send me an email.
CHAPTER 1. OLS: ALGEBRAIC AND GEOMETRIC PROPERTIES

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

$$X'X = \begin{bmatrix}
1917 & 1153 & 22493 & 52527 \\
1153 & 1153 & 13299 & 32691 \\
22493 & 13299 & 288731 & 594479 \\
52527 & 32691 & 594479 & 1666703
\end{bmatrix}$$

$$X'y = \begin{bmatrix}
4253.3716 \\
2633.5507 \\
51038.9769 \\
116972.6710
\end{bmatrix}$$

$$y'y = 9690.62$$

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

$$\begin{bmatrix}
1356.9719 & -120.7648 & -63.9068 & -17.6027 \\
-120.7648 & 220.4901 & 1.2056 & -0.9488 \\
-63.9068 & 1.2056 & 4.4094 & 0.4177 \\
-17.6027 & -0.9488 & 0.4177 & 0.4844
\end{bmatrix} \times 10^{-5}$$

and therefore we have

$$\hat{\beta} = \begin{bmatrix}
1.3289 \\
0.1757 \\
0.0526 \\
0.0061
\end{bmatrix}$$

$$e'e = 177.9738 \quad R^2 = 29.76\%$$

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

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>1.3289</td>
<td>0.035530</td>
<td>37.40</td>
</tr>
<tr>
<td>male</td>
<td>0.175656</td>
<td>0.0143224</td>
<td>12.26</td>
</tr>
<tr>
<td>educ</td>
<td>0.0526218</td>
<td>0.00202539</td>
<td>25.98</td>
</tr>
<tr>
<td>wexp</td>
<td>0.00608615</td>
<td>0.000671503</td>
<td>9.066</td>
</tr>
</tbody>
</table>

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

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:

$$\hat{w}_i = 1.33 + 0.176g_i + 0.053e_i + 0.006a_i$$
1.A. ASSORTED RESULTS

where $w_i$ is the log wage for individual $i$, $g_i$ is their gender, and the rest follows. 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 just used the average wage, we get

$$0.298 = 1 - \frac{e' e}{y'y} \implies e' e = 0.702 \cdot y'y,$$

not bad at all.

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).

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.\footnote{One of the reasons why we economists love logarithms is that they auto-magically turn absolute changes into relative ones: $\beta_2 = \frac{d\ln(W)}{d\epsilon} = \frac{1}{W} \frac{dW}{d\epsilon} = \frac{\Delta W}{W} = \frac{\Delta W/\Delta \epsilon}{\epsilon}$. 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.} 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 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 a copy of Horn and Johnson (2012) or Lütkepohl (1996).
CHAPTER 1. OLS: ALGEBRAIC AND GEOMETRIC PROPERTIES

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 $y = f(x)$, where you have a real number $y$ for every possible vector $x$. For example, if $y = x + wz$, you can define the vector $x = [x, w, z]'$. 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 $x$. We adopt the convention by which the gradient is a **row** vector; hence, for the example above, the gradient is

$$
\frac{\partial y}{\partial x} = \begin{bmatrix}
\frac{\partial y}{\partial x} & \frac{\partial y}{\partial w} & \frac{\partial y}{\partial z}
\end{bmatrix} = \begin{bmatrix} 1 & wz & z \end{bmatrix}
$$

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

$$
f(x) = a'x = \sum_{i=1}^{n} a_i x_i;
$$
evidently, the partial derivative of $f(x)$ with respect to $x_i$ is just $a_i$; by stacking all the partial derivatives into a vector, the result is just the vector $a$, and therefore

$$
\frac{d}{dx} a'x = a'
$$

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

As for the quadratic form

$$
f(x) = x'Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j;
$$
it can be proven easily (but it’s rather boring) that

$$
\frac{d}{dx} x'Ax = x'(A + A')
$$
and of course if $A$ is symmetric (as in most cases), then $\frac{d}{dx} x'Ax = 2 \cdot x' A$. Again, note that the scalar case $\frac{d}{dx} ax^2 = 2ax$ 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 $y = Ax$ and $z = By$; of course, if you need the derivative of $z$ with respect to $x$ you may proceed by defining $C = B \cdot A$ and observing that

$$
z = B(Ax) = Cx \Rightarrow \frac{\partial z}{\partial x} = C
$$
but you may also get the same result via the chain rule, as

$$
\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x} = B \cdot A = C.
$$
1.A. ASSORTED RESULTS

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 \( x \in \mathbb{R}^n \) is a concise way of saying that \( x \) 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:

\[
    z = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k = \sum_{j=1}^k \lambda_j x_j;
\]

note that the above could have been written more compactly in matrix notation as \( z = X\lambda \), where \( X \) is a matrix whose columns are the vectors \( x_j \) and \( \lambda \) 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 \( x_1, \ldots, x_k \) 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 \( z \) and \( x_1, x_2, \ldots, x_k \).

Begin by considering the special case \( k = 1 \). Here \( z \) is just a multiple of \( x_1 \); 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 \( z \) you can obtain by all possible choices for \( \lambda_1 \), you get a straight line going through the origin, and of course \( x_1 \); this set of points is called the space spanned, or generated by \( x_1 \); or, in symbols, \( \text{Sp}(x_1) \). It’s important to note that this won’t work if \( x_1 = 0 \): in this case, \( \text{Sp}(x_1) \) 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, \( \text{Sp}(x_1, x_2) \) is a plane and \( z = \lambda_1 x_1 + \lambda_2 x_2 \) is a point somewhere on that plane. Its exact location depends on \( \lambda_1 \) and \( \lambda_2 \), but note that

- by a suitable choice of \( \lambda_1 \) and \( \lambda_2 \), no point on the plane is unreachable;
- no matter how you choose \( \lambda_1 \) and \( \lambda_2 \), you can’t end up outside the plane.

However, if \( x_2 \) is a multiple of \( x_1 \), then \( x_2 \in \text{Sp}(x_1) \) and \( \text{Sp}(x_1, x_2) = \text{Sp}(x_1) \), that is a line, and not a plane. In this case, considering \( x_2 \) won’t make \( \text{Sp}(x_1) \) “grow” in dimension, since \( x_2 \) 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 \( x_1, \ldots, x_k \) is said to be linearly independent if none of them can be expressed as a linear combination of the remaining ones.\(^{30}\) The

\[^{30}\text{The usual definition is that } x_1, \ldots, x_k \text{ are linearly independent if no linear combination } \sum_{j=1}^k \lambda_j x_j \text{ is zero unless all the } \lambda_j \text{ are zero. The reader is invited to check that the two definitions are equivalent.}\]
case I called “standard” a few lines above happens when $x_1$ and $x_2$ are linearly independent.

1.2.3 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 $X$), the maximum number of linearly independent columns of $X$ is the rank of $X$ ($\text{rk}(X)$ in formulae). The rank function enjoys several nice properties:

1. $0 \leq \text{rk}(X) \leq k$ (by definition)

2. $\text{rk}(X) = \text{rk}(X^\prime)$ (I’m not proving this for the sake of brevity)

3. $0 \leq \text{rk}(X) \leq \min(k, n)$ (by putting together the previous two); but if $\text{rk}(X) = \min(k, n)$, 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 $A^\prime = B$, then equality holds, and $\text{rk}(B^\prime B) = \text{rk}(BB^\prime) = \text{rk}(B)$

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

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 any non-zero $A$, there are infinitely many pairs of matrices that satisfy the decomposition above. For example, the matrix

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

can be written as

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

as well as

$$A = \begin{bmatrix} -10 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} -0.1 & 0 \end{bmatrix}$$
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.31

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: $\mathbf{y} = A\mathbf{x}$; 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 $\text{rk}(A) = n$ 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. 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.

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:

---

31You almost never need to compute a determinant by hand, so I’ll spare you its definition. If you’re curious, there’s always Wikipedia.
CHAPTER 1. OLS: ALGEBRAIC AND GEOMETRIC PROPERTIES

1. when $A$ is $2 \times 2$. 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

$$
A = \begin{bmatrix}
A_1 & 0 & \cdots & 0 \\
0 & A_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & A_m
\end{bmatrix};
$$

if the inverse exists, it has the same structure:

$$
A^{-1} = \begin{bmatrix}
A_1^{-1} & 0 & \cdots & 0 \\
0 & A_2^{-1} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & A_m^{-1}
\end{bmatrix}.
$$

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 next section), then its inverse is positive definite too;
- if $A$ is invertible, then the only solution to $Ax = 0$ is $x = 0$; conversely, if $A$ is singular, then there exists at least one non-zero vector such that $Ax = 0$.
- 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 $\beta$ is

$$
L(\beta) = e(\beta)'e(\beta);
$$

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

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

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 $\mathbf{e}(\beta) = y - X\beta$, we have

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

and the necessary condition for minimisation is $\frac{\partial L(\beta)}{\partial \beta} = -2\mathbf{X}'\mathbf{e} = 0$, 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.\(^{32}\) For example, suppose you have the following matrix $\mathbf{X}$:

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

For $\epsilon > 0$, the rank of $\mathbf{X}$ is, clearly, 2; nevertheless, if $\epsilon$ is a very small number, a computer program\(^{33}\) 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 $\epsilon$; Table 1.3 contains the results. Ideally, the right-hand side column in the table should only contain identity matrices. Instead, results are

\(^{32}\)If you find this kind of things intriguing I cannot but recommend chapter 1 in Epperson (2013); actually, the whole book!
\(^{33}\)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.
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 $X$ 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 $A$ is positive definite (pd for short) if the quadratic form $x'Ax$ returns a positive number for any choice of $x$ and positive semi-definite (psd for short) if $x'Ax \geq 0$.

If $A$ is positive (semi-)definite, then $-A$ is negative (semi-)definite. Of course, it is entirely possible that $x'Ax$ can take positive or negative values depending on $x$, in which case $A$ is said to be indefinite. If $A$ is semi-definite and invertible, then it’s also definite. Figure 1.3 may be helpful.

There are many interesting facts on psd matrices. A nice one is: if a matrix $H$ exists such that $A = HH'$, then $A$ is psd. This, for example, gives you a quick way to prove that that $I$ is pd and $PX$ is psd.

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

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

---

34In fact, figure 1.3 contains a slight inaccuracy. Finding it is left as an exercise to the reader.

35Easy to prove, too. Try.
We would like to say something about the space spanned by matrices defined as $W = X \cdot A$. There are two cases of interest. The first one arises when $A$ is square and invertible: in this case, $\text{Sp}(X) = \text{Sp}(W)$, so $P_X = P_W$. The result is easy to prove: for any $y \in \text{Sp}(X)$, there must be a vector $\beta$ such that $X \beta = y$. But then, by choosing $\gamma = A^{-1} \beta$, it’s easy to see that $y$ can also be written as $W \gamma$ and therefore $y \in \text{Sp}(W)$; by a similar reasoning, it can also be proven that if $y \in \text{Sp}(W)$ then $y$ also belongs to $\text{Sp}(X)$, and therefore $y \in \text{Sp}(X) \iff y \in \text{Sp}(W)$

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):

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

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 $W$ is also a linear combination of the columns of $X$, and therefore each column of $W$ is an element of $\text{Sp}(X)$. As a consequence, any vector that belongs to $\text{Sp}(W)$ also belongs to $\text{Sp}(X)$.

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

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>$P_W$</th>
<th>$M_W$</th>
<th>$P_X$</th>
<th>$M_X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_W$</td>
<td>$P_W$</td>
<td>$0$</td>
<td>$P_W$</td>
<td>$0$</td>
</tr>
<tr>
<td>$M_W$</td>
<td>$0$</td>
<td>$M_W$</td>
<td>$P_X - P_W$</td>
<td>$M_X$</td>
</tr>
<tr>
<td>$P_X$</td>
<td>$P_W$</td>
<td>$P_X - P_W$</td>
<td>$P_X$</td>
<td>$0$</td>
</tr>
<tr>
<td>$M_X$</td>
<td>$0$</td>
<td>$M_X$</td>
<td>$0$</td>
<td>$M_X$</td>
</tr>
</tbody>
</table>

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

Table 1.4: Projection matrices “multiplication table”

In this situation, the property $P_XW = P_XXA = XA = W$ implies some interesting consequences on the projection matrices for the spaces $\text{Sp}(W)$ and $\text{Sp}(X)$, 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.
1.A.9 Proof of equation (1.21)

Suppose our model is

\[ y_i = z_i’ \beta + u_i \gamma + e_i \]

where \( z_i \) is a \( k \)-variable vector and \( u_i \) is a dummy variable such that \( u_i = 1 \) if \( i = m \) and 0 otherwise. Therefore, the matrix of regressors is

\[ X = [Z \quad u]. \]

The Frisch-Waugh theorem (section 1.5) gives an analytical expression for \( \hat{\beta} \) as

\[ \hat{\beta} = (Z'M_uZ)^{-1}Z'M_uy. \]

Now, observe that \( M_u = I - P_u. \) In turn, \( P_u \) equals

\[ P_u = u(u'u)^{-1}u' = uu' = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \]

since \( u'u = 1 \) and \( uu' \) is a matrix full of zeros except for the \( m \)-th diagonal element, which is 1. Therefore,

\[ M_u = I - uu' = \begin{bmatrix} 1 & 1 & \cdots & 0 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix}, \]

that is an identity matrix with a "gap" at row \( m. \) Consequently, \( M_uZ \) equals

\[ M_uZ = \begin{bmatrix} 1 & z_1' \\ 1 & z_2' \\ \vdots & \vdots \\ 0 & z_m' \\ \vdots & \vdots \\ 1 & z_n' \end{bmatrix} = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_m' \\ \vdots \\ z_n' \end{bmatrix}. \]

Therefore,

\[ Z'M_uZ = \sum_{i=1}^{m-1} z_i'z_i' + 0 \cdot 0' + \sum_{i=m+1}^{n} z_i'z_i' = \sum_{i \neq m} z_i'z_i'. \]
the expression $Z'M_uy = \sum_{i \neq m} z_i y_i$ follows by a totally analogous argument. Therefore,

$$\hat{\beta} = \left[ \sum_{i \neq m} z_i z_i \right]^{-1} \sum_{i \neq m} z_i y_i.$$ 

I’ll leave it to the reader, as an exercise, to prove that

$$\hat{\gamma} = \frac{e_m}{1 - z_m (Z'Z)^{-1} z_m},$$

where $e_m$ is the $m$-th element of the vector of residuals that you get from the regression of $y$ on $Z$ on the full sample.
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.

\[That \text{ the sun will not rise tomorrow is no less intelligible a proposition, and implies no more contradiction, than the affirmation, that it will rise. We should in vain, therefore, attempt to demonstrate its falsehood. Were it demonstratively false, it would imply a contradiction, and could never be distinctly conceived by the mind.}\]

\[D. \text{ Hume, An Enquiry Concerning Human Understanding (1772).}\]

Inductive reasoning can be broadly formalised as follows:

1. Event X has always happened in the past.
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;\footnote{To be fair, this only applies to what Immanuel Kant called “synthetic” propositions. But maybe I’m boring you?} 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 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

\[ x = [1, 5, 4, 6, 3, 2, 6, 3, 4] \]

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 \( x \). 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 \( T(x) \) whose aim is not as much to describe the available data \( x \), but rather to describe the DGP that generated \( x \), 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.

### 2.2 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
2.2. A CRASH COURSE IN PROBABILITY

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): Gal-
lant (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 philo-
sophical 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 prob-
ability a proper branch of measure theory.

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 usu-
ally 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

\[ P(Ω) = 1, \quad P(\emptyset) = 0, \quad P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

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, intersec-
tion) 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
Ω to R; or, to put it differently, for any ω in Ω you get a corresponding real num-
ber X(ω). 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

\[ a < X(ω) \leq b \iff ω \in A, \]

3The 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.
4Warning: 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.
5In 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.
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, $X : \Omega \rightarrow S \subseteq \mathbb{R}$, and the set $S$ is sometimes indicated as $S(X)$. For a six-sided die, $S(X) = \{1, 2, 3, 4, 5, 6\}$; if $X$ is the time before our car breaks down, then $S(X) = [0, \infty)$, 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:

- $\lim_{a \rightarrow -\infty} F_X(a) = 0$;
- $\lim_{a \rightarrow \infty} F_X(a) = 1$;
- $F_X(a)$ 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 $\theta$.

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,

$$P(a < X \leq b) = \int_a^b f_X(z)dz;$$

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

---

$^6$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.
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 \( x \) with \( k \) elements you have that

\[
F_x(a) = P[(x_1 \leq a_1) \cap (x_2 \leq a_2) \cap \ldots \cap (x_k \leq a_k)]
\]

If all the \( k \) elements of \( x \) are continuous random variables, then you can define the joint density as

\[
f_x(x) = \frac{\partial^k F_x(x)}{\partial z_1 \partial z_2 \cdots \partial z_k}.
\]

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

\[
f_Y(a) = \int_{S(X)} \int_{S(Z)} f_w(x, a, z) \, dz \, dx
\]

### 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 are given the news that event \( B \) had occurred, then we can safely exclude the event \( B \) from \( \Omega \).\(^7\) In fact, after receiving the message “\( B \) is true”, the state space \( \Omega \) shrinks to \( B \), because states in which \( B \) is false are no longer conceivable as possible.

The consequence for \( A \) is that the subset \( A \cap B \) is no longer possible, so we must update our probability measure so that \( P(B) \) becomes 1, and consequently\(^8\) \( P(A) \) must be revised as

\[
P(A|B) = P(A \cap B) / P(B)
\]

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.”

---

\(^7\)When speaking about sets, I use the bar \( \bar{\cdot} \) to indicate the complement.

\(^8\)Technically, it’s more complicated than 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.
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.

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 \( F_Y(Y | X > 6) \) should be different from \( F_Y(Y | X < 3) \), 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

\[
f_{Y|X=x}(z) = \frac{f_{Y,X}(z,x)}{f_X(x)}
\]

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 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:

\[
E[X] = \sum_{x \in S(X)} x \cdot p(x) \quad \text{for discrete rvs (2.2)}
\]

\[
E[X] = \int_{S(X)} z \cdot f_X(z) \, dz \quad \text{for continuous rvs. (2.3)}
\]

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

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

\(^9\)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, \( \sigma \)-algebras and other tools that I’m not willing to use in this book.
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 $E[X]$ 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

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

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

$$E[1/X] = \int_0^1 \frac{1}{x} \cdot 1 \, dx = \left[ \log x \right]_0^1 = \infty.$$
The properties of $\text{Cov}[x]$ should be well known, but let's briefly mention the most important ones: if $\Sigma = \text{Cov}[x]$, then

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

Definition 2.5 makes it quite easy to calculate the covariance matrix of an affine transform:\(^{13}\)

$$\text{Cov}[Ax + b] = A \cdot \text{Cov}[x] \cdot A'.$$

(2.6)

Note that this result makes it quite easy to prove that if $X$ and $Y$ are independent rvs, then $V[X + Y] = V[X] + V[Y]$ (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

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

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 $E[y|x] = m(x)$ as a function of $x$. This function is sometimes called the regression function of $Y$ on $X$.

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

**Example 2.2**

Assume that you have a bivariate variable $(Y, X)$ where $Y$ is 1 if an individual is a fan of “Breaking Bad” and 0 otherwise, and $X$ is 1 if the same individual is a meth user (and 0 otherwise). Suppose that the joint probability is

<table>
<thead>
<tr>
<th></th>
<th>$X = 0$</th>
<th>$X = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 0$</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>0.6</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The probability of being a Breaking Bad fan among meth users is $\frac{0.1}{0.1+0.1} = 50\%$, while for non meth users it’s $\frac{0.6}{0.2+0.8} = 75\%$. The same statement could have been stated in formulae as

$$E[Y|X] = 0.75 - 0.25X,$$

which gives 0.5 if $X = 1$ and 0.75 if $X = 0$. We may say that the regression function of $Y$ on $X$ is linear ($E[Y|X] = \theta_0 + \theta_1 X$), and that it depends on the vector $\theta = [\theta_0, \theta_1]$.\(^{12,13}\)

\(^{12}\)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.
2.3. ESTIMATION

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

\[
E[E[y|x]] = E[y].
\]

(2.7)

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:

\[
f(y) = E[f(y|x)]
\]

To continue with example 2.2, note that \( E[Y] = E[0.75 - 0.25 \cdot X] = 0.75 - 0.25 \cdot E[X] = 0.7 \).

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 that if you start doing meth you grow a dislike for that certain TV series. However, this does not mean that it’s not worth examining. 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 \( x \), and that the DGP which generated \( x \) can be described by means of a vector of parameters \( \theta \). An estimator is a statistic \( \hat{\theta} = T(x) \) that should be “likely” to yield a value “close to” the parameters of interest \( \theta \).

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

In very simple cases, we may be able to compute \( E[\hat{\theta}] \) and perhaps even \( V[\hat{\theta}] \), which leads us to the well known concepts of unbiasedness and efficiency:

- the bias of \( \hat{\theta} \) is the difference \( E[\hat{\theta}] - \theta \); therefore \( \hat{\theta} \) is said to be unbiased if \( E[\hat{\theta}] = \theta \);
- \( \hat{\theta} \) is more efficient than \( \hat{\theta} \) if \( V[\hat{\theta}] - V[\hat{\theta}] > 0 \) (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.

\(^{14}\)Notably, when \( \hat{\theta} \) is an affine function of \( x \).
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 probability:

$$X_n \xrightarrow{p} X \iff \lim_{n \to \infty} P[|X_n - X| < \epsilon] = 1 \quad (2.8)$$

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

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

the sequence above can be read as a sequence of events, where $|X_i - X| < \epsilon$ means "$X_i$ is more or less $X$". 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.

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 $X_n \xrightarrow{p} a$, 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,

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

does the sequence $\bar{X}_n$ have a limit in probability? The conditions for this are studied by the so-called Laws of Large Numbers, or LLNs for short.

There are many LLNs, and they 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 Khinchin's, which sets very strong bounds on the first two conditions and 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 $E[x_i] =$

---

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.
2.3. ESTIMATION

$m$, then $\bar{X} \xrightarrow{P} m$. 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.

**Example 2.3**

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:

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

Not that the probability $\pi$ 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: $E[x_i] = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi$. Together with the iid property, this is enough for invoking the LLN and establishing that $\bar{X} \xrightarrow{P} \pi$. Therefore, we can take the empirical frequency $\hat{p}$ as a consistent estimator of the true probability $\pi$.

Another common case when it’s easy to prove convergence in probability occurs when each member of the sequence $x_i$ possesses finite second moments and

$$
\lim_{n \to \infty} E[x_i] = m \\
\lim_{n \to \infty} V[x_n] = 0;
$$

if $m$ is finite, there is a result called **Chebyshev’s inequality**, using which you can easily prove that $\bar{X} \xrightarrow{P} m$. This case is nearly useless in practice 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 you can use to prove consistency.

The LLN becomes enormously powerful when coupled with another wonderful result, known as the **Slutsky’s Theorem**: if $X_n \xrightarrow{P} a$ and $g(\cdot)$ is continuous at $a$, then $g(X_n) \xrightarrow{P} g(a)$ (note how much easier this property makes it to work with probability limits rather than expectations).

We will want our estimators to be consistent:

$$
\hat{\theta} \xrightarrow{P} \theta \iff \lim_{n \to \infty} P[|\hat{\theta} - \theta| < \epsilon] = 1; \quad (2.9)
$$

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 construct 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 $E[x_i] = 1/\theta$, then $\hat{\theta} = 1/\bar{X}$; if $E[x_i] = e^{\theta}$, then $\hat{\theta} = \log(\bar{X})$; if $E[x_i] = \theta^2$, then $\hat{\theta} = \sqrt{\bar{X}}$; and so on. The extension to the case when $\theta$ is a vector is technically messier, but conceptually identical.

**Example 2.4**

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

$$E[X] = \frac{p}{\alpha}$$
$$E[X^2] = \frac{p(p+1)}{\alpha^2};$$

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

$$m_1 \xrightarrow{p} \frac{p}{\alpha}$$
$$m_2 \xrightarrow{p} \frac{p(p+1)}{\alpha^2}.$$

Now consider the statistic $\hat{p} = m_2^2 / (m_2 - m_1^2)$. Clearly $\hat{p}$ is a continuous function of both $m_1$ and $m_2$; therefore,

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

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

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

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

### 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 from the same quantity: that is, two different statistics $\hat{\theta}$ and $\bar{\theta}$ that have the same probability limit $\bar{\theta}$.

How do we choose which one to use? Consistency can’t be a criterion, since they both are: if we define

$$\bar{\theta}_n = P[|\hat{\theta} - \theta| < \epsilon]$$
$$\bar{\theta}_n = P[|\bar{\theta} - \theta| < \epsilon]$$
clearly \( \lim_{n \to \infty} \hat{P}_n = \lim_{n \to \infty} \tilde{P}_n = 1 \), so a decision can’t be made on these grounds. Nevertheless, if we could establish that, for \( n \) large enough, \( \hat{P}_n > \tilde{P}_n \), 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: \( \hat{P}_n \) is defined as

\[
\hat{P}_n = \int_{\theta - \epsilon}^{\theta + \epsilon} \hat{f}(x) \, dx,
\]

where \( \hat{f}(x) \) is the density function for \( \hat{\theta} \) (clearly, a parallel definition holds for \( \tilde{P}_n \)). In most cases, the analytical form of \( \hat{f}(x) \) is very hard to establish, if not at all impossible. However, it may be possible to find approximations to the actual densities that are 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 \( f(x) \) 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? 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 and 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:

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

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

\[
X_n \overset{d}{\to} X \iff F_{X_n}(z) \to F_X(z) \quad (2.10)
\]

\(^{18}\) I assume that the reader is reasonably comfortable with the Gaussian distribution, but section 2.A.2 is there, just in case.

\(^{19}\) 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.*
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 $X$. It can be proven that $X_n \xrightarrow{p} X$ implies $X_n \xrightarrow{d} X$, but not vice versa.

Note the fundamental difference between convergence in probability and in distribution: if $X_n \xrightarrow{p} X$, 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 $X_n \xrightarrow{d} X$ instead, there is no guarantee that the two numbers will be close: the only thing we know is that they come from (nearly) the same probability distribution.

Now imagine that the LLN holds and $\bar{X} \xrightarrow{p} m$ Clearly, $\bar{X} - m \xrightarrow{p} 0$. 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 analyses the conditions under which

$$\sqrt{n}(\bar{X} - m) \xrightarrow{d} N(0, \nu), \quad (2.11)$$

where $N(0, \nu)$ is a Gaussian random variable. A multivariate version also exists, which is slightly more intricate from a technical point of view, but the intuition carries over straightforwardly.

The approximation provided by the CLT can also be stated by using the symbol $\sim$, which means “approximately distributed as” (where the approximation gets better and better as $n$ grows):

$$\sqrt{n}(x - m) \xrightarrow{d} N(0, \Sigma) \implies x \sim N\left( \frac{m}{n}, \frac{1}{n} \Sigma \right)$$

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, $E[x_i] = m$, and $V[x_i] = \nu$, then equation (2.11) 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 2.5**

Let’s go back to example 2.3 (the coin-tossing experiment). Here not only the mean exists, but also the variance:

$$V[x_i] = E[x_i^2] - E[x_i]^2 = \pi - \pi^2 = \pi(1 - \pi)$$

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

$$\sqrt{n}(\hat{p} - \pi) \xrightarrow{d} N(0, \pi(1 - \pi)),$$
so the corresponding asymptotic approximation is
\[ \hat{p} \overset{a}{\sim} N\left( \pi, \frac{\pi(1-\pi)}{n} \right). \]

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 you’re interested in reproducing these numbers, Section 2.A.3 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 \( X_n \overset{p}{\rightarrow} a \) and \( Y_n \overset{d}{\rightarrow} Y \), then
\[ X_n \cdot Y_n \overset{d}{\rightarrow} a \cdot Y. \tag{2.12} \]

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 \( \hat{\theta} \);
\[
\left\{ \begin{array}{l}
\bar{X} \overset{p}{\rightarrow} m \\
\sqrt{n}(X - m) \overset{d}{\rightarrow} N(0, \Sigma)
\end{array} \right\} \implies \left\{ \begin{array}{l}
\hat{\theta} = g(\bar{X}) \overset{p}{\rightarrow} \theta = g(m) \\
\sqrt{n}(\hat{\theta} - \theta) \overset{d}{\rightarrow} N(0, J\Sigma J')
\end{array} \right\} \tag{2.13}
\]
where \( n \) is the sample size and \( J \) is the Jacobian \( \frac{\partial g(x)}{\partial x} \bigg|_{x=m} \).

**Example 2.6**
Given a sample of iid random variables \( x_i \) for which \( E[x_i] = 1/a \) and \( V[x_i] = 1/a^2 \), it is straightforward to construct a consistent estimator of the parameter \( a \) as
\[
\hat{a} = \frac{1}{\bar{X}} \overset{p}{\rightarrow} \frac{1}{1/a} = a.
\]
Its asymptotic distribution is easy to find: start from the CLT:
\[ \sqrt{n}(\bar{X} - 1/a) \overset{d}{\rightarrow} N(0, 1/a^2). \]

All we need is the Jacobian term, which is
\[
j = \operatorname{plim} \frac{d\hat{a}}{d\bar{X}} = -\operatorname{plim} \frac{1}{\bar{X}^2} = -\frac{1}{1/a^2} = -a^2; \]
therefore, the variance of the asymptotic distribution of $\hat{a}$ is given by

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

and therefore

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

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

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 two 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 $AV[b] - AV[a]$ is positive semi-definite (see page 34). This criterion is easy to understand when $a$ and $b$ are scalars: $AV[b] \geq AV[a]$. 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 $AV[g(b)] \geq AV[g(a)]$ in all cases.

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.

In addition, 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

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \omega^2);$$

this means that, for a decently large value of $n$, we can approximate the distribution of $\hat{\theta}$ as $\hat{\theta} \xrightarrow{a} N\left(\theta, \frac{\omega^2}{n}\right)$. This, in turn, implies that

$$P\left[\frac{|\hat{\theta} - \theta|}{\sqrt{\omega^2/n}} > 1.96\right] \approx 5\%;$$

therefore, the chances that the interval

$$\hat{\theta} \pm 1.96 \times \sqrt{\omega^2/n}$$

\footnote{It is common to call $AV[\hat{\theta}]$ the “asymptotic variance” of $\hat{\theta}$; this is technically incorrect, since $\hat{\theta}$ may not have a variance for any $n$, although its limit distribution does. But let’s not be pedantic.}
contains the true value of $\theta$ 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 $\theta$ is a 2-parameter vector, the confidence set would be an ellipse.

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; 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. 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. The probability of the coin showing the “heads” side when we flip it (say, $\pi$) is a number between 0 and 1. Our conjecture is $H_0: \pi = 0.5$. After flipping the coin $n$ times, we obtain a vector $x$ of zeros and ones. What we want is a function $T(x)$ (called a test statistic) such that we can decide whether to reject $H_0$ or not.

![Figure 2.1: Types of Error](image)

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

---

21 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.
BYfixing 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 \( H_0 \). 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 \textbf{type I} and \textbf{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:

\[
\begin{array}{ccc}
| & \text{Not Reject} & \text{Reject} \\
\hline
H_0 \text{ true} & \text{ok} & \text{Type I error} \\
H_0 \text{ false} & \text{Type II error} & \text{ok} \\
\end{array}
\]

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 \textit{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.

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 \textit{falsified}: we don’t use evidence to \textit{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, then 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.

\[\text{\textsuperscript{22}}\text{Unless, of course, we do something silly such as deciding to always reject, or never. But in that case, what’s the point of performing the experiment at all?}\]
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”.

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 $\theta$ 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 $H_0$, $g(\hat{\theta}) \xrightarrow{p} 0$.

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:

$$g(\hat{\theta}) \xrightarrow{a} N(g(\theta), \Sigma);$$

under consistency, $\Sigma$ should tend to a zero matrix; as for $g(\theta)$, that should be 0 if and only if $H_0$ is true. These two statement 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 should diverge to infinity, since $\text{plim}[g(\hat{\theta})] \neq 0$; if $H_0$ is true, instead, approximate normality implies that:

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

where $p = \text{rk}(\Sigma)$. Hence, under $H_0$, the $W$ statistic should take values typical of a $\chi^2$ random variable. 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, we would be committing 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

---

23Some people use the word “retain” instead of “accept”, which is certainly more correct, but unfortunately not very common.

24To see why, see section 2.A.2

25The 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.
probability of a type I error (called the size of the test, and usually denoted by the Greek letter $\alpha$) is a small number, typically 5%.

What people do in most cases is deciding which $\alpha$ they want to use and then set $c$ accordingly, so that in many cases you see $c$ expressed as a function if $\alpha$ (and written $c_\alpha$), 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. There are many interesting things that could and should be said about the power of tests, especially a truly marvelous result known as the Neyman-Pearson lemma, but I’m afraid this is not the place for this. See the literature cited at footnote 21.

Example 2.7

Let’s continue example 2.5 here. So we have that the relative frequency is a CAN estimator for the probability of a coin showing “heads”:

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

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

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

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

$$g(\hat{p}) \xrightarrow{d} N(0, n^{-1})$$

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 $W = 100 \cdot 0.08^2 = 6.4$. 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 $n = 1000$.  

\footnote{When the power of a test goes to 1 asymptotically, the test is said to be consistent. I know, it’s confusing.}
2.4. HYPOTHESIS TESTING

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 $\alpha$ is complicated: you compile a table of critical values for a given $\alpha$ 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 $\alpha$, as well as performing the inverse calculation, since algorithms for computing the cdf of a $\chi^2$ random variable are fast, precise and efficient. Thus, an equivalent route is often followed: after computing $W$, you calculate the probability that a $\chi^2$ variable should take values greater than $W$. If you prefer, the lowest significance level you would need to reject. That number is called the $p$-value for the test statistic $W$. Clearly, if $W > c$, the $p$-value must be smaller than $\alpha$, so the decision rule can be more readily stated as “reject $H_0$ if the $p$-value is smaller than $\alpha$.”

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;\(^\text{27}\) in Fisher’s own words, when we see a small $p$-value, “either an exceptionally rare chance has occurred or the theory […] is not true.”\(^\text{28}\) Figure 2.2 shows an example in which $W = 9$, and is compared against a $\chi^2_3$ 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.

To make results even quicker to read, most statistical packages adopt a graphical convention, based on ornatning the test statistic with a variable number of ‘*’ characters, usually called “stars”. Their meaning is as follows:

\(^{27}\)Thanks to Sven Schreiber for putting it so clearly and concisely.

**Table 2.2:** P-value example

<table>
<thead>
<tr>
<th>Stars</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>(none)</td>
<td>p-value greater than 10%</td>
</tr>
<tr>
<td>*</td>
<td>p-value between 5% and 10%</td>
</tr>
<tr>
<td>**</td>
<td>p-value between 1% and 5%</td>
</tr>
<tr>
<td>***</td>
<td>p-value under 1%</td>
</tr>
</tbody>
</table>

Therefore, when you see 3 stars you “strongly” reject the null, but you don’t reject \( H_0 \) 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 \( \chi^2 \) distribution is just an approximation to the actual density of the test. Therefore, the quantiles of the \( \chi^2 \) 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 \( \chi^2 \) 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 \( \theta \);
- 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 \( \theta \).

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 \( \theta \) 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** \( \psi \). That is, the parameters \( \theta \) characterise the DGP, while \( \psi \) is a formalised description of the aspect of reality we are trying to analyse. For example, in a typical econometric model, \( \psi \) may contain quantities such as the elasticity of demand for a certain good to its own price, or to the price of a substitute good, and so on.

What is the relationship between \( \psi \) and \( \theta \)? If we take \( \psi \) as being a stylised description of reality, and \( \theta \) as a stylised description of what we observe, then \( \theta \) should be a known function of \( \psi \), 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 \( \theta \); is this enough to estimate \( \psi \)? It depends on the function \( M(\cdot) \); if the function is invertible, and we have a CAN estimator \( \hat{\theta} \), a possible estimator for \( \psi \) 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 \( \psi \) give rise to the same \( \theta \).\(^{29}\) 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 \( \theta \)) in both cases; this situation is known as observational equivalence, and \( \psi_1 \) and \( \psi_2 \) are said to be observationally equivalent. In these case, being able to estimate \( \theta \), even in an arbitrarily precise way, doesn’t tell us if the “true” description of the world is \( \psi_1 \) or \( \psi_2 \). This unfortunate case is known in econometrics as under-identification.

**Example 2.8**

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 \( r \).

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

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

In this case, the probability distribution of our data is completely characterised by the parameter \( \pi \); as we know, we have a perfectly good way to estimate \( \pi \), since \( \hat{\pi} \) is a consistent estimator and testing hypotheses on \( \pi \) 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

\[
\pi = \frac{r}{w + r}
\]

\(^{29}\) The technical way to say this would be “the \( M(\cdot) \) function is not injective".
is one equation in two unknowns. Therefore, in the absence of extra information we are able to estimate \( \pi \) (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 \( \pi \), 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.

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 \( \mathrm{E}[g(X)] \neq g(\mathrm{E}[X]) \). However, when \( g(\cdot) \) is concave, we have a usable result that comes as an inequality:

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

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

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

\[
g(x) \leq g(x^*) + g'(x^*)(x-x^*) \tag{2.14}
\]

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, \( a < \mu = \mathrm{E}[X] < b \); this implies that equation (2.14) holds when \( x^* = \mu \), and therefore

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

Since obviously \( \mathrm{E}[X-\mu] = 0 \), it follows that

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

as required. Note that
• by linearity, $E[-X] = -E[X]$, so if the function is convex instead of concave, you can flip the inequality, because the negative of a concave function is convex: $E[g(X)] \geq g(E[X])$;

• 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 The normal and $\chi^2$ 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.

![Standard normal density function](image)

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\(^{30}\) by C. F. Gauss (the guy on page 12), so it’s also also known as a Gaussian random variable. By

\(^{30}\)Or invented? Interesting point.
playing with integrals a little\textsuperscript{31}, it can be proven that \( E[X] = 0 \) and \( V[X] = 1 \). 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

\[
f(y) = \frac{1}{\sqrt{2\pi s^2}} \exp\left\{ -\frac{(y - m)^2}{2s^2} \right\}
\]

A compact way to say this is \( Y \sim N(m, s^2) \).

In fact, one can define a multivariate normal random variable as a random vector \( x \) with density

\[
f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{ (x - m)'\Sigma^{-1}(x - m) \right\},
\]

or, in short, \( x \sim N(m, \Sigma) \), where \( n \) is the dimension of \( x \), \( m \) is its expectation and \( \Sigma \) its covariance matrix. The multivariate version of this random variable also enjoys the linearity property, so if \( x \sim N(m, \Sigma) \), then

\[
y = Ax + b \sim N(Am + b, A\Sigma A').
\]

(2.15)

It is easy to overlook how amazing this result is: the fact that \( E[Ax + b] = AE[x] + b \) 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 \textit{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 \( x = [Y, X] \) 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 \( y \) and \( x \) are jointly Gaussian, then the conditional density \( f(y|x) \) is Gaussian as well. In formulae:

\[
f(y|x) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{ (y - m)'\Sigma^{-1}(x - m) \right\},
\]

where \( m = E[y|x] = E[y] + B'(x - E[x]) \)

\[
B = \Sigma_x^{-1}\Sigma_{xy}
\]

\[
\Sigma = V(y|x) = \Sigma_y - \Sigma_{xy}\Sigma_x^{-1}\Sigma_{xy}
\]

where \( \Sigma_y \) is the covariance matrix of \( y \), \( \Sigma_x \) is the covariance matrix of \( x \) and \( \Sigma_{xy} \) is the matrix of covariances between \( x \) and \( y \).

\textsuperscript{31}If you want to have some fun with the moments of the standard normal distribution, you’ll find the result \( \frac{d\phi(x)}{dx} = -x\phi(x) \) very useful, because it implies that \( \int x\phi(x)dx = -\phi(x) \).
Example 2.9

For example, suppose that the joint distribution of \( y \) and \( x = [x_1, x_2] \) is normal, with

\[
E \begin{bmatrix} y \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}
\]
\[
V \begin{bmatrix} y \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}
\]

then you have

\[
E[y] = 1 \quad E[x] = (2,3)'
\]
\[
\Sigma_y = 3 \quad \Sigma_x = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \Sigma_{xy} = (0,1)'
\]

and therefore

\[
B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}
\]
\[
\Sigma = 3 - \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 3 - 1 = 2,
\]

since \( \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \). As a consequence, the conditional expectation of \( y \) given \( x \) equals

\[
E[y|x] = 1 + [-1,1]' \begin{bmatrix} x_1 - 2 \\ x_3 - 3 \end{bmatrix} = -x_1 + x_2
\]

and in conclusion

\[
y|x \sim N[x_2 - x_1, 2].
\]

Note that:

- the conditional mean is a linear function of the \( x \); this needn’t happen in general: it’s a miraculous property of Gaussian random variables;

- the conditional variance is not a function of the \( x \) 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. (2.7)) you get

\[
E[E[y|x]] = -E[x_1] + E[x_2] = -2 + 3 = 1 = E[y];
\]

quelle surprise!
If one instead needs to investigate the distribution of quadratic forms of Gaussian rvs, then another distribution arises, namely the \textit{chi-square} distribution ($\chi^2$ in symbols). The general result is that, if $\mathbf{x} \sim N(0, \Sigma)$, 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 $\chi^2$ density is over the non-negative reals; its shape depends on $n$ (the degrees of freedom) in the following way:

$$f(x) = \frac{0.5^{n/2}}{\Gamma(n/2)} x^{n/2-1} e^{-x/2}.$$ 

The most common cases, where $n$ ranges from 1 to 4, are shown in Figure 2.4; the 95% critical values for these cases are

<table>
<thead>
<tr>
<th>degrees of freedom</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>critical value at 95%</td>
<td>3.84</td>
<td>5.99</td>
<td>7.81</td>
<td>9.49</td>
</tr>
</tbody>
</table>

For example, a $\chi^2$ random variable takes values from 0 to 3.84 with probability 95%. Memorising them may turn out to be handy from time to time.

\footnote{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.}
2.A.3 Gretl script to reproduce example 2.5

Input:

```
set verbose off
clear

# characteristics of the event
scalar p = 0.5
scalar n = 100
scalar lo = 36
scalar hi = 45

# 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
scalar appr = cnorm(z1) - cnorm(z0) # "cnorm" = Normal distribution function

# printout
printf "probability of "heads" = %g\n", p
printf "number of tosses = %g\n", n
printf "probability of heads between %d and %d: \ntrue = %g, approximate via CLT = %g\n", lo, hi, 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 will 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 $x$, 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 $x$, so further insight can be gained by studying the conditional distribution $f(y|x)$.

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.

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 $f(y|x)$. 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 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 $x = 65.5$, 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 $\hat{c}_i = 23.9 + 0.646p_i$, 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 $f(y|x)$, we could focus on the conditional expectation $E[y|x]$: 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 have an upward-sloping line, that suits very well our belief that taller parents should have, as a rule, taller children.

The next ingredient for making this intuition operational is to assume that

---

1 An “outlier” is a data point that is far away from the rest.
3.1. THE REGRESSION FUNCTION

$E[y|x]$ is a simple function, whose shape is governed by a few parameters.\(^2\) The choice that is nearly universally made is that of a linear\(^3\) function: $E[y|x] = x'\beta$. Now define the random variable $\epsilon = y - x'\beta$, so that $y$ can be written (by definition) as $x'\beta + \epsilon$. If we observe multiple realisations of $y$ and $x$, then we can write

$$y_i = x'_i\beta + \epsilon_i$$

or, in matrix form,

$$y = X\beta + \epsilon$$

(3.2)

For historical reasons, the random variable $\epsilon$ is called the disturbance; the idea here is that if $\epsilon_i$ was 0, by observing $y_i$ we would observe $E[y_i|x_i]$ in its “uncontaminated” form.

Like many other terms in use in econometrics, the word comes from the statistical literature that predated econometrics proper. In other disciplines, mainly experimental ones, it is quite natural to think of the regression function as a “Law of Nature” and of $\epsilon$ as some random fluctuation in our data, coming from experimental errors, bad measurement, and so on. These considerations make it quite natural, in experimental science, to surmise that the distribution of $\epsilon$ should be roughly normal, as a consequence of the CLT. Unfortunately, this position is untenable in economics.

A very important property of the random variable $\epsilon$ is that it’s orthogonal to $x$ by construction:\(^4\)

$$E[x \cdot \epsilon] = \mathbf{0}$$

(3.3)

The proof is simple: since $\epsilon = y - x'\beta$, clearly

$$E[\epsilon|x] = E[y|x] - E[x'\beta|x] = x'\beta - x'\beta = 0;$$

therefore, by the law of iterated expectations (see Section 2.2.4),

$$E[x \cdot \epsilon] = E[x \cdot E[\epsilon|x]] = E[x \cdot 0] = \mathbf{0}.$$

Note the difference between equation (3.2) and the parallel OLS decomposition $y = X\hat{\beta} + e$, where everything on the right-hand side of the equation is an observable statistic. Instead, the only observable item in the right-hand side of (3.2) is $X$: $\beta$ is an unobservable vector of parameters and $\epsilon$ is an unobservable vector of random variables. Still, the similarity is striking; it should be no surprise that, under appropriate conditions, $\hat{\beta}$ is a CAN estimator of $\beta$, which we prove in the next section.

\(^2\)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.

\(^3\)As for what we mean exactly by “linear”, see 1.3.2.

\(^4\)Warning: as shown in the text, $E[\epsilon|x] = 0 \implies E[x \cdot \epsilon] = \mathbf{0}$, but the converse is not necessarily true.
3.2 Main statistical properties of OLS

A handy consequence of equation (3.2) is that the OLS statistic can be written as
\[ \hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon \]  
(3.4)

As any estimator, \( \hat{\beta} \) has a distribution, and its finite-sample properties can be studied, in some cases. For example, its unbiasedness is very easy to prove: if \( E[\varepsilon|X] = 0 \), then

\[ E[\hat{\beta}|X] = \beta + E[(X'X)^{-1}X'\varepsilon|X] = \beta + (X'X)^{-1}X'E[\varepsilon|X] = \beta. \]

And therefore, by the law of iterated expectations,

\[ E[\hat{\beta}] = E[E[\hat{\beta}|X]] = E[\beta] = \beta. \]

However, nobody cares about unbiasedness nowadays. Moreover, in order to say something on the distribution of \( \hat{\beta} \) we'd need assumptions on the distribution of \( \varepsilon \), 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:

\[ \hat{\beta} = \beta + \left[ \sum_i x_ix_i' \right]^{-1} \sum_i x_i\varepsilon_i = \beta + \left[ \frac{1}{n} \sum_i x_ix_i' \right]^{-1} \frac{1}{n} \sum_i x_i\varepsilon_i. \]  
(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

\[ z_i = x_i\varepsilon_i; \]  
(3.6)

given equation (3.3), \( E[z] = 0 \), so a straightforward application of the LLN gives

\[ \frac{1}{n} \sum_i x_i\varepsilon_i \xrightarrow{p} 0. \]  
(3.7)

As for the limit of the first term, assume that \( n^{-1}X'X \) has one, and call it \( Q \):

\[ \frac{1}{n} \sum_i x_ix_i' \xrightarrow{p} Q; \]  
(3.8)

Ordinarily, \( Q \) will be equal to \( E(x_ix_i') \).
3.2. MAIN STATISTICAL PROPERTIES OF OLS

if \( Q \) is invertible, then we can exploit the fact that inversion is a continuous transformation as follows

\[
\left[ \frac{1}{n} \sum_{t} x_t x_t' \right]^{-1} \xrightarrow{P} Q^{-1},
\]

so, after putting the two pieces together,

\[
\hat{\beta} \xrightarrow{P} \beta + Q^{-1} \cdot 0 = \beta.
\]

The OLS statistic, therefore, is a consistent estimator of the parameters of the conditional mean, or to be more technical, of the derivative of \( E[y|x] \) with respect to \( x \), 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 \( E[y|x] \) is non linear, it may be proven that \( \hat{\beta} \) 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 \( \frac{1}{n} \sum_{t} x_t x_t' \) converges to an invertible matrix; this may not happen for two reasons:

1. it may not converge to any limit; this would be the case if, for example, the vector \( x \) 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),

\[
\sqrt{n}(\hat{\beta} - \beta) = \left[ \frac{1}{n} \sum_{t} x_t x_t' \right]^{-1} \frac{1}{\sqrt{n}} \sum_{t} x_t \epsilon_t
\]

we already know from the previous subsection that \( \left[ \frac{1}{n} \sum_{t} x_t x_t' \right]^{-1} \xrightarrow{P} Q^{-1} \), but what happens to the second term as \( n \) grows to infinity? Define \( z_t \) as in equation (3.6). Therefore, by the CLT,

\[
\frac{1}{\sqrt{n}} \sum_{t} x_t \epsilon_t = \sqrt{n} \tilde{Z} \xrightarrow{d} N[0, \Omega],
\]

\(^6\)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.
CHAPTER 3. USING OLS AS AN INFERENTIAL TOOL

where \( \Omega \equiv \text{V}[z_i] = \text{E}[z'_ix'_i] \). Therefore, we can use Cramér’s theorem (see (2.12)) as follows: since

\[
\sqrt{n}(\hat{\beta} - \beta) = \left( \frac{1}{n} \sum_i x'_ix'_i \right)^{-1} \frac{1}{\sqrt{n}} \sum_i x'_i \varepsilon_i \\
\left( \frac{1}{n} \sum_i x'_ix'_i \right)^{-1} \xrightarrow{p} Q^{-1} \\
\frac{1}{\sqrt{n}} \sum_i x'_i \varepsilon_i \xrightarrow{d} N(0, \Omega)
\]

the quantity \( \sqrt{n}(\hat{\beta} - \beta) \) converges to a normal rv multiplied by the nonstochastic matrix \( Q^{-1} \); therefore, the linearity property for Gaussian rvs applies (see eq. (2.15)), and as a consequence

\[
\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1} \sigma^2 Q^{-1}) \quad (3.9)
\]

In order to say something on \( \Omega \), we can use the law of iterated expectations (see section 2.2.4) to re-write it as follows:

\[
\Omega = \text{E}[z'_iz'_i] = \text{E}[\varepsilon'_i|x'_i] = \text{E}[\varepsilon'_i|x'_i] \cdot x'_i.
\]

The quantity \( \text{E}[\varepsilon'_i|x_i] \) is a bit of a problem here. We proved quite easily that \( \text{E}[\varepsilon|x] = 0 \) as a natural consequence of the way \( \varepsilon \) is defined (see section 3.1, near the end). However, we know nothing about its second conditional moment (more succinctly definable as 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 \( x \).

In the present chapter, we will assume that \( \text{E}[\varepsilon^2|x] \) is a positive constant, traditionally labelled \( \sigma^2 \):

\[
\text{E}[\varepsilon^2|x] = \sigma^2; \quad (3.10)
\]

the most important implication of this assumption is that the conditional variance \( \text{V}[y_i|x_i] \) is constant for all observations \( i = 1 \ldots n \); this idea is known as homoskedasticity. Therefore, under the homoskedasticity assumption,

\[
\Omega = \text{E}[z'_iz'_i] = \text{E}[\sigma^2|x'_i] = \sigma^2 \cdot \text{E}[x'_i] = \sigma^2 Q
\]

and equation (3.9) simplifies to

\[
\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 Q^{-1}) \quad (3.11)
\]

This result is also important because it provides the basis for justifying the usage of OLS as an estimator of \( \beta \) 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. 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. Suffice it to say that, under homoskedasticity, OLS is hard to beat in terms of efficiency.

We can estimate consistently $Q$ via $n^{-1}XX'$ and $\sigma^2$ via

$$\hat{\sigma}^2 = \frac{e'e}{n} \xrightarrow{p} \sigma^2$$

so the approximate distribution for OLS that we use is

$$\hat{\beta} \xrightarrow{a} N[\beta, \hat{V}]$$

where $\hat{V} = \hat{\sigma}^2(X'X)^{-1}$.

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 packages. Instead, most people usually prefer a variant in which an alternative estimator of $\sigma^2$ is used:

$$s^2 = \frac{e'e}{n-k},$$

where $k$ is the number of elements of $\beta$; 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: $E[s^2] = \sigma^2$.

The difference between the two variants is negligible for any reasonable sample size $n$, 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.

### 3.3 Specification testing

#### 3.3.1 Tests on a single coefficients

Sometimes, we would like to test hypotheses on elements of $\beta$, so that we can decide optimally on which explanatory variables we must include in our regression function. This is often called “specification testing”.

---

7 If you really really care, a proof is given in section 3.A.2, but I don’t care about it very much myself.

8 See Hansen (2019), sections 7.20 and 7.21 if you’re interested.

9 Proof of this is unnecessary, but if you insist, go to subsection 3.A.1.
Let’s begin by testing a very simple hypothesis: $H_0 : \beta_i = 0$. The practical implication of $H_0$ in a model like

$$E[y|x] = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k = \sum_{j=1}^{k} x_j\beta_j$$

is that the impact of $x_i$ on $E[y|x]$ 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 $k = 3$,

Model A  
$y_i = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \epsilon_i$  
(3.13)

Model B  
$y_i = x_1\beta_1 + x_3\beta_3 + \epsilon_i$  
(3.14)

Clearly, if $H_0$ was true, model B would be preferable, chiefly on the grounds of parsimony; 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 $\beta$ such that $g(\beta) = 0$ if and only if $H_0$ is true. In this case, this is very easy: define $\mathbf{u}_i$ the $i$-th column of the identity matrix (aka “extraction vector”); equivalently, we may say that $\mathbf{u}_i$ is a vector of zeros, except for the $i$-th element, which is 1. 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 $\mathbf{a}$. For example, the product $A \cdot \mathbf{u}_2$ yields the second column of $A$, while $\mathbf{u}_2' A$ yields its second row.

By defining $g(\beta) = \mathbf{u}_i' \beta$, the hypothesis $H_0 : \beta_i = 0$ can be written as $H_0 : \mathbf{u}_i' \beta = 0$, and the Jacobian term is simply $\mathbf{u}_i'$. Hence

$$\sqrt{n} \left[ \mathbf{u}_i' \hat{\beta} - \mathbf{u}_i' \beta \right] \overset{d}{\rightarrow} N \left[ 0, \mathbf{u}_i' V \mathbf{u}_i \right]$$

so our test statistic is

$$\sqrt{n} \left[ \mathbf{u}_i' \hat{\beta} - \mathbf{u}_i' \beta \right] = \sqrt{n} \left[ \mathbf{u}_i' \hat{\beta} - \mathbf{u}_i' \beta \right] = t_i$$

(3.15)

where $t_i = \frac{\hat{\beta}_i}{\se_i}$, $\se_i$ is the $i$-th element on the diagonal of $\hat{V}$ and $\se_i = \sqrt{\se_i}$. The quantity $\se_i$ is usually referred to as the standard error of $\hat{\beta}_i$.

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

$$t_{\beta_i=a} = \frac{\hat{\beta}_i - a}{\se_i}$$

(3.16)
3.3. SPECIFICATION TESTING

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.

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 \( \beta \) 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:

\[
E[y_i | x_i] = x_{1i} \beta_1 + x_{2i} \beta_2 + x_{3i} \beta_3
\]

in the previous subsection, we saw the the restricted model corresponding to the constraint \( \beta_2 = 0 \) is

\[
E[y_i | x_i] = x_{1i} \hat{\beta}_1 + x_{3i} \hat{\beta}_3.
\]

Suppose now that instead of \( \beta_2 = 0 \), our conjecture was \( \beta_1 = 1 \); the restricted version of (3.17) would become

\[
E[y_i - x_{1i} | x_i] = x_{2i} \hat{\beta}_2 + x_{3i} \hat{\beta}_3,
\]

and we would be in fact 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 redefine the dependent variable of our model.

One more example: suppose we combine (3.17) with the restriction \( \beta_2 + \beta_3 = 0 \) (or, equivalently, \( \beta_2 = -\beta_3 \)); in this case, the constrained model turns out to be

\[
E[y_i | x_i] = x_{1i} \hat{\beta}_1 + (x_{2i} - x_{3i}) \hat{\beta}_2.
\]

Of course you can combine more than one restriction into a system:

\[
\begin{cases}
\hat{\beta}_1 = 1 \\
\hat{\beta}_2 + \hat{\beta}_3 = 0,
\end{cases}
\]

and if you applied these to (3.17), the constrained model would turn into

\[
E[y_i - x_{1i} | x_i] = (x_{2i} - x_{3i}) \hat{\beta}_2.
\]
The best way to represent constraints of the kind we just analysed is via the matrix equation

\[ R \beta = d, \]

where the matrix \( R \) and the vector \( d \) are chosen so as to express the constraints we want to test. The examples above on model (3.17) are translated into the \( R \beta = d \) form in the following table:

<table>
<thead>
<tr>
<th>Constraint</th>
<th>( R )</th>
<th>( d )</th>
<th>Restricted model</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_3 = 0 )</td>
<td>[0 0 1]</td>
<td>0 ( \mathbb{E}[y_i</td>
<td>x_i] = x_1 \beta_1 + x_2 \beta_2 )</td>
</tr>
<tr>
<td>( \beta_1 = 1 )</td>
<td>[1 0 0]</td>
<td>1 ( \mathbb{E}[y_i - x_1</td>
<td>x_i] = x_2 \beta_2 + x_3 \beta_3 )</td>
</tr>
<tr>
<td>( \beta_2 + \beta_3 = 0 )</td>
<td>[0 1 1]</td>
<td>0 ( \mathbb{E}[y_i</td>
<td>x_i] = x_1 \beta_1 + (x_2 - x_3) \beta_2 )</td>
</tr>
<tr>
<td>( \beta_1 = 1 ) ( \beta_2 = \beta_3 )</td>
<td>[1 0 0] [0 1 -1]</td>
<td>( \mathbb{E}[y_i - x_1</td>
<td>x_i] = (x_2 - x_3) \beta_2 )</td>
</tr>
</tbody>
</table>

Since the Jacobian of \( R \beta - d \) with respect to \( \beta \) is just the matrix \( R \), we could decide on the appropriateness of the restriction by computing the statistic

\[ W = (\hat{R} \beta - d)' (\hat{R} \hat{V} \hat{R})^{-1} (\hat{R} \hat{\beta} - d) \quad (3.18) \]

and matching it to the \( \chi^2 \) 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 \( R \beta = d \) 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 \( \beta_j \) 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

\[ \beta = \begin{pmatrix} a \\ \alpha_1 \\ \alpha_2 \end{pmatrix} \]
by means of an OLS regression, and the hypothesis “constant returns to scale”
would simply amount to $R\beta = d$, with

$$R = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \quad d = 1.$$  

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 $\beta$ to incorporate the information $a_1 + a_2 = 1$, and there is no
reason why $\hat{\beta}$ should.

In section 3.6, 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:

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; 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:

$$p_i = \beta_0 + \beta_1 s_i + \beta_2 b_i + \beta_3 a_i + \beta_4 x_i + \epsilon_i$$  

(3.19)

where $p_i$ is the log price of the $i$-th house and the explanatory variables are:

<table>
<thead>
<tr>
<th>Legend</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>lsize</td>
<td>log of living area, hundreds of square feet</td>
</tr>
<tr>
<td>baths</td>
<td>number of baths</td>
</tr>
<tr>
<td>age</td>
<td>age of home at time of sale, years</td>
</tr>
<tr>
<td>pool</td>
<td>= 1 if home has pool, 0 otherwise</td>
</tr>
</tbody>
</table>

Models of this type, where the dependent variable is the price of a good and
the explanatory variables its features, are commonly called hedonic models.
In this case (like in most hedonic models), the dependent variable is in loga-
rithm; 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 29 in
Chapter 1).

---

13 After reading this section, the reader might want to go back to section 1.6 and read it again from a different perspective.
14 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

<table>
<thead>
<tr>
<th></th>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>8.85359</td>
<td>0.0483007</td>
<td>183.3</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>lsize</td>
<td>1.03696</td>
<td>0.0232121</td>
<td>44.67</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>baths</td>
<td>-0.00515142</td>
<td>0.0130688</td>
<td>-0.3942</td>
<td>0.6935</td>
</tr>
<tr>
<td>age</td>
<td>-0.00238675</td>
<td>0.000270997</td>
<td>-8.807</td>
<td>2.29e-18 ***</td>
</tr>
<tr>
<td>pool</td>
<td>0.106793</td>
<td>0.0226757</td>
<td>4.710</td>
<td>2.61e-06 ***</td>
</tr>
</tbody>
</table>

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

As you can see, the output is divided into two tables; the most interesting is the top one, which contains $\hat{\beta}$ 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 $\hat{\beta}$, that is $\hat{\beta}_i$;
3. the corresponding standard error, that is $se_i = \sqrt{s^2 \cdot (\hat{X}'\hat{X})^{-1}}$;
4. the ratio of those two numbers, that is the t-ratio (see eq. 3.15)
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.\footnote{Why not the constant? you may ask. Nobody cares about the constant.} 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
\[
\eta = \frac{d y}{d x} \cdot \frac{x}{y} = \frac{d \log y}{d \log x}
\]
because \( \frac{d \log y}{d y} = \frac{1}{y} \) and therefore \( d \log y = \frac{dy}{y} \). So, every 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
error for that coefficient is about 0.023, which gives a \( t \)-ratio of 44.67, and the
related \( p \)-value is such an infinitesimal number that the software just
gives you 0.16. 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 meaningful hypotheses \( H_0: \beta_1 = 1 \), it
would be quite easy to compute the appropriate \( t \) statistic as per equation (3.16):
\[
t = \frac{\hat{\beta}_1 - 1}{se_{\hat{\beta}_1}} = \frac{1.03696 - 1}{0.0232121} = 1.592
\]
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
(\(-0.0051514\)). 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 prop-
erty 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 \( \text{baths} \) 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 1.5). 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 \( \text{age} \) and \( \text{pool} \) are highly significant. The
coefficient for age, for example, is about -0.002: each year of age decreases the

---

\(^{16}\)In case you're curious: I can't compute the number exactly, but it's smaller than \( 10^{-310} \).
house value by about 0.2%, which makes sense; the coefficient for the dummy variable pool is about 0.107; therefore, it would seem that having a pool increases the house value by a little over 10%, which makes sense, again.

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 $e'e$ (157.88) and the square root of $s^2 = e'e/(n-k)$, 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%, which means that our model, all in all, does a fair job at describing price differentials between houses, given how little information on each properties we have. However, our estimate of $\sigma^2$ is quite sizeable, so we shouldn’t expect our model to give us a detailed description of individual house prices.

The line below contains a test\(^{17}\) 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;\(^{18}\) 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 3.6. However, we may have to choose between the two alternatives in which nesting is impossible:

$$y_i \approx x_i' \beta$$

$$y_i \approx z_i' \gamma$$

\(^{17}\)This test, technically, is of the $F$ variety — see section 3.6.1 for its definition.

\(^{18}\)If the data are iid and $f(y|x)$ is Gaussian, then $\hat{\beta}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators of $\beta$ and $\sigma^2$. 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.
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:

\[
\begin{align*}
\text{AIC} & = -2L + 2k \\
\text{BIC} & = -2L + k \log n \\
\text{HQC} & = -2L + 2k \log \log n
\end{align*}
\]

The rule is to pick the model that minimises information criteria. 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”. 19

### 3.5 Exogeneity and causal effects

This section is short, but very important. It is not about the statistical properties of \( \hat{\beta} \), but rather on its interpretation as an estimator of \( \beta \), 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 \( \beta \) measure? If \( \mathbb{E}[y|x] = x'\beta \), then \( \beta \) is simply defined as

\[ \beta = \frac{\partial \mathbb{E}[y|x]}{\partial x}; \]

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.20 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

---

19If you want some more detail on this, have a look at 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).

20The 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”.

---
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 \( x \) 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 “\( x \) 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 \( \beta \) are not those that describe the conditional mean, and therefore, if you define \( \epsilon_i = y_i - x_i' \beta \), the first consequence is that the property \( E[\epsilon_i|x_i] = 0 \) doesn’t hold anymore, and so \( E[\epsilon_i x_i] = \text{Cov}[\epsilon_i, x_i] \neq 0 \). 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.

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 \( \frac{\partial E[y|x]}{\partial x} \). If the nature of the problem is such that our parameters of interest \( \beta \) 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.

### 3.6 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 \( R\beta = d \). In other words, we are looking for a vector \( \hat{\beta} \) that minimises the SSR under the condition that a certain set of linear constraints are satisfied:

\[
\hat{\beta} = \text{Argmin}_{R\beta = d} \| y - X\beta \|;
\]  

(3.23)

compare (3.23) with equation (1.14): OLS is defined as the unconstrained SSR minimiser (we can choose \( \hat{\beta} \) among all \( k \)-element vectors); RLS, instead, can only be chosen among those vectors \( \beta \) that satisfy \( R\beta = d \). Figure 3.2 exemplifies the situation for \( k = 2 \).
Define the restricted residuals as \( \hat{e} = y - X\hat{\beta} \); we will be interested in comparing them with the OLS residuals, so in this section we will denote them as \( \tilde{e} = y - X\tilde{\beta} = M_Xy \) to make the distinction typographically evident.

Figure 3.2: Example: two-parameter vector

The ellipses are the contour lines of the function \( e' e \). 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 \( \tilde{\beta}_1, \tilde{\beta}_2 \).

A couple of remarks can already be made even without knowing what the solution to the problem in (3.23) is. First, since \( \hat{\beta} \) is an unrestricted minimiser, \( \hat{e}' \hat{e} \) cannot be larger than the constrained minimum \( \tilde{e}' \tilde{e} \). However, the inequality \( \hat{e}' \hat{e} \leq \tilde{e}' \tilde{e} \) can be made more explicit by noting that

\[
M_X \tilde{e} = M_X (y - X\tilde{\beta}) = M_X y = \hat{e}
\]

and therefore

\[
\hat{e}' \hat{e} = \tilde{e}' M_X \hat{e} = \tilde{e}' \tilde{e} - \tilde{e}' P_X \tilde{e}
\]

so that

\[
\tilde{e}' \tilde{e} - \hat{e}' \hat{e} = \tilde{e}' P_X \tilde{e} \tag{3.24}
\]

where the right-hand side of the equation is non-negative, since \( \tilde{e}' P_X \tilde{e} \) can be written as \( (P_X \tilde{e})' (P_X \tilde{e}) \), which is a sum of squares.\(^2\)

In order to solve (3.23) for \( \tilde{\beta} \), we need to solve a constrained optimisation problem, which is not complicated once you know how to set up a Lagrangean.

\(^2\)In fact, the same claim follows more elegantly by the fact that \( P_X \) is, by construction, positive semi-definite.
CHAPTER 3. USING OLS AS AN INFERENTIAL TOOL

The details, however, are not important here and I’ll give you the solution straight away:

\[ \tilde{\beta} = \hat{\beta} - (X'X)^{-1}R' \left[ R(X'X)^{-1}R' \right]^{-1} (R\hat{\beta} - d); \]  

(3.25)

derivation of this result is provided in the separate subsection 3.A.3, so you can skip it if you want.

The statistical properties of \( \tilde{\beta} \) are proven in section 3.A.4, but the most important point to make here are: if \( R\hat{\beta} = d \), then \( \tilde{\beta} \) is consistent just like the OLS estimator \( \hat{\beta} \), but has the additional advantage of being more efficient. If, on the contrary, \( R\hat{\beta} \neq d \), then \( \tilde{\beta} \) is inconsistent. The practical consequence of this fact is that, if we were certain that the equation \( R\hat{\beta} = d \) 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 \( \tilde{\beta} \) is to run OLS on the restricted model and then “undo” the restrictions: for example, if you take model (3.17), reproduced here for your convenience

\[ y_i = x_{1i} \beta_1 + x_{2i} \beta_2 + x_{3i} \beta_3 + \epsilon_i \]

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

\[ y_i - x_{1i} = (x_{2i} + x_{3i}) \hat{\beta}_2 + \epsilon_i, \]  

(3.26)

that can be “unravelled” as

\[ y_i = x_{1i} \cdot 1 + x_{2i} \hat{\beta}_2 + x_{3i} \hat{\beta}_2 + \epsilon_i \]

and then forming \( \hat{\beta} \) as \([1, \hat{\beta}_2, \hat{\beta}_2]\), where \( \hat{\beta}_2 \) is the OLS estimate of equation (3.26).

Nevertheless, equation (3.25) is useful for proving an important result. Let’s define the vector \( \lambda \) as

\[ \lambda = [R(X'X)^{-1}R']^{-1} (R\hat{\beta} - d); \]

by premultiplying (3.25) by \( X \) we get:

\[ X\tilde{\beta} = \tilde{y} = \bar{y} - X(X'X)^{-1}R' \lambda \]

which in turn implies

\[ \hat{\theta} = \hat{\theta} + X(X'X)^{-1}R' \lambda \]

By using \( \hat{\theta} = M_X y \implies \hat{\theta} X = 0 \), we can use the above equation to get

\[ \hat{\theta}' \hat{\theta} = \hat{\theta}' \hat{\theta} + \lambda' R(X'X)^{-1}R' \lambda \]

but by the definition of \( \lambda \),

\[ \lambda' R(X'X)^{-1}R' \lambda = (R\hat{\beta} - d)' [R(X'X)^{-1}R']^{-1} (R\hat{\beta} - d) \]
3.6. RESTRICTED LEAST SQUARES AND HYPOTHESIS TESTING

so finally

\[ \hat{\varepsilon}'\hat{\varepsilon} - \hat{\varepsilon}'\hat{\varepsilon} = (R\hat{\beta} - d)' \left[ R(X'X)^{-1}R' \right]^{-1} (R\hat{\beta} - d). \]  (3.27)

Note that the right-hand side of equation (3.27) is very similar to (3.18). In fact, if our estimator for \( \sigma^2 \) is \( \hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/n \), we can combine equations (3.18), (3.24) and (3.27) to write the \( W \) statistic as:

\[ W = \frac{\hat{\varepsilon}'\hat{\varepsilon} - \hat{\varepsilon}'\hat{\varepsilon}}{\hat{\sigma}^2} \cdot \frac{n - k}{p}; \]  (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 \( W \).

3.6.1 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

\[ F = \frac{\hat{\varepsilon}'\hat{\varepsilon} - \hat{\varepsilon}'\hat{\varepsilon}}{p} = \frac{n - k}{p}; \]  (3.29)

it can be easily seen that there are two differences between \( F \) and \( W \): \( F \) uses \( s^2 \), the unbiased estimator of \( \sigma^2 \) 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:

\[ W = p \cdot F \cdot \frac{n}{n - k} \iff F = \frac{n - k}{n} (W / p) \]

so in the standard case, when \( n \) is much larger than \( k \), you have that \( W = p \cdot F \). 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 \( \varepsilon_i \) 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 \( \varepsilon_i \) 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 \( \sigma^2 \) by using \( \hat{\varepsilon} \) instead of \( \hat{\varepsilon} \):

\[ \hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} \]  (3.30)
It can be proven that intuition is right, and if $H_0$ is true $\hat{\sigma}^2$ is indeed consistent for $\sigma^2$. If we use $\hat{\sigma}^2$ instead of $\hat{\sigma}^2$ in equation (3.28), we obtain the so-called **LM statistic**:\(^{22}\)

$$LM = n \frac{\hat{e}'\hat{e} - \hat{\hat{e}}'\hat{\hat{e}}}{\hat{\hat{e}}'\hat{\hat{e}}} = n \frac{\hat{e}'P_X\hat{e}}{\hat{\hat{e}}'\hat{\hat{e}}},$$  \hspace{1cm} (3.31)

where equality comes from (3.24). Since $\hat{e}'\hat{e}$ cannot be less than $\hat{\hat{e}}'\hat{\hat{e}}$, 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 $\chi^2_p$.\(^{23}\)

The nice feature of the $LM$ statistic is that it can be computed via a neat trick, known as **auxiliary regression**:

1. run OLS on the constrained model and compute the residuals $\hat{e}$;
2. run OLS on a model where the dependent variable is $\hat{e}$ 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 $\frac{\hat{e}'P_X\hat{e}}{\hat{\hat{e}}'\hat{\hat{e}}}$, so in the present case $\frac{\hat{e}'P_X\hat{e}}{\hat{\hat{e}}'\hat{\hat{e}}}$ 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 restricted model as follows: combine equation (3.19) with $\beta_1 = 1$ and $\beta_2 = 0$. The restricted model becomes

$$p_i - s_i = \beta_0 + \beta_3 a_i + \beta_4 x_i + \varepsilon_i. \hspace{1cm} (3.32)$$

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

---

\(^{22}\)The reason for the name is that this test statistic can be shown to be a “Lagrange Multiplier” test if normality of $\varepsilon_i$ 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 18 in Section 3.4.2.

\(^{23}\)The reader may want to verify that alternative formulations of the $W$ and $LM$ statistics are possible using $\hat{\sigma}^2$ and $\hat{\sigma}^2$, or the $R^2$ indices from the two models.
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

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>8.94603</td>
<td>0.00833147</td>
<td>1074</td>
</tr>
<tr>
<td>age</td>
<td>-0.00247396</td>
<td>0.000232353</td>
<td>-10.65</td>
</tr>
<tr>
<td>pool</td>
<td>0.115810</td>
<td>0.0221914</td>
<td>5.219</td>
</tr>
</tbody>
</table>

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 $\chi^2_2 = 6.99$, 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 from model (3.32) and regress them against the explanatory variables of the unrestricted model (3.19). In this case, you get

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>-0.0924377 0.0483007</td>
<td>-1.914</td>
<td>0.0588 *</td>
</tr>
<tr>
<td>lsize</td>
<td>0.0398664 0.0232121</td>
<td>1.592</td>
<td>0.1115</td>
</tr>
<tr>
<td>baths</td>
<td>-0.06515142 0.0130688</td>
<td>-0.3942</td>
<td>0.6935</td>
</tr>
<tr>
<td>age</td>
<td>8.72115e-05 0.000270997</td>
<td>0.3218</td>
<td>0.7476</td>
</tr>
<tr>
<td>pool</td>
<td>-0.00901706 0.0226757</td>
<td>-0.3977</td>
<td>0.6909</td>
</tr>
</tbody>
</table>
SSR = 157.884, R-squared = 0.001493

The auxiliary regression per se is not particularly interesting: its parameters don't have any meaningful interpretation.\footnote{For curiosity: the SSR is the same as in the restricted model. Can you prove why analytically? It's not difficult} 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. \footnote{If you need to refresh the notion of confidence interval, go back to the end of section 2.3.2.}

\section*{3.7 Prediction}

Once a model is estimated and we have CAN estimators of $\beta$ and $\sigma^2$, we may want to answer the following question: if we were a given a new datapoint for which the vector of covariates is known and equal to $\hat{x}$, what could we say about the value of the dependent variable $\hat{y}$ 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 $\beta$ and $\sigma^2$ we could say, however, that

$$E[\hat{y}|\hat{x}] = \hat{x}'\beta \quad V[\hat{y}|\hat{x}] = \sigma^2$$

If we were willing to entertain the claim that $\epsilon$ is normal, we could even build a confidence interval\footnote{This may seem obvious, but actually is less simple than one may think: 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.} and say

$$P(|\hat{y} - m| < 1.96\sigma) \approx 95\% \quad (3.33)$$

where $m = \hat{x}'\beta$. The problem is that we don't observe $\beta$; instead, we observe $\hat{\beta}$, so, assuming that the best way to make a point prediction of a random variable is to take its expectation,\footnote{For curiosity: the SSR is the same as in the restricted model. Can you prove why analytically? It’s not difficult} the best we can do to predict $y$ is computing

$$\hat{y} = \hat{x}'\hat{\beta}.$$ 

Note, however, that $\hat{\beta}$ 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

$$e^* = \hat{y} - \hat{\hat{y}} = (\hat{x}'\beta + \hat{\epsilon}) - \hat{x}'\hat{\beta} = \hat{\epsilon} + \hat{x}'(\beta - \hat{\beta});$$
the expression above reveals that our prediction can be wrong for two reasons:
(a) because $\hat{\epsilon}$ is inherently unpredictable: our model does not contain all the
features 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 $\beta$, but only its estimate $\hat{\beta}$.

If $\hat{\epsilon}$ can be assumed independent of $\hat{\beta}$ (as is normally safe to do), then the
variance of the difference is the sum of the variances:

$$V[e^∗] = V[(\hat{y} - \hat{y})|\hat{x}] = V[\hat{\epsilon}] + V[\hat{x}'(\beta - \hat{\beta})] = \sigma^2 + \hat{x}'V\hat{x}.$$  

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 $\sigma^2$, 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,  

$$\hat{x}' = [1 \ 2.708 \ 2 \ 5 \ 0];$$

(the number 2.708 is just log(1500/100); since

$$\hat{\beta}' = [8.8536 \ 1.037 \ -0.00515 \ -0.00239 \ 0.107]$$

simple multiplication yields $\hat{y} = 11.6395$. Therefore the predicted price is about
US$ 113491 (that is, $\exp(11.6395)$).

As for the variance,\(^{27}\) we need $\hat{V}$, that is $s^2(X'X)^{-1}$:

$$\hat{V} = 0.0001 \times \begin{bmatrix}
23.330 & -9.914 & 2.288 & -0.025 & 1.724 \\
-9.914 & 5.388 & -2.245 & -0.010 & -0.705 \\
2.288 & -2.245 & 1.708 & 0.016 & -0.021 \\
-0.025 & -0.010 & 0.016 & 0.001 & -0.001 \\
1.724 & -0.705 & -0.021 & -0.001 & 5.142
\end{bmatrix}$$

It turns out that

$$V[\hat{y}|x] = \hat{\sigma}^2 + \hat{x}'\hat{V}\hat{x} = 0.0606082 + 6.28845 \cdot 10^{-05} = 0.0606711;$$

since $\sqrt{0.0606711} \approx 0.2463$, we can even calculate a 95% confidence interval
around our prediction as

$$\hat{y} \pm 1.96 \sqrt{V[\hat{y}]} = 11.6395 \pm 1.96 \times 0.2463$$

\(^{27}\)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.

\(^{28}\)Of course, we could have used the asymptotic version $\hat{\sigma}^2(X'X)^{-1}$ and very little would have
changed.
so we could expect that, with a probability of 0.95, the log price of our hypothet-
cal 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 vari-
ance of the predictor depends almost entirely on the “model uncertainty” com-
ponent $\hat{\sigma}^2$ and very little on the “parameter uncertainty” component $\hat{x}' \hat{V} \hat{x}$. This
is not surprising, in the light of the fact that, as $n \to \infty$, the latter component
should vanish, since $\hat{\beta}$ 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 $\sigma^2$.

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

$$y_i = x_i \beta_1 + z_i \beta_2 + \epsilon_i;$$  \hspace{1cm} (3.34)

if you try to estimate $\beta_1$ 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:29

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i (x_i \beta_1 + z_i \beta_2 + \epsilon_i)}{\sum_{i=1}^{n} x_i^2} = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_i z_i}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i \epsilon_i}{\sum_{i=1}^{n} x_i^2}$$  \hspace{1cm} (3.35)

If you take probability limits, you’ll find that, even if $E[\epsilon_i|x_i] = 0$,

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{E[x_i z_i]}{E[x_i^2]}$$

and the estimator would be inconsistent unless $\beta_2 = 0$ and/or $E[x_i z_i] = 0$. The
solemn maxim the student receives at this point is “if you omit a relevant regres-
sor ($z_i$ in our case, that is relevant if $\beta_2 \neq 0$) then your estimates will be incorrect,
unless the omitted variable miraculously happens to be uncorrelated with $x_i$”.
The phenomenon is usually called omitted variable bias.

29 I could extend this example to matrices, but it would be totally unnecessary to grasp my point.
3.8. THE SO-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 \( \beta_1 \) 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 \( z_i \). 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

\[
E[y_i|x_i] = \beta_1 x_i + \beta_2 E[z_i|x_i].
\]

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.

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\(^\text{32}\) website. The variables I’m using for this example are

\(^{30}\)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 \( \beta_1 \) consistently via a technique called instrumental variable estimation, which is the object of Chapter 6.

\(^{31}\)In fact, there is an interesting link the bias you get from variable omission and the one you get from endogeneity (see section 3.5). Maybe I’ll write it down at some point.

\(^{32}\)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](http://wdi.worldbank.org) or through DBnomics [https://db.nomics.world/].
CHAPTER 3. USING OLS AS AN INFERENTIAL TOOL

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NY.GDP.PCAP.PP.KD</td>
<td>GDP per capita based on purchasing power parity (PPP).</td>
</tr>
<tr>
<td>SH.MED.BEDS.ZS</td>
<td>Hospital beds (per 1,000 people)</td>
</tr>
<tr>
<td>NV.AGR.TOTL.ZS</td>
<td>Agriculture, value added (% of GDP)</td>
</tr>
</tbody>
</table>

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 $l_{\text{gdp}}$, $l_{\text{hbeds}}$ and $l_{\text{agri}}$. 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 $l_{\text{agri}}$ becomes insignificant (and besides, its sign changes). On the contrary, you find that GDP matters a lot.

### Table 3.2: Two regressions on WDI data

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>1.090**</td>
<td>-4.876**</td>
</tr>
<tr>
<td></td>
<td>(0.1168)</td>
<td>(1.231)</td>
</tr>
<tr>
<td>$l_{\text{agri}}$</td>
<td>-0.2916**</td>
<td>0.08467</td>
</tr>
<tr>
<td></td>
<td>(0.05794)</td>
<td>(0.09217)</td>
</tr>
<tr>
<td>$l_{\text{gdp}}$</td>
<td>0.5655**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1163)</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.2635</td>
<td>0.4496</td>
</tr>
</tbody>
</table>

(standard error in parentheses)

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 $l_{\text{hbeds}}$ on a wider information set, which includes GDP per capita,\(^\text{33}\) the effect disappears. That is, for a given level of economic development\(^\text{34}\) there is no visible link between hospital beds and agriculture. To put it more explicitly: if you compare two countries

---

\(^{33}\)In the applied economic jargon: "once you control for GDP".

\(^{34}\)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?
3.A. ASSORTED RESULTS

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).

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 (3.2), $e = M_Xy = M_X\varepsilon$. Therefore, the sum of squared residuals can be written as

$$e'e = e'M_X\varepsilon = e'\varepsilon - e'X(X'X)^{-1}X'\varepsilon;$$

now, given the definition of $\hat{\sigma}^2$,

$$\hat{\sigma}^2 = \frac{e'e}{n},$$

divide everything by $n$ and take probability limits; the first bit is easy:

$$\frac{e'e}{n} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \xrightarrow{P} \mathbb{E}[\varepsilon_i^2] = \sigma^2.$$

On the other hand, equations (3.7) and (3.8) say that

$$\frac{e'X}{n} \xrightarrow{p} 0 \quad \frac{XX'}{n} \xrightarrow{p} Q$$

and therefore

$$\hat{\sigma}^2 = \frac{e'M_X\varepsilon}{n} = \frac{e'e}{n} - \frac{e'X(X'X)^{-1}X'\varepsilon}{n} \xrightarrow{p} \sigma^2 - 0'Q^{-1} \cdot 0 = \sigma^2.$$

3.A.2 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 $\mathbb{E} [\hat{\beta}] = \beta$; linearity means that $\hat{\beta}$ can be written as $\hat{\beta} = L'y$.

The OLS statistic enjoys both properties ($L'$ being equal to $(X'X)^{-1}X'$ for OLS), but

---

35The discerning reader will doubtlessly spot the parallel with the discussion we had on the Frisch-Waugh theorem in section 1.5.
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 $X$ is a matrix of fixed constants and does not contain random variables, because in this case we can shuffle $X$ in and out of the expectation operator $E[\cdot]$ any way we want. Considering the case when $X$ contains random variables makes the proof more involved.

Here goes: take a linear estimator $\hat{\beta}$ defined as $\hat{\beta} = L'y$. In order for it to be unbiased, the following must hold

$$E[\hat{\beta}] = E[L'(X\beta + \varepsilon)] = L'X\beta + L'E[\varepsilon] = \beta;$$

it is safe to assume that $E[\varepsilon] = 0$, so the unbiasedness requirement amounts to $L'X = I$. Note that, in the standard case, there are infinitely many matrices that satisfy this requirement, since $n > k$ and $X$ is a “tall” matrix. In the OLS case, $L' = (X'X)^{-1}X'$, and the requirement is met. Therefore, under unbiasedness, $\hat{\beta}$ can be written as $\hat{\beta} = \beta + L'\varepsilon$.

Now consider the variance of $\hat{\beta}$:

$$V[\hat{\beta}] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[L'\varepsilon\varepsilon'L] = L'E[\varepsilon\varepsilon']L = L'V[\varepsilon]L';$$

under homoscedasticity, $V[\varepsilon] = \sigma^2I$, so

$$V[\hat{\beta}] = \sigma^2L'L; \quad (3.36)$$

again, OLS is just a special case, so the variance of $\hat{\beta}$ is easy to compute as $V[\hat{\beta}] = \sigma^2[(X'X)^{-1}X'][X(X'X)^{-1}] = \sigma^2(X'X)^{-1}$.

The gist of the theorem lies in proving that the difference

$$V[\hat{\beta}] - V[\hat{\beta}] = \sigma^2L'L - \sigma^2(X'X)^{-1} = \sigma^2[L'L - (X'X)^{-1}]$$

is positive semidefinite any time $L \neq (X'X)^{-1}X'$, and therefore OLS is more efficient than $\hat{\beta}$. This is relatively easy: define $D = L' - (X'X)^{-1}X'$, which has to be nonzero unless $\hat{\beta} = \beta$. Therefore, $D'D$ must be positive semidefinite (see section 1.A.7).

$$D'D = [L' - (X'X)^{-1}X'][L - X(X'X)^{-1}] = L'L - (X'X)^{-1}X'L - L'X(X'X)^{-1} + (X'X)^{-1};$$

under unbiasedness, $L'X = X'L = I$, so the expression above becomes

$$D'D = L'L - (X'X)^{-1} \quad (3.37)$$

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 $X$ 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.3 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\(^\text{36}\), but here I’ll give you a super-simplified account for your convenience.

If you have to find maxima and/or minima of a function \(f(x)\) subject to a system of constraints \(g(x) = 0\), you set up a function, called the **Lagrangean**, in which you sum the objective function to a linear combination of the constraints, like this:

\[
\mathcal{L}(x, \lambda) = f(x) + \lambda' g(x).
\]

The elements of the vector \(\lambda\) are known as “Lagrange multipliers”.

For example, the classic microeconomic problem of a utility-maximising consumer is represented as

\[
\mathcal{L}(x, \lambda) = U(x) + \lambda \cdot (Y - p'x)
\]

where \(x\) is the bundle of goods, \(U(\cdot)\) is the utility function, \(Y\) is disposable income and \(p\) is the vector of prices. In this example, the only constraint you have is the budget constraint, so \(\lambda\) is a scalar.

The solution has to obey two conditions, known as the “first order conditions”:

\[
\frac{\partial \mathcal{L}}{\partial x} = 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda} = 0,
\]

so in practice you differentiate the Lagrangean with respect to your variables and \(\lambda\), and then check if there are any solutions to the system of equations you get by setting the partial derivatives to 0. If the solution is unique, you’re all

\(^{36}\text{One I especially like is Dixit (1990), but for a nice introductory treatment I find Dadkhah (2011) hard to beat.}\)
set. In the utility function example, applying equations (3.38) gives the standard microeconomic textbook solution to the problem:

\[
\frac{\partial U(x)}{\partial x} = \lambda p \quad Y = p'x.
\]

in words: at the maximum, (a) marginal utilities are proportional to prices (or \( \frac{\partial U}{\partial x_i} / \frac{\partial U}{\partial x_j} = \frac{p_i}{p_j} \) if you prefer) and (b) you should spend all your income.

In the case of RLS, the Lagrangean is

\[
\mathcal{L}(\beta, \lambda) = \frac{1}{2} e' e + \lambda' (R\beta - d).
\]

where of course \( e = y - X\beta \). The derivative of \( \mathcal{L} \) with respect to \( \lambda \) is just the constraint; as for the other one, since the derivative of \( e \) with respect to \( \beta \) is \(-X\), 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

\[
\frac{\partial \mathcal{L}}{\partial \beta} = -X' e + R' \lambda
\]

so the first order condition with respect to \( \beta \) can be written as

\[
X' \hat{e} = R' \lambda, \tag{3.39}
\]

where \( \hat{e} \) is the vector that satisfies equation (3.39), defined as \( y - X\hat{\beta} \). By premultiplying (3.39) by \((X'X)^{-1}\) we get

\[
(X'X)^{-1} X'(y - X\hat{\beta}) = \hat{\beta} - \beta = (X'X)^{-1} R' \lambda,
\]

which of course implies

\[
\hat{\beta} = \beta - (X'X)^{-1} R' \lambda \tag{3.40}
\]

So the constrained solution \( \hat{\beta} \) can be expressed as the OLS vector \( \hat{\beta} \), plus a “correction factor”, proportional to \( \lambda \). If we premultiply (3.40) by \( R \) we get

\[
\lambda = [R(X'X)^{-1} R']^{-1} (R\hat{\beta} - d) \tag{3.41}
\]

because \( R\hat{\beta} = d \) by construction. Interestingly, \( \lambda \) itself is proportional to \( (R\hat{\beta} - d) \), 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 (3.40) and (3.41):

\[
\hat{\beta} = \beta - (X'X)^{-1} R' [R(X'X)^{-1} R']^{-1} (R\hat{\beta} - d).
\]

\(^{37}\)Note that I divided the objective function by 2. Clearly, the solution is the same, but the algebra is somewhat simplified.
3.A. ASSORTED RESULTS

3.A.4 Asymptotic properties of the RLS estimator

Begin by (3.25)

\[ \hat{\beta} = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1} R' [R(\mathbf{X}'\mathbf{X})^{-1} R']^{-1} (R\hat{\beta} - \mathbf{d}); \]

from this equation, it is quite easy to see that \( \tilde{\beta} \) is an affine function of \( \hat{\beta} \); this will be quite useful. Now define \( H \) as

\[ H = \operatorname{plim} \left( (\mathbf{X}'\mathbf{X})^{-1} R' [R(\mathbf{X}'\mathbf{X})^{-1} R']^{-1} \right) = Q^{-1} R' [RQ^{-1} R']^{-1} \]

(3.42)

normally, this is a \((k \times p)\) matrix with rank \( p \); also note that \( HR \) is idempotent.

If \( \hat{\beta} \) is consistent, then \( \tilde{\beta} \rightarrow H \cdot (R\beta - \mathbf{d}) \)

As for efficiency, let me briefly remind you what we mean by “efficiency”: 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 \( V(b) - V(a) \) is a positive semi-definite matrix (psd for short — see section 1.A.7, especially figure 1.3).

There are a few matrix algebra results that we are going to need here:

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 \( \tilde{\beta} \) is smaller (in a matrix sense) that that for \( \hat{\beta} \):

\[ \operatorname{AV}[\tilde{\beta}] = (I - HR) \cdot \operatorname{AV}[\hat{\beta}] \cdot (I - R' H') \]

if \( \operatorname{AV}[\hat{\beta}] = \sigma^2 Q^{-1} \), then note that

\[ HRQ^{-1} = Q^{-1} R' [RQ^{-1} R']^{-1} RQ^{-1}, \]

(3.43)

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 \( \tilde{\beta} \) can be written as

\[ \operatorname{AV}[\tilde{\beta}] = \sigma^2 \{ Q^{-1} - HRQ^{-1} - Q^{-1} R'H' + HRQ^{-1} R'H' \} = \sigma^2 \{ Q^{-1} - HRQ^{-1} \} = \operatorname{AV}[\hat{\beta}] - \sigma^2 HRQ^{-1} \]

---

38 A slightly fuller discussion is at the end of section 2.3.2.
39 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 (3.43).

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 $[RQ^{-1}R']^{-1}$ (property 1). Finally, by using property 1 again, we find that $Q^{-1}R'[RQ^{-1}R']^{-1}RQ^{-1}$ 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 \( E[y|x] \) exists and is linear: \( E[y|x] = x'\beta \);

4. the conditional variance \( V[y|x] \) exists and does not depend on \( x \) at all, so it’s a positive constant: \( V[y|x] = \sigma^2 \).

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 \( \hat{\beta} \) 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 \( \hat{\beta} \xrightarrow{p} \beta \) was that \( E[y|x] = x'\beta \) (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,

\[
E[y_i|x_i] = m(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_q x_i^q = \sum_{j=0}^{q} \beta_j x_i^j.
\]

2. Our data comprises observation for which the DGP is partly different. That is, we have \( j = 1 \ldots m \) separate sub-populations, for which

\[
E[y_i|x_i] = x'_i \beta_j
\]

where \( j \) is the class that observation \( i \) belongs to. For example, we have data on European and American firms, and the vector \( \beta \) is different on the two sides of the Atlantic (in this case, \( m = 2 \)).

#### 4.1.1 The RESET test

As I repeated several times over, the hypothesis of linearity simply means that \( E[y_i|x_i] \) 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 (see Section 1.3.4 for a fuller discussion).

For example, suppose \( x_i \) is a scalar; it is perfectly possible to accommodate something like

\[
E[y_i|x_i] = \beta_1 x_i + \beta_2 x_i^2;
\]

(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

\[
\hat{\gamma} = (x'x)^{-1}x'y = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.
\]
4.1. DIAGNOSTICS FOR THE CONDITIONAL MEAN

Clearly, there is no value of $\gamma$ that can make $u_i$ the difference between $y_i$ and $E[y_i|x_i]$, 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 $\beta_1$ nor a simple function of it, so technically there is no way we can use $\hat{\gamma}$ to estimate $\beta_1$ 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 $\beta_1$ and $\beta_2$. However, in a real-life case, where you have a vector of explanatory variables $x_i$, 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

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$

would become

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i^2 + \beta_4 x_i \cdot z_i + \beta_5 z_i^2 + \epsilon_i$$

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. 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 REgression Specification Error Test) is a way to check whether a given specification is likely to be rich enough to need no additional nonlinear effects. 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

$$\hat{y}_i^2 = (\hat{\beta}_1 x_i + \hat{\beta}_2 z_i)^2.$$  

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 3.6.1).

---

1Of 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?

2The reader is invited to work out what happens for various values of $k$. 

---

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).
CHAPTER 4. DIAGNOSTIC TESTING IN CROSS-SECTIONS

1. Run OLS, save residuals $e_i$ and fitted values $\hat{y}_i$;

2. generate squares and cubes $\hat{y}_i^2, \hat{y}_i^3$;

3. run the auxiliary regression

$$e_i = \gamma'x_i + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3 + u_i;$$

4. compute $LM = n \cdot R^2 \sim \chi^2_2$

**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

<table>
<thead>
<tr>
<th>Dependent variable: $\hat{e}$</th>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>195.825</td>
<td>81.6622</td>
<td>2.398</td>
<td>0.0166 **</td>
</tr>
<tr>
<td>lsize</td>
<td>38.9198</td>
<td>17.1107</td>
<td>2.275</td>
<td>0.0230 **</td>
</tr>
<tr>
<td>baths</td>
<td>-0.205739</td>
<td>0.0861412</td>
<td>-2.388</td>
<td>0.0170 **</td>
</tr>
<tr>
<td>age</td>
<td>-0.0900385</td>
<td>0.0393760</td>
<td>-2.287</td>
<td>0.0223 **</td>
</tr>
<tr>
<td>pool</td>
<td>3.97729</td>
<td>1.76245</td>
<td>2.257</td>
<td>0.0241 **</td>
</tr>
<tr>
<td>$\hat{y}_i^2$</td>
<td>-3.42319</td>
<td>1.40727</td>
<td>-2.433</td>
<td>0.0151 **</td>
</tr>
<tr>
<td>$\hat{y}_i^3$</td>
<td>0.103615</td>
<td>0.0399778</td>
<td>2.592</td>
<td>0.0096 ***</td>
</tr>
</tbody>
</table>

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 $\chi^2_2$ 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 “splines” or “fractional polynomials”; if you’re into kinky stuff, try “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 sub-populations. A dummy variable $d_i$ tells us which group observation $i$ belongs
4.1. DIAGNOSTICS FOR THE CONDITIONAL MEAN

to. Suppose that the DGP could be described as follows:

\[
\begin{align*}
\text{Subsample 1} & \quad y_i = \beta_0 + \beta_1 x_i + \epsilon_i \\
\text{Subsample 2} & \quad y_i = \beta_0 + \beta_2 x_i + \epsilon_i
\end{align*}
\]

with \( d_i = 0 \) in subsample 1 and \( d_i = 1 \) in subsample 2. Note that the model could be rewritten as

\[
y_i = \beta_0 + \left[ \beta_1 + (\beta_2 - \beta_1) d_i \right] x_i + \epsilon_i = \beta_0 + \beta_1 x_i + \gamma d_i \cdot x_i + \epsilon_i \tag{4.1}
\]

where \( \gamma = \beta_2 - \beta_1 \). 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 \( \beta_1 \) and \( \beta_2 \) is easy: all you need to do is check whether the regressor \( d_i \cdot x_i \) 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

\[
y_i = \beta_0 + \beta_1 x_i + \gamma_0 d_i + \gamma_1 d_i \cdot x_i + \epsilon_i,
\]

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 that you need to remember this when computing marginal effects. The marginal effect for \( x_i \) in equation (4.1), for example, would be

\[
\frac{\partial \mathbb{E}[y_i|x_i]}{\partial x_i} = \beta + \gamma d_i,
\]

that is, \( \beta \) if \( d_i = 0 \) and \( \beta + \gamma \) if \( d_i = 1 \).

Generalising this idea is not difficult: if you have \( m \) categories and you split the vector \( x_i \) into \( \tilde{x}_i \) (constant effect across sub-populations) and \( \dot{x}_i \) (possibly different effect across sub-populations) you can write the model as

\[
y_i = \tilde{x}_i' \gamma + \sum_{j=1}^{m} (d_{ij} \dot{x}_i) \gamma_j + \epsilon_i
\]

where \( d_{jj} = 1 \) if observation \( i \) belongs to sub-population \( j \), and 0 otherwise. An equivalent (and much more often used) formulation is

\[
y_i = \tilde{x}_i' \beta + \sum_{j \neq q} (d_{ij} \dot{x}_i) \gamma_j + \epsilon_i; \tag{4.2}
\]

the proof of the equation above, if you’re interested, is in section 4.A.1. In practice, you choose one subpopulation as the “reference” category (\( q \) in the equation above), and the \( \gamma_j \) coefficients measure the difference between the \( j \)-th category and the benchmark.
Example 4.2 (R&D and size)

Figure 4.1 depicts the relationship between the size of a firm (log of total revenue, $x_i$ in formulae) and its expenditure in Research and Development (in logs, $y_i$ in formulae). The data come from a sample of 1354 Italian manufacturing firms.\(^3\)

![Figure 4.1: Relationship between revenues and R&D expenditure (in logs)](image)

$$
\hat{y}_i = -2.5141 + 0.7531 x_i
$$

$$
Y = -2.51 + 0.753X
$$

The model looks basically OK: $R^2$ is not at all bad, and the fact the the estimated elasticity of R&D with respect to revenues is positive and very significant,\(^4\) but slightly lower than one, goes well with our economic intuition (bigger firms spend more on R&D than smaller ones; however, the ratio between R&D expenditure and revenues decreases with firm size). However, we may surmise that different firms act differently. Suppose we divide the sample into 3 groups:

1. producers of “basic stuff” (NACE codes 15 to 19)\(^5\)
2. producers of “not-so-basic stuff” (NACE codes 20 to 28)
3. producers of “machines” (NACE codes 29 to 36).

---

\(^3\)Thanks to Alessandro Sterlacchini, who gave me the data for this example.

\(^4\)The corresponding $t$-ratio is roughly 28.5; computing it with the data show in Figure 4.1 is an easy exercise left to the reader.

\(^5\)NACE is the standard EU industry classification (the acronym is for *Nomenclature statistique des Activités économiques dans la Communauté Européenne*). For a list of NACE codes, just google for it.
In our sample, about half of the firms belong to category 2, while the rest is split almost equally between categories 1 and 3 (see Figure 4.2). What happens if we run different models for the three subsamples? See figure 4.3. You can see the resulting coefficients in the three subgraphs. So for example, the elasticity of R&D with respect to revenues for subsample 3 only turns out to be 0.948 (as opposed to 0.753 for the whole sample). It’s no surprise that the estimated parameters are different from each other; the relevant question is: are they significantly different? Or, in other words: sure, the $\hat{\beta}_1$ coefficients are different across groups. But are the $\beta_1$ coefficients (without hats) different from one another? For even if they were not, their estimates would be different from each other anyway. So, the hypothesis we would like to test is: are the estimated $\hat{\beta}_1$ coefficients consistent estimators for the same quantity? If this hypothesis were rejected, we would have conclusive evidence that the conditional expectation $E[y|x]$ is different across subpopulations.

If we create two dummy variables for NACE classes 1 and 3 (we keep 2 as the reference class) we can estimate the three coefficient at the same time by interacting $x_i$ by the two dummies. See Table 4.1 for the results. Take for example...
the elasticity for NACE class 3: the value 0.948 is the result of adding the elasticity for the reference category (0.724) and the coefficient for the specific interaction (0.224); proceeding in the same way, the elasticity for NACE class 1 turns out to be \(0.724 - 0.135 = 0.588\) (give or take some rounding).

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>(z)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_i)</td>
<td>0.723872</td>
<td>0.0353120</td>
<td>20.50</td>
</tr>
<tr>
<td>(d_{3i} \cdot x_i)</td>
<td>0.223810</td>
<td>0.0618350</td>
<td>3.619</td>
</tr>
<tr>
<td>(d_{1i} \cdot x_i)</td>
<td>0.135458</td>
<td>0.0667693</td>
<td>2.029</td>
</tr>
<tr>
<td>(d_{3i} \cdot x_i)</td>
<td>0.837215</td>
<td>0.607043</td>
<td>1.379</td>
</tr>
<tr>
<td>(d_{3i})</td>
<td>1.76931</td>
<td>0.562752</td>
<td>-3.144</td>
</tr>
</tbody>
</table>

\[ R^2 = 0.404707 \]

\[ e'e = 2060.20 \]

Table 4.1: revenues and R&D — with NACE subgroups interactions

However, here we get some extra info: since the coefficient on \(d_{3i} \cdot x_i\) is the difference between the elasticity for subsample 3 (machine makers) and the elasticity for subsample 2 (the reference category), the fact that it is significantly different from 0 (the corresponding p-value is 0.0003) indicates that the two subgroups do in fact behave differently, and therefore, incorporating the extra terms into the original model is a good idea. The Wald test for comparing the two models would give

\[ W = 1354 \cdot \frac{2159.074 - 2060.202}{2060.202} = 224.602 \]

that is a rather astronomical number for a \(\chi^2_4\) distribution (the corresponding p-value is \(1.91703 \times 10^{-47}\)). Therefore, we may safely reject the null, and we should conclude that the elasticities do vary across NACE groups.

When you interact all the parameters by a dummy, then the test is particularly simple (and amounts to what is known as the **Chow test**), since unrestricted SSR 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

\[ W = n \cdot \frac{S_T - (S_1 + S_2)}{S_1 + S_2} \]
because the SSR for the model with all the interactions is equal to the sum of the SSRs for the separate submodels. In fact, this result holds for any number of subcategories; the proof is in subsection 4.A.2.

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 \( E[y|x] = x'\beta \) makes consistency almost automatic (\( E[\varepsilon \cdot x] = 0 \) implies \( \hat{\beta} \xrightarrow{p} \beta \) if some kind of LLN can be invoked), it’s impossible to derive the parallel result for asymptotic normality without assuming homoskedasticity, that is \( E[\varepsilon^2|x] = \sigma^2 \).

Like all assumptions, homoskedasticity is easier to justify in certain circumstances than others. If data come from controlled experiments, \( \varepsilon_i \) can often be interpreted as the disturbance term that contaminated the \( i \)-th experiment; it is normally safe to think that \( \varepsilon_i \) should be independent from \( x_i \), the conditions under which the experiment was performed. Therefore, if \( \varepsilon_i \perp \perp x_i \), no moments of \( \varepsilon_i \) can depend on \( x_i \), 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 \( x_i \) that contain a description of that individual (education, age, work experience and so on), and \( \varepsilon_i \) 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 \( E[\varepsilon_i|x_i] = 0 \), which holds by construction under the linearity hypothesis.

Therefore, assuming that \( \varepsilon_i \) has a finite second moment, in general all we can say is that \( E[\varepsilon_i^2|x_i] \) is some function of \( x_i \):

\[
E[\varepsilon_i^2|x_i] = h(x_i) = \sigma_i^2,
\]

where the function \( h(\cdot) \) is of unknown form (but certainly non-linear, since \( \sigma_i^2 \) can never be negative). Since the variances \( \sigma_i^2 \) may be different across observations, we use the term **heteroskedasticity**.

To simplify notation, in this section all expectation operators will be implicitly understood as conditional on \( X \). In other words, we will treat \( X \) as if it were a matrix of constants. Therefore,

\[
y = X\beta + \varepsilon \quad V[\varepsilon] = E[\varepsilon \varepsilon'] = \Sigma
\]
If our model is heteroskedastic, then $\Sigma$ is a diagonal matrix, where elements along the main diagonal need not be equal to each other, and it would look like this:

$$
\begin{bmatrix}
\sigma_1^2 \\
\sigma_2^2 \\
\vdots \\
\sigma_n^2
\end{bmatrix}
$$

The variance of $\hat{\beta}$ can be simply computed as

$$
V[\hat{\beta}] = V[\beta + (X'X)^{-1}X'\varepsilon] = (X'X)^{-1}X'\Sigma X (X'X)^{-1}
$$

and clearly if $\Sigma = \sigma^2 I$ we go back to $\sigma^2 (X'X)^{-1}$. But under heteroskedasticity the $\hat{V} = \hat{\sigma}^2 (X'X)^{-1}$ matrix may be very far from the actual asymptotic covariance matrix of $\hat{\beta}$ (shown in equation (4.4)), even asymptotically; therefore, our test statistics are very unlikely to be $\chi^2$-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 $\Sigma$ is known.

### 4.2.1 If $\Sigma$ were known

If the matrix $\Sigma$ were observable, then the $\sigma_i^2$ variances would be known (they are just the diagonal elements of $\Sigma$), and getting rid of heteroskedasticity would be easy: define

$$
\hat{y}_i = \frac{y_i}{\sigma_i} \quad \hat{x}_i = \frac{1}{\sigma_i} x_i \quad u_i = \frac{\varepsilon_i}{\sigma_i}
$$

and the model becomes

$$
\hat{y}_i = \hat{x}_i' \beta + u_i
$$

but clearly $V[u_i] = 1$ by construction,\(^7\) so you can happily run OLS on the transformed variables, since $\hat{y}_i$ and $\hat{x}_i$ would both be observable. The resulting estimator

$$
\hat{\beta} = \left[ \sum_{i=1}^{n} \hat{x}_i \hat{x}_i' \right]^{-1} \sum_{i=1}^{n} \hat{x}_i \hat{y}_i = (X'X)^{-1}X'y
$$

could also be written as

$$
\hat{\beta} = \left[ \sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i x_i' \right]^{-1} \sum_{i=1}^{n} \frac{1}{\sigma_i} x_i y_i = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y
$$

\(^6\)In fact, some of our considerations carry over to more general cases, in which $\Sigma$ is a generic symmetric, positive semi-definite matrix, but let’s not complicate matters.

\(^7\)Readers would hopefully not feel offended if I reminded them that a straightforward application of equation (2.6) yields $V[\frac{X}{b}] = V[\frac{X}{b}]b^2$. 

and is called GLS. It can be proven that GLS is more efficient than OLS (the proof is in subsection 4.A.3), and that its covariance matrix equals

$$\text{Var}[\hat{\beta}] = (X'\Sigma^{-1}X)^{-1}$$

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.\(^8\)

Of course, in ordinary circumstances \(\Sigma\) is unknown, but we could use this idea to explore alternative avenues:

1. In some cases, you may have reason to believe that \(\sigma^2_i\) 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 \(\sigma^2_i = Kn_i^{-1}\), 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

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$$

not only leads to a more natural interpretation of the parameters (since \(\beta_1\) can be read as an elasticity), but also alleviates heteroskedasticity problems.

3. Even more generally, it can be proven that, if we have \(\hat{\Sigma} \overset{P}{\rightarrow} \Sigma\), we can use it in the so-called “feasible” version of GLS, or FGLS for short:

$$\hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y$$

but this is normally accomplished by setting an explicit functional form for the conditional variance (the function \(h(\cdot)\) in 4.3). It can be done, but

---

\(^8\)In fact, we wouldn’t even have to observe \(\Sigma\), as long as we had a matrix which is proportional to \(\Sigma\) by a (possibly unknown) scalar factor. If we had \(\Omega = c \cdot \Sigma\), where \(c\) is an unknown positive scalar, we could use \(\Omega\) instead of \(\Sigma\) in equation (4.5), since the scalar \(c\) would cancel out.
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 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 $\hat{V}$, that is

$$\hat{V} = \hat{\sigma}^2 (X'X)^{-1}$$

for hypothesis testing yields statistics that are not asymptotically $\chi^2$-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)

$$V[\hat{\beta}] = (X'X)^{-1} (X'\Sigma X)(X'X)^{-1},$$

or anything asymptotically equivalent, inference would be perfectly standard. This seems impossible, given that $\Sigma$ is unobservable: the middle matrix in the equation above could be written as

$$X'\Sigma X = [x_1 \ x_2 \ \ldots \ x_n] \left( \begin{array}{ccc} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_n^2 \end{array} \right) \left( \begin{array}{c} x_1' \\ x_2' \\ \vdots \\ x_n' \end{array} \right) = \sum_{i=1}^{n} \sigma_i^2 x_i x_i'$$

and it would seem that in order to compute an asymptotically equivalent expression we would need the $\sigma_i^2$ variances (or at the very least consistent estimates).

However, things are different: although $\Sigma$ does in fact contain $n$ distinct unknown elements, the size of $X'\Sigma X$ is $k \times k$,

\footnote{In fact, it’s a symmetric matrix, so the number of its distinct elements is $k(k+1)/2$.}

so in principle we may hope to estimate consistently its elements as $n \to \infty$.

In other terms, even if we can’t estimate consistently the individual variances $\sigma_i^2$, we may be able to estimate consistently the individual elements of the matrix $X'\Sigma X$. H A L W H I T E
4.2. HETEROSKEDASTICITY AND ITS CONSEQUENCES

This is the basic idea that was put forward in White (1980): under heteroskedasticity, OLS is still consistent, so

\[ \hat{\beta} \xrightarrow{p} \beta \implies e_i = y_i - x_i' \hat{\beta} \xrightarrow{p} e_i = y_i - x_i' \beta \implies e_i^2 \xrightarrow{p} \epsilon_i^2 : \]

the difference between the OLS residuals \(e_i\) and the disturbances \(\epsilon_i\) should be “small” in large samples. Since \(\sigma_i^2 = E(\epsilon_i^2)\), we’d expect that the average of \(e_i^2 - \sigma_i^2\) should also be a small quantity. Therefore,

\[ \frac{1}{n} \sum_{i=1}^{n} (e_i^2 - \sigma_i^2) x_i x_i' \xrightarrow{p} [0], \]

where I’m using the [0] notation for “a matrix full of zeros”. Now rewrite (4.4) as

\[ V[\hat{\beta}] = (X'X)^{-1} \left( \sum_{i=1}^{n} \sigma_i^2 x_i x_i' \right) (X'X)^{-1} \]

Therefore, asymptotically you can estimate \(V[\hat{\beta}]\) via

\[ \bar{V} = (X'X)^{-1} \left( \sum_{i=1}^{n} e_i^2 x_i x_i' \right) (X'X)^{-1} \tag{4.6} \]

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 still the original one.\(^\text{10}\)

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.

**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

\(^{10}\)A clever variation on the same principle that goes under the name of **cluster-robust** variance estimation has become inexplicably popular in recent years. I’m not going to describe it in this book, but you should be aware that in some circles you will be treated like the village idiot if you don’t “cluster your standard errors”. For more details, read Cameron and Miller (2015).
Heteroskedasticity-robust standard errors, variant HC0

<table>
<thead>
<tr>
<th></th>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>8.85359</td>
<td>0.0557726</td>
<td>158.7</td>
<td>0.0000  ***</td>
</tr>
<tr>
<td>lsize</td>
<td>1.03696</td>
<td>0.0270429</td>
<td>38.34</td>
<td>1.85e-255 ***</td>
</tr>
<tr>
<td>baths</td>
<td>-0.00815142</td>
<td>0.0150608</td>
<td>-0.3420</td>
<td>0.7323</td>
</tr>
<tr>
<td>age</td>
<td>-0.00239646</td>
<td>0.0150608</td>
<td>-7.943</td>
<td>2.92e-15 ***</td>
</tr>
<tr>
<td>pool</td>
<td>0.106793</td>
<td>0.0239646</td>
<td>4.456</td>
<td>8.69e-06 ***</td>
</tr>
</tbody>
</table>

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)

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

The idea is both simple and powerful: if ε_i is homoskedastic, then both estimators of the parameters covariance matrix are consistent, so \( \hat{V} = \hat{\sigma}^2 (X'X)^{-1} \) and its robust counterpart \( \tilde{V} \) 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.\(^{11}\)

If we re-write \( \hat{V} \) as

\[
\hat{V} = (X'X)^{-1} (\hat{\sigma}^2 X'X) (X'X)^{-1} = (X'X)^{-1} \left( \sum_{i=1}^{n} \hat{\sigma}^2 x_i x_i' \right) (X'X)^{-1},
\]

and compare this expression to (4.6), it’s clear to see that any difference between the two variance estimators comes from the matrix in the middle, which equals \( \sum_{i=1}^{n} \hat{\sigma}^2 x_i x_i' \) for \( \hat{V} \) and \( \sum_{i=1}^{n} e_i^2 x_i x_i' \) for \( \tilde{V} \). Therefore, the difference between them

\[
\frac{1}{n} \sum_{i=1}^{n} (e_i^2 - \hat{\sigma}^2) x_i x_i'
\]

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

\(^{11}\)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.
two estimators would converge to the same limit, and therefore the two esti-
mators would coincide asymptotically; this, of course, wouldn't happen under heteroskedasticity. Therefore, the null hypothesis of White's test is *homoskedasticity*.

It would seem that performing such a test is difficult; fortunately, an asymp-
totically equivalent test is easy to compute by means of an auxiliary regression:

\[ e_i^2 = \gamma_0 + z_i' \gamma + u_i; \]

the vector \( z_i \) contains the cross-products of \( x_i \), that is all combinations of the kind \( x_{il} \cdot x_{mi} \) (with \( l, m = 1 \ldots k \)) that wouldn't yield collinearity (an example is given below). Of course, if \( x_i \) contains a constant term, then \( z_i \) would contain all the elements of \( x_i \), 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 \( \chi^2_p \), where \( p \) is the size of the vector \( z_i \).

For example: suppose that \( x_i \) 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 ta-
table”

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>( x_i )</th>
<th>( w_i )</th>
<th>( d_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_i )</td>
<td>( x_i )</td>
<td>( x_i^2 )</td>
<td>( x_i \cdot w_i )</td>
<td>( x_i \cdot d_i )</td>
</tr>
<tr>
<td>( w_i )</td>
<td>( w_i )</td>
<td>( x_i \cdot w_i )</td>
<td>( w_i^2 )</td>
<td>( w_i \cdot d_i )</td>
</tr>
<tr>
<td>( d_i )</td>
<td>( d_i )</td>
<td>( x_i \cdot d_i )</td>
<td>( w_i \cdot d_i )</td>
<td>( d_i )</td>
</tr>
</tbody>
</table>

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 \( z_i \) would make the auxiliary regression collinear.

Therefore, the vector \( z_i \) would contain

\[ z_i' = [ x_i, \ w_i, \ d_i, \ x_i^2, \ x_i \cdot w_i, \ x_i \cdot d_i, \ w_i^2, \ w_i \cdot d_i ] \]
so the auxiliary regression would read

\[ e_i^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 w_i + \gamma_3 d_i + \gamma_4 x_i^2 + \gamma_5 x_i w_i + \gamma_6 x_i d_i + \gamma_7 w_i^2 + \gamma_8 w_i d_i + u_i \]

where \( u_i \) is the error term of the auxiliary regression. In this case, \( p = 8 \).\(^{12}\)

**Example 4.4**

Running White’s heteroskedasticity test on the hedonic model for houses (see section 3.4) yields:

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>0.806236</td>
<td>0.160351</td>
<td>5.028</td>
</tr>
<tr>
<td>lsize</td>
<td>-0.710993</td>
<td>0.13434</td>
<td>-5.289</td>
</tr>
<tr>
<td>baths</td>
<td>0.0936181</td>
<td>0.0514158</td>
<td>2.181</td>
</tr>
<tr>
<td>age</td>
<td>0.020464135</td>
<td>0.0019047</td>
<td>2.309</td>
</tr>
<tr>
<td>pool</td>
<td>0.267089</td>
<td>0.010599</td>
<td>2.310</td>
</tr>
<tr>
<td>sq_lsize</td>
<td>0.159277</td>
<td>0.0307527</td>
<td>5.179</td>
</tr>
<tr>
<td>X2_X3</td>
<td>-0.0409056</td>
<td>0.0240448</td>
<td>-0.851</td>
</tr>
<tr>
<td>X2_X4</td>
<td>-0.0163418</td>
<td>0.0005749</td>
<td>-3.099</td>
</tr>
<tr>
<td>X2_X5</td>
<td>-0.164525</td>
<td>0.0499579</td>
<td>-3.297</td>
</tr>
<tr>
<td>sq_baths</td>
<td>0.00360873</td>
<td>0.0013648</td>
<td>2.309</td>
</tr>
<tr>
<td>X3_X4</td>
<td>0.000204179</td>
<td>0.0000669</td>
<td>1.091</td>
</tr>
<tr>
<td>X3_X5</td>
<td>0.065657</td>
<td>0.0252631</td>
<td>2.546</td>
</tr>
<tr>
<td>X4_X5</td>
<td>0.00245662</td>
<td>0.0007339</td>
<td>3.258</td>
</tr>
</tbody>
</table>

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 \( \chi^2_{13} \) 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.

\(^{12}\)It’s easy to prove that, if you have \( k \) regressors, then \( p \leq \frac{k(k+1)}{2} − 1 \).
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 4.4.

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.

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. This is especially true in some cases, where you know from the outset what the situation is. The 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:

\[
 y_i = \begin{cases} 
 1 & \text{with probability } \pi_i \\
 0 & \text{with probability } 1 - \pi_i 
\end{cases} 
\]  

(4.7)

The linearity hypothesis implies that \( E[y_i|x_i] = \pi_i = x'_i \beta \), since for a Bernoulli rv the expected value is, by construction, the probability of success. This is quite weird already, because \( \pi \) is a probability, and therefore has to be between 0 and 1, whereas if it really was a linear function of the \( x \) 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.

But then, you also have that for a Bernoulli rv \( V[y_i] = \pi_i \cdot (1 - \pi_i) \), and therefore

\[
 V[y_i|x_i] = \pi_i = x'_i \beta \cdot (1 - x'_i \beta_i) 
\]

so the conditional variance cannot be constant unless the conditional mean is constant too. The vector of parameters \( \beta \) enjoys a special nature, being the vector of parameters that determine both the conditional mean and the conditional variance. In theory, it is be possible to estimate \( \beta \) by an elaborate FGLS strategy, but in these cases practitioners always just use OLS with robust standard errors.

---

13In 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).

14Models 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.
4.A Assorted results

4.A.1 Proof of equation (4.2)

We would like to prove that the two models

$$y_i = \tilde{x}_i'\gamma + \sum_{j=1}^{m} (d_{ji}\hat{x}_i)'\gamma_j + \varepsilon_i$$

and

$$y_i = x_i'\beta + \sum_{j \neq q} (d_{ji}\hat{x}_i)'\gamma_j + \varepsilon_i$$

are equivalent. In order to do so, we will prove that the space spanned by the regressors of the two equations is the same. Without loss of generality, suppose that the observations are sorted so that all the cases for which $d_{1i} = 1$ are on top, and then come those for which $d_{2i} = 1$, and so on, so that the bottom rows of the dataset are the observations for which $d_{mi} = 1$.

The regressors matrix for the first equation can then be written as

$$Z = \begin{bmatrix}
\tilde{X}_1 & \hat{X}_1 & 0 & \cdots & 0 \\
\tilde{X}_2 & 0 & \hat{X}_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\tilde{X}_m & 0 & 0 & \cdots & \hat{X}_m
\end{bmatrix}$$

Suppose, again without loss of generality, that the reference category is the first one; now substitute the second block of columns by the sum of the blocks for 2 to the last. Call the resulting matrix $W$, where

$$W = \begin{bmatrix}
\tilde{X}_1 & \hat{X}_1 & 0 & \cdots & 0 \\
\tilde{X}_2 & \hat{X}_2 & \tilde{X}_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\tilde{X}_m & \hat{X}_m & 0 & \cdots & \hat{X}_m
\end{bmatrix} = \begin{bmatrix}
X_1 & 0 & \cdots & 0 \\
X_2 & X_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
X_m & 0 & \cdots & \hat{X}_m
\end{bmatrix}$$

where the second equality trivially come from $X_j = [\tilde{X}_j \hat{X}_j]$. It is clear that

- $W$ is the regressors matrix for the second equation, and
- you can compute $W$ from $Z$, but also $Z$ from $W$; therefore, $W$ can be written as $W = Z \cdot A$, where $A$ is invertible.

Therefore, $\text{Sp}(Z) = \text{Sp}(W)$ (see section 1.A.8), so $P_Z = P_W$ and the fitted values and the residuals from the two regressions are the same (see section 1.A.8).

For example, take the model

Subsample 1 \[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \]

Subsample 2 \[ y_i = \beta_0 + \beta_2 x_i + \varepsilon_i \]
with \( d_i = 0 \) in subsample 1 and \( d_i = 1 \) in subsample 2. In this case we have

\[
Z = \begin{bmatrix}
x_1 & 0 \\
x_2 & 0 & x_2
\end{bmatrix}
\]

Since \( Z \) has three columns, \( C \) should be \( 3 \times 3 \). The reader is invited to check that by the choice

\[
C = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 1
\end{bmatrix}
\]

you get

\[
W = \begin{bmatrix}
x_1 & 0 \\
x_1 & x_2
\end{bmatrix}
\]

### 4.A.2 Proof that full interactions are equivalent to split-sample estimation

Suppose you have \( m \) categories. If you interact your dummies with the full set of regressors \( x_i \), the resulting model is

\[
y_i = \sum_{j=1}^{m} (d_j x_i' \gamma_j) + \varepsilon_i
\]

which can be written in matrix notation as

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_m
\end{bmatrix} = \begin{bmatrix}
X_1 & 0 & \ldots & 0 \\
0 & X_2 & \ldots & 0 \\
0 & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & X_m
\end{bmatrix} \begin{bmatrix}
\gamma_1 \\
\gamma_2 \\
\vdots \\
\gamma_m
\end{bmatrix} + \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_m
\end{bmatrix}
\]
where \( y_j \) is the segment of the \( y \) vector containing the observations for the \( j \)-th subsample, and so forth. If you apply the OLS formula to equation (4.9), you get

\[
\begin{bmatrix}
\hat{\gamma}_1 \\
\hat{\gamma}_2 \\
\vdots \\
\hat{\gamma}_m
\end{bmatrix} = \begin{bmatrix}
x'_1 \qquad 0 \qquad \cdots \qquad 0 \\
0 \qquad x'_2 \qquad \cdots \qquad 0 \\
\vdots \\
0 \qquad 0 \qquad \cdots \qquad x'_m
\end{bmatrix} \begin{bmatrix}
x_1 \qquad 0 \qquad \cdots \qquad 0 \\
0 \qquad x_2 \qquad \cdots \qquad 0 \\
\vdots \\
0 \qquad 0 \qquad \cdots \qquad x_m
\end{bmatrix}
\]

\[
\times 
\begin{bmatrix}
x_1y_1 \\
x_2y_2 \\
\vdots \\
x_my_m
\end{bmatrix} = 
\begin{bmatrix}
x'_1x_1^{-1}x'_1y_1 \\
x'_2x_2^{-1}x'_2y_2 \\
\vdots \\
x'_mx_m^{-1}x'_my_m
\end{bmatrix}
\]

So clearly each \( \hat{\gamma}_j \) coefficient can be calculated by an OLS regression using the data for subsample \( j \) only. Therefore, the residuals for subsample \( j \) are \( e_j = y_j - X_j \hat{\gamma}_j \). As a consequence,

\[
e'e = \sum_{j=1}^{m} e_j'e_j,
\]

which in words reads: the SSR for model (4.8) is the same as the sum of the SSRs you get for the \( m \) separate submodels.

Now note that if you take \( q \) to be the “reference” category\(^{15}\), you can rewrite equation (4.8) as

\[
y_i = x'_i\beta + \sum_{j \neq q} (d_{ij}x'_j\beta_j) + \epsilon_i
\]

where \( \beta_j = (\gamma_j - \gamma_q) \) (see section 4.A.1). As a consequence, you can compare the model above with the model where all the \( \beta_j \) vectors are 0 by comparing the SSR for the restricted model \( y_i = x'_i\beta + \epsilon_i \) (call it \( e'e \)) against the sum of the SSRs of the \( m \) separate submodels (call them \( e_j'e_j \), with \( j = 1 \ldots m \)). A Wald-type statistic would be

\[
W = n \cdot \frac{e'e - \sum_j (e_j'e_j)}{\sum_j (e_j'e_j)}
\]

The classic Chow test is just the special case \( m = 2 \).

\(^{15}\)I will not offend the reader’s intelligence by writing the obvious double inequality \( 1 \leq q \leq m \).
CHAPTER 4. DIAGNOSTIC TESTING IN CROSS-SECTIONS

4.A.3 Proof that GLS is more efficient than OLS

In order to prove that \( V[\hat{\beta}] - V[\tilde{\beta}] \) is psd, I’ll use the properties on psd matrices that I listed in section 3.A.4, 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 \( A = HH' \);\(^{16}\)
3. all idempotent matrices are psd.

Therefore, to check the relative efficiency of \( \hat{\beta} \) and \( \tilde{\beta} \), we’ll perform an equivalent check on

\[ \Delta \equiv V[\tilde{\beta}]^{-1} - V[\tilde{\beta}]^{-1} \]

(by property 1 above). To prove that \( \Delta \) is psd, start from its definition:

\[ \Delta = X'(H')^{-1}H^{-1}X - (X'X)(X'H'X)^{-1}(X'X) \]

Now define \( W = H'X \) and re-express \( \Delta \) as:

\[ \Delta = (H^{-1}X)'[I - W(W'W)^{-1}W'](H^{-1}X) = (H^{-1}X)'M_W(H^{-1}X), \]

since \( M_W \) is idempotent, it is psd (property 3); but then, the same is true of \( (H^{-1}X)'M_W(H^{-1}X) \); therefore, the claim follows.

Note that under heteroskedasticity \( \Sigma \) is assumed to be diagonal, but the above proof holds for any non-singular covariance matrix \( \Sigma \).

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

\[ \hat{\theta} = T(X), \]

where \( X \) is a data matrix matrix with \( n \) rows, and the covariance matrix of \( \hat{\theta} \) is unknown, or difficult to compute, or perhaps your sample size is not large enough to trust the asymptotic approximation given by the CLT.

Of course, given your data \( X \) you can compute \( \hat{\theta} \) just once, but if you could observe \( H \) different datasets with the same distribution, then you could compute your estimator \( H \) times and have a look at how different \( \hat{\theta} \) is each time.

\(^{16}\)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.
The idea is to use your observed data $\mathbf{X}$ to generate, with the aid of computer-generated pseudo random numbers, $H$ alternative datasets $\mathbf{X}_h$, with $h = 1 \ldots H$, and compute your estimator for each of them, so you end up with a collection $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_H$. This procedure is what we call bootstrapping, and the $H$ realizations 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:

$$\hat{\theta} = \frac{1}{H} \sum_{h=1}^{H} \hat{\theta}_h$$

$$\tilde{V}(\hat{\theta}) = \frac{1}{H} \sum_{h=1}^{H} (\hat{\theta}_h - \hat{\theta})^2$$

How do you generate your artificial datasets $\mathbf{X}_h$? There is a myriad of ways you can do this, but when the observations are iid, 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. 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 initially, but I can assure you that’s the whole point.

---

17 According to Efron, “[i]t’s 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.

18 When data are not independent, things get a bit more involved.

19 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.
# 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 $X_h$ 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 $x_i$. 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 $E[y_t|x_t]$, 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 $y_1, \ldots, y_T$.

$^2$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 1970 and 2017 ($y$ and $c$, respectively). 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.

Therefore:

- the choice of $x_t$ as the conditioning set for $E[y_t|x_t]$ 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 $y_t - E[y_t|x_t]$ 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

$$y_t = x_t' \beta + \varepsilon_t$$

and working out solutions to deal with the fact that $E[\varepsilon_t \varepsilon_t'] = \Sigma$ 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, let us begin by defining what the *autocovariance* of a sequence of random variables is as the covariance of the sequence evaluated at two points:

$$\text{Cov}_t\{y_t, y_{t+k}\} = E[y_t y_{t+k}] - E[y_t] E[y_{t+k}] = \text{cov}_t\{y_t, y_{t+k}\}$$

for $k = 1, 2, \ldots$.

For example, if $y_t = y_{t-1} + \varepsilon_t$, we have

$$\text{Cov}_t\{y_t, y_{t+k}\} = \text{Cov}_t\{y_{t-1} + \varepsilon_t, y_{t+k}\} = \Sigma = \text{cov}_t\{y_t, y_{t+k}\}$$

where $\Sigma$ is the covariance matrix of $y_t$. In this case, the autocovariance is constant for all $k$, and the sequence is said to be *stationary*.

However, if $y_t = y_{t-1} + \phi y_{t-2} + \varepsilon_t$, the autocovariance depends on $k$ and the sequence is said to be *non-stationary*. The presence of non-zero entries outside the diagonal was commonly called the “autocorrelation” or “serial correlation” problem. In order to define this concept, let us begin by defining what the *autocovariance* of a sequence of random variables is as the covariance of the sequence evaluated at two points:

$$\text{Cov}_t\{y_t, y_{t+k}\} = E[y_t y_{t+k}] - E[y_t] E[y_{t+k}] = \text{cov}_t\{y_t, y_{t+k}\}$$

for $k = 1, 2, \ldots$.

For example, if $y_t = y_{t-1} + \varepsilon_t$, we have

$$\text{Cov}_t\{y_t, y_{t+k}\} = \text{Cov}_t\{y_{t-1} + \varepsilon_t, y_{t+k}\} = \Sigma = \text{cov}_t\{y_t, y_{t+k}\}$$

where $\Sigma$ is the covariance matrix of $y_t$. In this case, the autocovariance is constant for all $k$, and the sequence is said to be *stationary*.

However, if $y_t = y_{t-1} + \phi y_{t-2} + \varepsilon_t$, the autocovariance depends on $k$ and the sequence is said to be *non-stationary*. The presence of non-zero entries outside the diagonal was commonly called the “autocorrelation” or “serial correlation” problem. In order to define this concept, let us begin by defining what the *autocovariance* of a sequence of random variables is as the covariance of the sequence evaluated at two points:

$$\text{Cov}_t\{y_t, y_{t+k}\} = E[y_t y_{t+k}] - E[y_t] E[y_{t+k}] = \text{cov}_t\{y_t, y_{t+k}\}$$

for $k = 1, 2, \ldots$.

For example, if $y_t = y_{t-1} + \varepsilon_t$, we have

$$\text{Cov}_t\{y_t, y_{t+k}\} = \text{Cov}_t\{y_{t-1} + \varepsilon_t, y_{t+k}\} = \Sigma = \text{cov}_t\{y_t, y_{t+k}\}$$

where $\Sigma$ is the covariance matrix of $y_t$. In this case, the autocovariance is constant for all $k$, and the sequence is said to be *stationary*.

However, if $y_t = y_{t-1} + \phi y_{t-2} + \varepsilon_t$, the autocovariance depends on $k$ and the sequence is said to be *non-stationary*. The presence of non-zero entries outside the diagonal was commonly called the “autocorrelation” or “serial correlation” problem.
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

$$\rho_{t,s} = \frac{\text{Cov}[z_s, z_t]}{\sqrt{\text{V}[z_t]} \sqrt{\text{V}[z_s]}}$$

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.

Clearly, if the autocorrelation between $\varepsilon_t$ and $\varepsilon_s$ is nonzero for some $t$ and $s$, $\Sigma$ cannot be diagonal, so GLS solutions have been devised\footnote{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.}, 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 devising clever ways to perform acceptable inference on equation (5.1), we’d be much better off if we redefined our object of interest.

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\footnote{To be rigorous, we should define the information set by using a technical tool called a $\sigma$-field. This ensures that $\mathcal{I}_t$ contains all possible functions of the elements listed above ($\Delta y_{t-1}$, for example). But in an introductory treatment such as this, I’ll just use the reader’s intuition and use $\mathcal{I}_t$ as “all the things we know at time $t$”.}

$$\mathcal{I}_t = \{x_1, x_2, \ldots, x_t, y_1, y_2, \ldots, y_{t-1}\};$$

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 $\mathcal{I}_t$.

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 $x_t$, but also the past of both $y_t$ and $x_t$. 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 (e.g. $x_t = t$), seasonal dummy
variables (e.g., $x_t = 1$ 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) $\mathcal{I}_t$ is an element of a sequence where $\mathcal{I}_{t-1} \subseteq \mathcal{I}_t \subseteq \mathcal{I}_{t+1}$; in other words, the sequence of information sets is increasing.\(^7\)

Now consider the conditional expectation $E[y_t|\mathcal{I}_t]$; even under the linearity assumption, this object may have two potentially troublesome characteristics:

1. since the sequence $\mathcal{I}_t$ is increasing, $E[y_t|\mathcal{I}_t]$ may contain information that goes indefinitely back into the past, and

2. $E[y_t|\mathcal{I}_t]$ 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,

$$E[y_t|\mathcal{I}_t] = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{q} \beta_i' x_{t-i},$$

where $p$ and $q$ are finite numbers. Although in principle $\mathcal{I}_t$ contains all the past, no matter how remote, only the most recent elements of $\mathcal{I}_t$ 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 $\mathcal{I}_t$ (call it $\mathcal{F}_t$), that contains only recent information, such that conditioning on $\mathcal{I}_t$ or $\mathcal{F}_t$ makes no difference:

$$E[y_t|\mathcal{I}_t] = E[y_t|\mathcal{F}_t], \quad (5.2)$$

where $\mathcal{I}_t \supseteq \mathcal{F}_t$. In practice, $\mathcal{F}_t$ is the relevant information at time $t$.

The linearity assumption makes the regression function of $y_t$ on $\mathcal{I}_t$ 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 $x_t$;\(^8\) if we proceed in a similar way as in chapter 3, and define $\varepsilon_t = y_t - E[y_t|\mathcal{I}_t]$, we can write the so-called ADL model:

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{q} \beta_i' x_{t-i} + \varepsilon_t. \quad (5.3)$$

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

\(^7\)Or, if you will, we are assuming that we always learn and never forget.

\(^8\)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.
this is done, we may construct our \( y \) and \( X \) matrices as

\[
\begin{bmatrix}
y_{p+1} \\
y_{p+2} \\
y_{p+3} \\
\vdots 
\end{bmatrix} =
\begin{bmatrix}
y_p & y_{p-1} & \ldots & y_1 & x'_{p+1} & x'_{p} & \ldots & x'_{p-q+1} \\
y_{p+1} & y_p & \ldots & y_2 & x'_{p+2} & x'_{p+1} & \ldots & x'_{p-q+2} \\
y_{p+2} & y_{p+1} & \ldots & y_3 & x'_{p+3} & x'_{p+2} & \ldots & x'_{p-q+3} \\
\vdots & \vspace{1pt} & \vspace{1pt} & \vspace{1pt} & \vdots & \vspace{1pt} & \vspace{1pt} & \vspace{1pt} & \vdots 
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_p \\
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_q 
\end{bmatrix} +
\begin{bmatrix}
e_{p+1} \\
e_{p+2} \\
\vdots \\
e_{p+q} 
\end{bmatrix}
\]

\[ y = W\gamma + \varepsilon \]

where \( w_t \) is defined as

\[ w'_t = [y_{t-1}, y_{t-2}, \ldots, y_{t-p}, x'_t, x'_{t-1}, \ldots, x'_{t-q}] , \]

and

\[ \gamma' = [\alpha_1, \alpha_2, \ldots, \alpha_p, \beta'_0, \beta'_1, \ldots, \beta'_q] . \]

Given this setup, clearly the OLS statistic can be readily computed with the usual formula \((W'W)^{-1}W'y\), but given the nature of the conditioning, one may wonder if OLS is a CAN estimator of the \( \alpha \) and \( \beta \) 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 \( E[y_t \mid \mathcal{I}_t] \) were known and didn’t have to be estimated, \textit{how do we interpret them?}

5.2 Manipulating difference equations

Given the difference equation

\[ y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{q} \beta_i x_{t-i} \]

we may ask ourselves: what is the effect of \( x \) on \( y \) after a given period? That is: how does \( x_t \) affect \( y_{t+h} \)? Since the coefficients \( \alpha_i \) and \( \beta_i \) 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 \( x_{t-h} \)? Clearly, if \( h = 0 \) we have a quantity that is straightforward to interpret, that is the instantaneous impact if \( x_t \) in \( y_t \), but much is to be gained by considering magnitudes like

\[ d_h = \frac{\partial y_t}{\partial x_{t-h}} = \frac{\partial y_{t+h}}{\partial x_t} , \]

\[ h \geq 0 \]

\[ \text{(5.4)} \]

1 assumed for simplicity that \( p \geq q \); of course, potentially collinear deterministic terms would have to be dropped.
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, 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 5.1**

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:

$$b_t = Lb_{t-1} + s_t \rightarrow b_t - Lb_t = (1 - L)b_t = \Delta b_t = s_t$$

The $\Delta$ 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 5.2**

Call $q_t$ the GDP for the Kingdom of Verduria in quarter $t$. Obviously, yearly GDP is given by

$$y_t = q_t + q_{t-1} + q_{t-2} + q_{t-3} = (1 + L + L^2 + L^3)q_t$$

\(^{10}\)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.
5.2. MANIPULATING DIFFERENCE EQUATIONS

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

\[
P(L)\mu = \sum_{j=0}^{n} p_j \mu = \mu \sum_{j=0}^{n} p_j = P(1)\mu.
\]

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:

\[
\frac{1}{P(x)} \text{ exists iff } P(x) = 0 \Rightarrow |x| > 1. \tag{5.6}
\]

The proofs are in subsection 5.A.1.

Example 5.3 (The Keynesian multiplier) \(^{11}\)

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 \(Y_t\) id GDP, \(C_t\) is aggregate consumption and \(I_t\) is investment; \(0 < \alpha < 1\) is the marginal propensity to consume.

\(^{11}\)To be precise, we should say: ‘if you apply the \((1 - L)\) operator to the expression above’. 

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 \( \alpha \), by force of equation (5.8), so at time 2 it increases by \( \alpha^2 \) and so on. Since \( 0 < \alpha < 1 \), the effect dies out eventually.

If investments were constant through time, then \( I_t = \bar{I} \) and the expression above becomes the familiar “Keynesian multiplier” formula:
\[ Y_t = \bar{I} \sum_{i=0}^{\infty} \alpha^i = \frac{\bar{I}}{1 - \alpha}. \]

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 5.4
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
\[ z_{t-1} = x_{t-1} y_{t-1} = Lx_t L y_t = L^2 x_t y_t = L^2 z_t = z_{t-2} \]
which is obviously absurd.

5.2.2 Dynamic multipliers
Manipulating polynomials in \( L \) is the technical tool we need in order to handle the problem that we are ultimately after. 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 \textit{stable}. In this case, we may define \( D(L) = A(L)^{-1}B(L) = B(L)/A(L) \); in general, \( D(L) \) may not be of finite order, so

\[
y_t = D(L)x_t = \sum_{i=0}^{\infty} d_i x_{t-i}
\]

This is all we need for dealing with the issue we started from, that is: how do we interpret the parameters of a dynamic model? The answer is easy 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 \( \beta_0 \) (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. Let me show you a practical example. For an ADL(1,1) model,

\[
y_t = \alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1},
\]

use the definition of a multiplier as a derivative and write

\[
\begin{align*}
d_0 &= \frac{\partial y_t}{\partial x_t} = \frac{\partial}{\partial x_t} (\alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1}) = \beta_0 \\
d_1 &= \frac{\partial y_t}{\partial x_{t-1}} = \frac{\partial}{\partial x_{t-1}} (\alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1}) = a \frac{\partial y_{t-1}}{\partial x_{t-1}} + \beta_1 = a d_0 + \beta_1,
\end{align*}
\]

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,

\[
d_2 = \frac{\partial y_t}{\partial x_{t-2}} = \frac{\partial}{\partial x_{t-2}} (\alpha y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1}) = a \frac{\partial y_{t-1}}{\partial x_{t-2}} = a d_1
\]

and so on, and on, and on, . . .

\(^{12}\)As we will see in section 5.3, invertibility of \( A(L) \) is not only required for the calculation of the multipliers, but also for the CAN property of OLS.
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

$$A(L)d_i = B(L)u_i,$$  \(5.10\)

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 5.5** (Inversion of polynomials)

Take for example the following difference equation:

$$y_t = 0.2y_{t-1} + 0.4x_t + 0.3x_{t-2}.$$  

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

$$\frac{B(L)}{A(L)} = 0.4 \times (1 + 0.2L + 0.04L^2 + 0.008L^3 + \cdots) + 0.3L^2 \times (1 + 0.2L + 0.04L^2 + 0.008L^3 + \cdots) = 0.4 + 0.08L + 0.016L^2 + 0.0032L^3 + \cdots + 0.3L^2 + 0.06L^3 + 0.012L^4 + 0.0024L^5 \cdots = 0.4 + 0.08L + 0.316L^2 + 0.0632L^3 + \cdots$$

but it’s really boring. The recursive way of calculating the coefficient is much quicker:

$$d_0 = B(0)/A(0) = 0.4/1 = 0.4$$
$$d_1 = 0.2 \cdot d_0 = 0.08$$
$$d_2 = 0.2 \cdot d_1 + 0.03 = 0.016 + 0.3 = 0.316$$
$$d_3 = 0.2 \cdot d_2 = 0.0632$$

and so on.
5.2. MANIPULATING DIFFERENCE EQUATIONS

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
\[ c_j = d_0 + d_1 + \cdots + d_j = \sum_{i=0}^{j} d_i. \quad (5.11) \]
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 5.6 (interim multipliers)**

Let’s go back to the difference equation we used in example 5.5:
\[ y_t = 0.2 y_{t-1} + 0.4 x_t + 0.3 x_{t-2}. \]

Interim multipliers are easily computed:
\[
\begin{align*}
c_0 &= d_0 = 0.4 \\
c_1 &= d_0 + d_1 = c_0 + d_1 = 0.48 \\
c_2 &= d_0 + d_1 + d_2 = c_1 + d_2 = 0.796
\end{align*}
\]
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 tells us what the proportion is between \( y_t \) and \( x_t \) in steady state: imagine 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.

Imagine that the system is in steady state: both variables are time-invariant, so $y_t = Y$ and $x_t = X$; as a consequence,

$$A(L)\ y_t = B(L)\ x_t \Rightarrow A(L)\ Y = B(L)\ X \Rightarrow A(1)\ Y = B(1)\ X \Rightarrow Y = \frac{B(1)}{A(1)}\ X = cX$$

so 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)

$$A(L)\ y_t = B(L)\ x_t + \varepsilon_t$$

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 $X$ matrix and the $y$ 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 $w_t$ like at the end of section 5.1, as

$$w_t' = [y_{t-1}, y_{t-2}, \ldots, y_{t-p}, x_{t-1}', x_{t-2}', \ldots, x_{t-q}]$$

so that we can write our dynamic model as

$$y_t = w_t' \gamma + \varepsilon_t$$

where of course $\gamma' = [\alpha_1, \alpha_2, \ldots, \alpha_p, \beta_0', \beta_1', \ldots, \beta_q']$.

The first important requirement is that the second moments of $w_t$ exist and that

$$T^{-1} \sum_{t=1}^{T} w_t w_t' \xrightarrow{p} Q$$

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.\textsuperscript{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 (5.3) is that $\epsilon_t$ be a martingale difference sequence (or MDS for short).\textsuperscript{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 $\mathcal{F}_{t-1}$ is a sequence of random variables $X_t$ such that $E[X_t|\mathcal{F}_{t-1}] = X_{t-1}$;
- If $X_t$ is a martingale, then $\Delta X_t$ is a MDS: $E[\Delta X_t|\mathcal{F}_{t-1}] = 0$

Of course, if we could condition $y_t$ on $\mathcal{F}_t$ then $\epsilon_t$ would be a MDS by construction:

$$E[\epsilon_t|\mathcal{F}_t] = E[y_t - E[y_t|\mathcal{F}_t]|\mathcal{F}_t] = E[y_t|\mathcal{F}_t] - E[y_t] = 0$$

(this is essentially the same argument we used in section 3.1). But of course we can’t use $\mathcal{F}_t$ in practice; however, we’re assuming (see equation 5.2) that there exists a subset $\mathcal{F}_t \subset \mathcal{F}_t$ such that $E[y_t|\mathcal{F}_t] = E[y_t|\mathcal{F}_t]$, so $\mathcal{F}_t$ (which is usable, because it’s finite) is just as good. So if you condition $y_t$ on $\mathcal{F}_t$, the quantity $\epsilon_t = y_t - E[y_t|\mathcal{F}_t]$ is a MDS and all is well.

However, what happens if you use a conditioning set $\mathcal{G}_t$ that is “too small”? That is, that doesn’t contain $\mathcal{F}_t$? In that case, the difference $u_t = y_t - E[y_t|\mathcal{G}_t]$ is not a MDS with respect to $\mathcal{F}_t$: if $E[y_t|\mathcal{G}_t] \neq E[y_t|\mathcal{F}_t]$, then

$$E[u_t|\mathcal{F}_t] = E[y_t - E[y_t|\mathcal{G}_t]|\mathcal{F}_t] = E[y_t|\mathcal{F}_t] - E[y_t|\mathcal{G}_t] \neq 0.$$ 

On the contrary, it is easy to prove that in the opposite case, when you condition on a subset of $\mathcal{F}_t$ that is larger than $\mathcal{F}_t$, no problems arise.

This remark is extremely important in practice because the order of the polynomials $A(L)$ and $B(L)$ (or $p$ 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 $\mathcal{F}_t$, 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 $\mathcal{F}_t$ 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.

\textsuperscript{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).

\textsuperscript{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 context, 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, $\varepsilon_t$ is a MDS, if we pick $p$ and $q$ large enough. The obvious implication is that $E[\varepsilon_t | w_t] = 0$ (since $w_t$ is contained in $\mathcal{Z}_{t-1}$). If we also add a homoskedasticity assumption $E[\varepsilon_t^2 | w_t] = \sigma^2$, 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 $E[w_t \cdot \varepsilon_t | \mathcal{I}_t] = 0$, and therefore $\hat{\gamma} \xrightarrow{p} \gamma$; additionally,

$$
\sqrt{T}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, \sigma^2 Q^{-1})
$$

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 $\gamma$, we can simply compute the multipliers from the estimated parameters $\hat{\gamma}$ and get automatically CAN estimators of the multipliers.\footnote{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.}

The homoskedasticity assumption is not normally a problem, except for financial data at high frequencies (e.g., 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 $\varepsilon_t$ 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, $\varepsilon_t$ 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: 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 $\varepsilon_t$ 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.

\footnote{If I were insufferably pedantic, I would say “a MDS with finite second moments”.

15\footnote{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.}
(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:

\[ A(L)y_t = B(L)x_t + \gamma_1 e_{t-1} + \gamma_2 e_{t-2} + \cdots + \gamma_h e_{t-h} + \epsilon_t \]

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.\(^\text{17}\) 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 3.6.1) and is asymptotically distributed (under \( H_0 \)) as \( \chi^2_h \). In practice, you carry out an auxiliary regression of the OLS residuals \( e_t \) against \( w_t \) 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\(^\text{18}\) at each step to make sure your reduction was not too aggressive.

\(^{17}\)“Mi ha detto mio cugino 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 DGP at time $t$. If we did, we’d obtain a “static model”

$$c_t = \beta_0 + \beta_1 y_t + \epsilon_t,$$

whose output is in Table 5.1.

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>-0.128464</td>
<td>0.0533021</td>
<td>-2.410</td>
</tr>
<tr>
<td>$y$</td>
<td>0.968085</td>
<td>0.00377749</td>
<td>256.3</td>
</tr>
</tbody>
</table>

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 $\hat{\rho} = 0.944$ 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 $\epsilon_t$ 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(1,3) model. In practice:

$$c_t \approx k + \alpha c_{t-1} + \beta_0 y_t + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3};$$

Table 5.2 contains the OLS estimates, that is $\hat{\alpha} = 0.956$, $\hat{\beta}_0 = 0.609$, and so on.

Note that 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.
Example 5.7 (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

\[ A(L) = 1 - 0.956403L \]
\[ B(L) = 0.609142 - 0.746166L + 0.298760L^2 - 0.124319L^3, \]

so in this case we have

\[ d_i = 0.956403d_{i-1} + 0.609142u_i - 0.746166u_{i-1} + 0.298760u_{i-2} - 0.124319u_{i-3}, \]

where \( u_0 = 1 \) and \( u_i = 0 \) for \( i \neq 0 \). Therefore, \( d_0 \) equals

\[ d_0 = 0.609142 \]

while for \( d_1 \) and \( d_2 \) we have

\[ d_1 = 0.956403 \cdot d_0 - 0.746166 = -0.16358 \]
\[ d_2 = 0.956403 \cdot d_1 + 0.298760 = 0.14231, \]

and so on. With a little effort (and appropriate software), you get the following results:

Table 5.2: Dynamic regression example
Where I added a column for the interim (cumulated) multipliers. Moreover, you have that $A(1) = 1 - 0.956403 = 0.043597$, $B(1) = 0.037416$, and therefore the long-run multiplier equals $c = 0.037416/0.043597 = 0.85823$.

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 express a difference equation in a slightly modified form, so that certain quantities appear more clearly. In fact, one can also use the word "reparametrisation" for the ECM, 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. David Hendry, who is considered the father of ECM (or at least, one of the fathers) is adamant on Equilibrium Correction Mechanism. 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 that.

To illustrate how the ECM representation works, let’s start from the simple
5.5. THE ECM REPRESENTATION

case of an ADL(1,1) (here $x_t$ is a vector):

$$y_t = \alpha y_{t-1} + \beta'_0 x_t + \beta'_1 x_{t-1};$$

from the definition of the $\Delta$ operator, evidently $y_t = y_{t-1} + \Delta y_t$ and $x_t = x_{t-1} + \Delta x_t$. After substitution,

$$\Delta y_t = (\alpha - 1) y_{t-1} + \beta'_0 \Delta x_t + (\beta_0 + \beta_1)' x_{t-1}$$

which, after rearranging terms, yields

$$\Delta y_t = \beta'_0 \Delta x_t + (\alpha - 1) \left[ y_{t-1} - \frac{(\beta_0 + \beta_1)'}{1 - \alpha} x_{t-1} \right] \tag{5.12}$$

Which means: the time variation in $y_t$ (on the left-hand side) may come from variation in $x_t$, with response $\beta_0$ (the impact multiplier); however, even if $\Delta x_t = 0$ there may be some variation in $y_t$ if the term in square brackets is non-zero. This term can also be written as

$$y_{t-1} - c' x_{t-1}$$

where $c = \frac{\beta_0 + \beta_1}{1 - \alpha}$, 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 $x_{t-1}$) 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 $\Delta y_t$ 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 $\alpha$ 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

$$A(L) y_t = m_t + B(L) x_t + \varepsilon_t$$

where the order of $A(L)$ is $p$ and the order of $B(L)$ is $q$, then the ECM representation is

$$P(L) \Delta y_t = m_t + Q(L) \Delta x_t - A(1) y_{t-1} + B(1) x_{t-1} + \varepsilon_t,$$

where the orders of $P(L)$ and $Q(L)$ are $q - 1$ and $p - 1$, respectively.

For example:

**Example 5.8 (ECM Representation)**

Let’s take another look at the difference equation I used in example 5.6:

$$y_t = 0.2 y_{t-1} + 0.4 x_t + 0.3 x_{t-2}$$
and compute the ECM representation. The quickest way to do this is to re-express all the terms relative to time \( t - 1 \):

\[
\begin{align*}
y_t &= y_{t-1} + \Delta y_t \\
x_t &= x_{t-1} + \Delta x_t \\
x_{t-2} &= x_{t-1} - \Delta x_{t-1};
\end{align*}
\]

now substitute

\[
y_{t-1} + \Delta y_t = 0.2 y_{t-1} + 0.4(x_{t-1} + \Delta x_t) + 0.3(x_{t-1} - \Delta x_{t-1})
\]

and collect

\[
\Delta y_t = -0.8 y_{t-1} + 0.7 x_{t-1} + 0.4 \Delta x_t - 0.3 \Delta x_{t-1}
\]

so finally

\[
\Delta y_t = 0.4 \Delta x_t - 0.3 \Delta x_{t-1} - 0.8 \left[ y_{t-1} - 0.875 x_{t-1} \right];
\]

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. It isn’t hard, is it?

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.

**Example 5.9 (ECM on real data)**

The ECM representation of the model shown in table 5.2 is easily computed after performing the following substitutions:

\[
\begin{align*}
c_t &= c_{t-1} + \Delta c_t \\
y_t &= y_{t-1} + \Delta y_t \\
y_{t-2} &= y_{t-1} - \Delta y_{t-1} \\
y_{t-3} &= y_{t-1} - \Delta y_{t-1} - \Delta y_{t-2}.
\end{align*}
\]

Hence,

\[
\Delta c_t = k + (\alpha - 1)c_{t-1} + \beta_0 \Delta y_t + \left( \sum_{i=0}^{3} \beta_i \right) y_{t-1} - (\beta_2 + \beta_3) \Delta y_{t-1} - \beta_3 \Delta y_{t-2} + \varepsilon_t;
\]

that is

\[
\Delta c_t = k + \beta_0 \Delta y_t - A(1) \left[ c_{t-1} - c y_{t-1} \right] - (\beta_2 + \beta_3) \Delta y_{t-1} - \beta_3 \Delta y_{t-2} + \varepsilon_t;
\]
5.6. HYPOTHESIS TESTS ON THE LONG-RUN MULTIPLIER

ecm: OLS, using observations 1970:4-2016:4 (T = 185)
Dependent variable: d_c

<table>
<thead>
<tr>
<th></th>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>0.064</td>
<td>0.014</td>
<td>4.293</td>
<td>2.88e-05 ***</td>
</tr>
<tr>
<td>d_y</td>
<td>0.609</td>
<td>0.050</td>
<td>12.13</td>
<td>4.36e-25 ***</td>
</tr>
<tr>
<td>d_y_1</td>
<td>-0.174</td>
<td>0.052</td>
<td>-3.324</td>
<td>0.0011   ***</td>
</tr>
<tr>
<td>d_y_2</td>
<td>0.124</td>
<td>0.049</td>
<td>2.489</td>
<td>0.0137   **</td>
</tr>
<tr>
<td>c_1</td>
<td>-0.044</td>
<td>0.019</td>
<td>-2.205</td>
<td>0.0287   **</td>
</tr>
<tr>
<td>y_1</td>
<td>0.037</td>
<td>0.019</td>
<td>1.959</td>
<td>0.0517   *</td>
</tr>
</tbody>
</table>

Mean dependent var 0.005008 S.D. dependent var 0.005592
Sum squared resid 0.002388 S.E. of regression 0.003653
R-squared 0.584832 Adjusted R-squared 0.573235
F(5, 179) 50.43006 P-value(F) 1.99e-32
Log-likelihood 778.8122 Akaike criterion -1545.624
Schwarz criterion -1526.302 Hannan-Quinn -1537.794

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 5.3: Dynamic regression in ECM form</td>
<td></td>
</tr>
</tbody>
</table>

so, after substituting the estimated numerical values (and rounding results a little),

\[
\Delta c_t = 0.064 + 0.609 \Delta y_t - 0.174 \Delta y_{t-1} + 0.124 \Delta y_{t-2} + \
-0.044 \left[ c_{t-1} - 0.858 y_{t-1} \right] + \epsilon_t.
\]

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 0.002388); 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

\[
\hat{c} = \frac{\hat{B}(1)}{\hat{A}(1)}
\]
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, so the delta method (see section 2.3.2, particularly equation (2.13)) would be required. A much simpler way comes from observing that

$$c = k \iff B(1) - k \cdot A(1) = 0$$

which is a linear test and, as such, falls under the $R\beta = d$ jurisdiction.

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 $-\hat{A}(1)$ (note the minus sign) and the parameter for $x_{t-1}$ is $\hat{B}(1)$.

**Example 5.10**

Suppose that we have the following estimates:

$$\hat{y}_t = 0.75y_{t-1} + 0.53x_t - 0.24x_{t-1}$$

with the following covariance matrix:

$$\hat{V}_{ADL} = 0.001 \times \begin{bmatrix} 5 & 0.5 & -2 \\ 0.5 & 5 & 4 \\ -2 & 4 & 5 \end{bmatrix}$$

The hypothesis $c = 1$ implies $\alpha + \beta_0 + \beta_1 = 1$. Therefore, a Wald test can be set up with $R = [1 \ 1 \ 1]$ and $d = 1$ (see section 3.3.2 for details). Therefore

$$R\hat{\beta} - d = \begin{bmatrix} 0.53 \\ -0.24 \end{bmatrix} - 1 = 0.04$$

$$R \cdot \hat{V}_{ADL} \cdot R' = 0.001 \times 20 = 0.02$$

$$W = \frac{0.04^2}{0.02} = 0.08$$

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:

$$\Delta y_t = 0.53\Delta x_t - 0.25y_{t-1} + 0.29x_{t-1}$$

$^{19}$You have $\hat{A}(1)$ in the denominator, so for example in an ADL(1,1) $c = \frac{\hat{B}_0 + \hat{B}_1}{1 - \hat{A}(1)}$, and the Jacobian term would be $f = \frac{1}{1 - \hat{A}(1)} [1 \ 1 \ c]$. 


with the associated covariance matrix

\[ \hat{V}_{ECM} = 0.001 \times \begin{bmatrix} 5 & 0.5 & 9 \\ 5 & -1.5 & 18 \\ 18 & 0.29 & \end{bmatrix} \]

In this case the hypothesis can be written as \( H_0 : B(1) - A(1) = 0 \), so for the ECM form

\[ R\hat{\beta} - d = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.53 \\ -0.25 \\ 0.29 \end{bmatrix} = 0.04 \]

\[ R \cdot \hat{V}_{ECM} \cdot R' = 0.001 \times 20 = 0.02 \]

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

\[ y_{T+1} = \alpha y_T + \beta_0 x_{T+1} + \beta_1 x_T + \epsilon_{T+1}; \quad (5.13) \]

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 \( x_{T+1} = \tilde{x}_{T+1} \). Therefore, since \( \epsilon_t \) is a martingale difference sequence, its conditional expectation with respect to \( \mathcal{F}_{T+1} \) is 0, so

\[ E[y_{T+1}|\mathcal{F}_T] = \alpha y_T + \beta_0 \tilde{x}_{T+1} + \beta_1 x_T \]

\[ V[y_{T+1}|\mathcal{F}_T] = \sigma^2 \]

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

\[ \hat{y}_{T+1} = \hat{\alpha} y_T + \hat{\beta}_0 \tilde{x}_{T+1} + \hat{\beta}_1 x_T \]

and a 95% confidence interval can be constructed as

\[ \hat{y}_{T+1} \pm 1.96 \times \hat{\sigma} \]

---

\(^{20}\) We’re sticking to the definition of \( \mathcal{F}_{T+1} \) introduced in Section 5, so \( \mathcal{F}_{T+1} \) includes \( x_{T+1} \) but not \( y_{T+1} \); later in this section, we’ll use a different convention.
where it is implicitly assumed that $\varepsilon_t$ 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 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 (5.13) 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

$$\hat{y}_{T+2} = \hat{\alpha} \hat{y}_{T+1} + \hat{\beta}_0 \hat{x}_{T+2} + \hat{\beta}_1 \hat{x}_{T+1},$$

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

$$V[\hat{y}_{T+k}] = \frac{1 - (\alpha^2)^k}{1 - \alpha^2} \sigma^2.$$ 

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 $\beta_0$ and $\beta_1$ coefficients were 0 in equation (5.13). 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 something that we may want to construct a hypothesis test for.

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 \textbf{Granger-causality} test, after the great Clive Granger, Nobel Prize winner in 2003.\footnote{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 \textit{cointegration}, that I will skip in this book, but is absolutely indispensable if you want to engage in applied macroeconomics.} 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 \textit{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 inference on an information set that collects everything that is known at time \( t - 1 \), namely

\[ \Im^*_{t-1} = \{ y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots \} ; \]

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 \( x_{t+1} \); in practice, it collects all information on \( y_t \) and \( x_t \) that is available up to time \( t \). Forecasting, therefore, amounts to finding

\[ \hat{y}_{T+1|T} = E[ y_{T+1} | \Im^*_T] . \]

The subscript “\( T + 1 | T \)” is customarily read as “at time \( T + 1 \), based on the information available at time \( T \).”

As a consequence, our ADL model

\[ A(L) y_t = B(L) x_t + \epsilon_t \]

will not include \( x_t \), but only its lags:

\[ y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \beta_1' x_{t-1} + \beta_2' x_{t-2} + \cdots + \epsilon^*_t \]

where \( \epsilon^*_t \) is defined as \( y_t - E[ y_{T+1} | \Im^*_T] \). 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

\[ H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0 \]

which is of course a system of linear restrictions, that we can handle just fine via the \( R\beta = d \) 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 those days, the concept of Granger-causality seemed to provide a convincing answer.\(^{22}\)

\[^{22}\text{Readers who are into the history of economics and econometrics might want to take a look at Sims (1972).}\]
After taking logs, we estimate by OLS the following 2 equations:

\[ c_t = m_1 + \alpha_1 c_{t-1} + \alpha_2 c_{t-2} + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \epsilon_t \]  
(5.14)

\[ e_t = \mu_1 + \gamma_1 c_{t-1} + \gamma_2 c_{t-2} + \lambda_1 e_{t-1} + \lambda_2 e_{t-2} + \eta_t \]  
(5.15)

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 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
## 5.7. FORECASTING AND GRANGER CAUSALITY

Table 5.4: Granger causality tests between chickens and eggs

### Dependent variable: l_chicken

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>2.1437</td>
<td>0.7715</td>
<td>2.7788</td>
</tr>
<tr>
<td>l_chicken_1</td>
<td>0.4037</td>
<td>0.1389</td>
<td>2.9054</td>
</tr>
<tr>
<td>l_chicken_2</td>
<td>0.4362</td>
<td>0.1320</td>
<td>3.3037</td>
</tr>
<tr>
<td>l_egg_1</td>
<td>0.8627</td>
<td>0.2011</td>
<td>4.2906</td>
</tr>
<tr>
<td>l_egg_2</td>
<td>-0.8724</td>
<td>0.1999</td>
<td>-4.3642</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>0.8816</td>
<td>0.5292</td>
<td>1.6660</td>
</tr>
<tr>
<td>l_chicken_1</td>
<td>-0.1196</td>
<td>0.0953</td>
<td>-1.2548</td>
</tr>
<tr>
<td>l_chicken_2</td>
<td>0.0695</td>
<td>0.0906</td>
<td>0.7673</td>
</tr>
<tr>
<td>l_egg_1</td>
<td>1.5302</td>
<td>0.1379</td>
<td>11.0961</td>
</tr>
<tr>
<td>l_egg_2</td>
<td>-0.5570</td>
<td>0.1371</td>
<td>-4.0624</td>
</tr>
</tbody>
</table>

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: \( F(2, 68) = 1.28907, \ p-value = 0.282174 \)
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: people normally do their Christmas shopping before Christmas, so it would be legitimate to say that gift purchases Granger-cause Christmas. 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).

5.A Assorted results

5.A.1 Inverting polynomials

Let us begin by noting that, for any $a \neq 1$,

$$\sum_{i=0}^{n} a^i = \frac{1-a^{n+1}}{1-a}, \quad (5.16)$$

which is easy to prove: call

$$S = \sum_{i=0}^{n} a^i = 1 + a + a^2 + \cdots + a^n; \quad (5.17)$$

of course

$$a \cdot S = a + a^2 + \cdots + a^{n+1} \quad (5.18)$$

and therefore, by subtracting (5.18) from (5.17), $S(1-a) = 1-a^{n+1}$, and hence equation (5.16).

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 $|a| < 1$, 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,

$$\frac{1}{1-\alpha L} = \sum_{i=0}^{\infty} a^i L^i$$

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:

$$P(x) = \prod_{j=1}^{n} \left( 1 - \frac{1}{\lambda_j} x \right) \quad (5.19)$$

\footnote{If you don’t believe me, google for “Fundamental theorem of algebra.”}
where the numbers $\lambda_j$ 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 $\lambda_j$ 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 5.12**

*Consider the polynomial* $A(x) = 1 - 1.2x + 0.32x^2$; *is it invertible? Let’s check its roots:*

$$A(x) = 0 \iff x = \frac{1.2 \pm \sqrt{1.44 - 1.28}}{0.64} = (1.2 \pm 0.4)/0.64$$

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*

$$\frac{1}{A(x)} = (1 - 0.4x)^{-1}(1 - 0.8x)^{-1} = (1 + 0.4x + 0.16x^2 + \cdots)(1 + 0.8x + 0.64x^2 + \cdots) = 1 + 1.2x + 1.12x^2 + 0.96x^3 + 0.7936x^4 + \cdots$$

---

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; if you want the real thing, go for Brockwell and Davis (1991).

Suppose you have an infinitely long sequence of random variables

$$\ldots, x_{t-1}, x_t, x_{t+1}, \ldots$$

\(^{24}\)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 $|z| = \sqrt{a^2 + b^2}$. 
where the index \( t \) is normally taken to mean “time” (although in fact it needn’t be). 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).

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:

\[
\gamma_h = \text{Cov}[x_t, x_{t+h}]
\]

note that \( \gamma_h \), the autocovariance of order \( h \), is a function of \( h \) only, not of \( t \); of course, \( \gamma_0 \) is just \( \text{V}[x_t] \). 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 number of children born in the Kingdom of Verduria 400 years ago should contain no useful information on the number of children born in the Kingdom of Verduria today. 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

\[
\sum_{i=0}^{\infty} |\gamma_i| = M < \infty
\]

\(^{25}\)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 \( \omega \in \Omega \).
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:

\[
\frac{1}{T} \sum_{t=1}^{T} x_t \xrightarrow{p} E[x_t].
\]

Note that the above expression can be considered as one of the many versions of the Law of Large Numbers.

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).

5.A.3 Why martingale difference sequences are serially uncorrelated

Here’s a rapid proof: if \( \epsilon_t \) is a MDS with respect to \( \mathcal{F}_t \), then

\[
E[\epsilon_t | \mathcal{F}_t] = E[\epsilon_t | x_t, x_{t-1}, \ldots, y_{t-1}, y_{t-2}, \ldots] = 0.
\]

Now, observe that \( \epsilon_{t-1} \) is defined as \( \epsilon_{t-1} = y_{t-1} - E[y_{t-1} | \mathcal{F}_{t-1}] \); since both elements of the right-hand side of the equation are contained in \( \mathcal{F}_t \), then \( \epsilon_{t-1} \in \mathcal{F}_t \); moreover, \( \mathcal{F}_t \supseteq \mathcal{F}_{t-1} \supseteq \mathcal{F}_{t-2} \ldots \), so clearly all lags of \( \epsilon_t \) are all contained in \( \mathcal{F}_t \) (see footnote 6 in this chapter). As a consequence, we can use the law of iterated expectations (2.7) as follows:

\[
\text{Cov} [\epsilon_t, \epsilon_{t-k}] = E[\epsilon_t \cdot \epsilon_{t-k}] = E[E[\epsilon_t \cdot \epsilon_{t-k}] | \mathcal{F}_t] = E[E[\epsilon_t | \mathcal{F}_t] \cdot \epsilon_{t-k}] = E[0 \cdot \epsilon_{t-k}] = 0
\]

A second argument, perhaps more intuitive, rests directly on the definition of a MDS: if \( \epsilon_t \) is a MDS with respect to \( \mathcal{F}_t \), then its expectation conditional on \( \mathcal{F}_{t-k} \) (for \( k > 0 \)) must also be 0, because \( \mathcal{F}_{t-k} \) is a subset of \( \mathcal{F}_t \). But that means that the expectation of any future element \( \epsilon_{t+k} \) conditional on the present information set \( \mathcal{F}_t \) is 0. In formulae:

\[
\begin{align*}
\mathcal{F}_{t-k} \subseteq \mathcal{F}_t & \text{ for } k > 0 \\
\implies E[\epsilon_t | \mathcal{F}_{t-k}] = E[\epsilon_{t+k} | \mathcal{F}_t] = 0
\end{align*}
\]

This is tantamount to saying that \( \epsilon_t \) is effectively unpredictable. But then, if \( \text{Cov} [\epsilon_t, \epsilon_{t-k}] \neq 0 \), \( \epsilon_t \) 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 \( k \).
CHAPTER 5. DYNAMIC MODELS

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 \( P(x) = 1 - 1.8x + 0.4x^2 \) 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 \( a = 1 \):

\[
\begin{align*}
P(L) & = P(0) - Q(L) \cdot L \\
Q(L) & = Q(1) + P^*(L)(1 - L)
\end{align*}
\]

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 (5.20) in \( L = 1 \), you have \( P(1) = P(0) - Q(1) \), so that equation (5.21) becomes

\[ Q(L) = P(0) - P(1) + P^*(L)(1 - L) \]

and therefore, using equation (5.20) again,

\[ P(L) = P(0) - [P(0) - P(1) + P^*(L)(1 - L)] \cdot L = P(0)\Delta + P(1)L - P^*(L)\Delta \cdot L. \]

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 \( A(L) y_t = B(L)x_t + \epsilon_t \):

\[ \Delta y_t + A(1)y_{t-1} - A^*(L)\Delta y_{t-1} = B(0)\Delta x_t + B^*(L)\Delta x_{t-1} + B(1)x_{t-1} + \epsilon_t; \]

(note that \( A(0) = 1 \) by construction). After rearranging terms, you obtain the ECM representation proper:

\[ \Delta y_t = B(0)\Delta x_t + A^*(L)\Delta y_{t-1} + B^*(L)\Delta x_{t-1} - A(1)\left[y_{t-1} - c'x_{t-1}\right] + \epsilon_t \]

where \( c' = R(1) \frac{\mu}{R(1)} \) 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 $\varepsilon_t$;

2. a short-run transitory component: $B(0)\Delta x_t + A^*(L)\Delta y_{t-1} + B^*(L)\Delta x_{t-1}$; the first coefficient, $B(0)$, gives you the instantaneous effect of $x_t$ on $y_t$;

3. a long-run component whose base ingredient is the long-run multiplier $c$. 
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

\[ y = X\beta + \varepsilon, \]

where \( \varepsilon \) is defined as \( y - E[y|X] \), with the appropriate adjustments for dynamic models; the derived property \( E[\varepsilon|X] = 0 \) is the key ingredient for guaranteeing consistency of OLS as an estimator of \( \beta \). 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 3.5, 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:

\[ y_i = x_i^* \beta + \varepsilon_i \]  

(6.1)
where $E[y_i|x_i^*] = x_i^* \beta$ and $\beta$ 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:

$$
x_i = x_i^* + \eta_i
$$

where $\eta_i$ is a zero-mean random variable, independent of $x_i^*$ and $\epsilon_i$, with variance $\sigma_\eta^2 > 0$; clearly, the larger $\sigma_\eta^2$ is, the worse is the quality of our measurement for the variable of interest $x_i^*$. One may think that, since $\eta_i$ 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 $\beta$. This is quite easy to prove: combine the two equations above to get

$$
y_i = x_i \beta + \epsilon_i - \beta \eta_i = x_i \beta + u_i
$$

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

$$
E[x_i u_i] = E[(x_i^* + \eta_i)(\epsilon_i - \beta \eta_i)] = E[x_i^* \epsilon_i] - \beta E[x_i^* \eta_i] + E[\eta_i \epsilon_i] - \beta E[\eta_i^2] = -\beta \sigma_\eta^2
$$

If we define $Q = E[x_i^2]$, clearly

$$
\hat{\beta} \xrightarrow{p} \beta - \frac{\beta \sigma_\eta^2}{Q} = \beta \left(1 - \frac{\sigma_\eta^2}{Q}\right) \neq \beta
$$

It can be proven that $0 < \sigma_\eta^2 < Q$, 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 $\sigma_\eta^2$ relative to $Q$; second, the asymptotic bias is such that $|\text{plim}(\hat{\beta})| < |\beta|$; 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 $E[y_i|x_i]$ is not equal to $x_i \beta$; if it were, $E[x_i u_i]$ would be zero, but we just showed it isn’t. Since OLS is programmed to estimated the parameters of a conditional expectation, you can’t expect it to come up with anything else.

---

1Come 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

\[ \text{The fundamental psychological law […] is that men are disposed, as a rule and on the average, to increase their consumption as their income increases but not by as much as the increase in the income.} \]

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.

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.

Consider one of the simplest examples of simultaneous equation system in economics: a two-equation linear model of supply and demand for a good.

\[
q_t = \alpha_0 - \alpha_1 p_t + u_t \quad (6.4)
\]

\[
p_t = \beta_0 + \beta_1 q_t + v_t, \quad (6.5)
\]

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 \(E[u_t] = E[v_t] = 0\).

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.

One simple way to convince yourself is simply to consider that if \(E[q_t | p_t]\) 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 \(E[p_t | q_t]\) 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 \(E[u_t | p_t]\) should be 0. Now substitute (6.4) into equation (6.5):

\[
p_t = \beta_0 + \beta_1 (\alpha_0 - \alpha_1 p_t + u_t) + v_t
\]

\[
= (\beta_0 + \beta_1 \alpha_0) - (\beta_1 \alpha_1) p_t + (v_t + \beta_1 u_t) \Rightarrow
\]

\[
(1 + \beta_1 \alpha_1) p_t = (\beta_0 + \beta_1 \alpha_0) + (v_t + \beta_1 u_t) \Rightarrow
\]

\[
p_t = \pi_1 + \eta_t, \quad (6.6)
\]

where the constant \(\pi_1\) 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:

\[
\text{Cov}[p_t, u_t] = E[p_t \cdot u_t] = E[u_t \pi_1 + u_t \eta_t] = 0 + E[u_t \cdot (v_t + \beta_1 u_t)] =
\]

\[
= \text{Cov}[v_t, u_t] + \beta_1 V(u_t)
\]
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), \( \text{Cov}[p_t, u_t] \neq 0 \). Borrowing on the definition I gave in Section 3.5, the variable \( p_t \) is clearly endogenous.

But if \( \text{Cov}[p_t, u_t] \neq 0 \), then \( E[u_t | p_t] \) can't be 0 either; therefore, \( (\alpha_0 - \alpha_1 p_t) \) can't be \( E[q_t | p_t] \), 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 \( \alpha_0 \) and \( \alpha_1 \).

To be more specific: even assuming that \( E[q_t | p_t] \) is a linear function like
\[
E[q_t | p_t] = \gamma_0 + \gamma_1 p_t,
\]
OLS gives you an excellent estimate of the coefficients \( \gamma_0 \) and \( \gamma_1 \); unfortunately, they are not the same thing as \( \alpha_0 \) and \( \alpha_1 \).

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 \( \beta_0 \) and \( \beta_1 \), either. This example will be generalised in section 6.6.2.

### 6.2 The IV estimator

In a standard linear model \( y_i = x_i' \beta + \epsilon_i \). As we argued in chapter 3, the assumption \( x_i' \beta = E[y_i | x_i] \) is crucial for the consistency of the OLS statistic as an estimator of \( \beta \); in fact, you could see this assumption as a definition of \( \beta \), in that \( \beta \) is the only vector for which the following equation is true:
\[
E[X' (y - X' \beta)] = 0. \tag{6.8}
\]
The OLS statistic \( \hat{\beta} \), instead, is implicitly defined by the relationship
\[
X' (y - X \hat{\beta}) = 0. \tag{6.9}
\]
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 \( \hat{\beta} \) works quite nicely as an estimator of its counterpart \( \beta \) just agrees with common sense.

If, on the contrary, the parameter of interest \( \beta \) satisfies an equation other than (6.8), then we may proceed by analogy and use, as an estimator, a statistic \( \tilde{\beta} \) that satisfies the corresponding sample property. In this chapter, we assume we have a certain number of observable variables \( W \) for which
\[
E[W' (y - X \tilde{\beta})] = 0. \tag{6.10}
\]
The corresponding statistic will then be implicitly defined by
\[
W' (y - X \tilde{\beta}) = 0. \tag{6.11}
\]
The variables $W$ are known as \textit{instrumental variables}, or, more concisely, \textit{instruments}. The so-called “simple” IV estimator can then be defined as follows: if we had a matrix $W$, of the same size as $X$, satisfying (6.10), then then we may define a statistic $\tilde{\beta}$ such that (6.11) holds:

$$W'(y - X\tilde{\beta}) = 0 \implies W'X\cdot \tilde{\beta} = W'y.$$ 

In a parallel fashion, the difference $\varepsilon = y - X\beta$ is not defined as the difference between $y$ and it conditional mean, but rather as a zero-mean random variable which describes how much $y$ deviates from its “standard” value, as described by the “structural” relationship $X\beta$. A term that we use in this context is \textit{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 $W$ has as many columns as $X$, then the matrix $W'X$ is square; if it’s also invertible, then $\tilde{\beta}$ is

$$\tilde{\beta} = (W'X)^{-1}W'y; \quad (6.12)$$

this is sometimes called the “simple” IV estimator.

The actual availability of the variables $W$ 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.

\subsection*{6.2.1 The generalised IV estimator}

What if the number of columns of $W$ (call it $m$) was different from number of columns from $X$ (call it $k$)? Of course, the matrix $W'X$ 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 $W$ 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 $W$?

Fortunately, there is a solution: assume, for simplicity, that the covariance matrix of the structural disturbances is a multiple of the identity matrix:\footnote{In fact, this assumption is not strictly necessary, but makes for a cleaner exposition.}

$$E[\varepsilon\varepsilon' | W] = \sigma^2 I.$$ 

By hypothesis, $E[\varepsilon | W] = 0$; therefore,

$$E[W'\varepsilon\varepsilon'W | W] = \sigma^2 W'W = \sigma^2 \Omega.$$
6.2. THE IV ESTIMATOR

Now define $v$, $C$ and $e$ as

$$
\begin{align*}
    v & = W'y \quad C = W'X \quad e = W'\varepsilon; \\
    m \times 1 & \quad m \times k & \quad m \times k
\end{align*}
$$

so the following equality holds:

$$
    v = C\beta + e, \quad (6.13)
$$

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 $X$) but the peculiar feature of this model is that the number of “observations” is $m$ (the column size of $W$).

Since $\Omega$ is observable (up to a constant), we may apply the GLS estimator (see 4.2.1) to (6.13) and write

$$
    \hat{\beta} = \left[ C'\Omega^{-1}C \right]^{-1} C'\Omega^{-1} v = \left[ X'W(W'W)^{-1}W'X \right]^{-1} X'W(W'W)^{-1} W'y = \left( X'P_W X \right)^{-1} X'P_W y. \quad (6.14)
$$

This clever idea is due to the English econometrician Denis Sargan, whose name will also crop up later, in section 6.7.1.

The estimator $\hat{\beta}$ 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 $WX$ 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**.

---

6For 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 $\beta$, you also get a residual vector $\tilde{\varepsilon} = y - X\tilde{\beta}$, so that an estimator of $\sigma^2$ is readily available:

$$\tilde{\sigma}^2 = \frac{\tilde{\varepsilon}'\tilde{\varepsilon}}{n}. $$

As shown in section 6.A.1, it can be proven that, under the set of assumptions I just made, the statistics $\tilde{\beta}$ and $\tilde{\sigma}^2$ are CAN estimators. Therefore, the whole testing apparatus we developed in Chapter 3 can be applied without modifications since

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, V).$$

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

$$\hat{V} = \tilde{\sigma}^2(X'PWX)^{-1}.$$

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:

$$\tilde{\beta} = \operatorname{Argmin}_{\beta \in \mathbb{R}^k} \varepsilon(\beta)'PWX(\beta)$$

(compare the above expression with equation (1.14)).

In this book, we will not make much use of this property. However, defining $\tilde{\beta}$ 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 $\tilde{\beta}$ and explore what requisites they imply for the variables we choose as instruments. The two conditions are:

1. $\frac{1}{n}\sum w_i \varepsilon_i \xrightarrow{p} 0$

2. $\frac{1}{n}\sum x_i w_i' \xrightarrow{p} A$, where $A$ is a $k \times m$ matrix with rank $k$.

Condition 1 is more or less guaranteed by (6.10), that is by $w_i$ being exogenous, which basically means “uncorrelated with the structural disturbance $\varepsilon_i$; if this requirement isn’t met, the limit in probability of $\tilde{\beta}$ is not $\beta$. 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 $E[x_i' \varepsilon_i]$ 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 \( x_i \) 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 \( x \). 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 \( \tilde{\beta} \) is not computable, since \( (X'P_WX) \) is singular and your software will complain about this.

The relevance condition, instead, is much trickier to spot, since (with probability 1) \( \frac{1}{n} \sum x_i w_i \) will have rank \( k \) even if \( A \) doesn’t. Hence, if \( \text{rk}(A) < k \), you will be able to compute \( \tilde{\beta} \), 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 \( \tilde{\beta} \) 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

\[
\tilde{\beta} = \frac{n^{-1} \sum_{i=1}^{n} x_i y_i}{n^{-1} \sum_{i=1}^{n} w_i x_i};
\]

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 \( \tilde{\beta} \) 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 \( n^{-1} \sum_{i=1}^{n} w_i x_i \overset{p}{\longrightarrow} A = 0 \) (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 \( \tilde{\beta} \) 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 \( X \) and \( \varepsilon \) are correlated, any variable perfectly correlated to \( X \) could not be orthogonal to \( \varepsilon \). The trick here is that \( W \) is not perfectly correlated to \( X \).
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:

\[ y_i = z_i' \beta_0 + e_i \beta_1 + \epsilon_i \]  (6.15)

where \( y_i \) is the log wage for an individual, \( e_i \) is their education level and the vector \( z_i \) contains other characteristics we want to control for (gender, work experience, etc). The parameter of interest is \( \beta_1 \), 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 1.6. 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 an endogeneity problem.  

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:

---

9The literature on this topic is truly massive. A good starting point is Card (1999).
6.4. THE HAUSMAN TEST

OLS, using observations 1-1917
Dependent variable: lw

<table>
<thead>
<tr>
<th></th>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>1.32891</td>
<td>0.0355309</td>
<td>37.40</td>
<td>2.86e-230 ***</td>
</tr>
<tr>
<td>male</td>
<td>0.175656</td>
<td>0.0143224</td>
<td>12.26</td>
<td>2.42e-33 ***</td>
</tr>
<tr>
<td>wexp</td>
<td>0.00606615</td>
<td>0.000671303</td>
<td>9.066</td>
<td>2.97e-19 ***</td>
</tr>
<tr>
<td>educ</td>
<td>0.0526218</td>
<td>0.00202539</td>
<td>25.98</td>
<td>1.02e-127 ***</td>
</tr>
</tbody>
</table>

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
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-
CHAPTER 6. INSTRUMENTAL VARIABLES

TSLS, using observations 1-1917
Dependent variable: lw
Instrumented: educ
Instruments: const male wexp age peduc

<table>
<thead>
<tr>
<th></th>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>0.943553</td>
<td>0.0684539</td>
<td>13.78</td>
<td>2.80e-41 ***</td>
</tr>
<tr>
<td>male</td>
<td>0.182926</td>
<td>0.0149929</td>
<td>12.20</td>
<td>5.02e-33 ***</td>
</tr>
<tr>
<td>wexp</td>
<td>0.00860475</td>
<td>0.000795375</td>
<td>10.82</td>
<td>1.62e-26 ***</td>
</tr>
<tr>
<td>educ</td>
<td>0.0792106</td>
<td>0.00449758</td>
<td>17.61</td>
<td>1.85e-64 ***</td>
</tr>
</tbody>
</table>

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 $X$ is exogenous, OLS is more efficient that IV under standard conditions.\(^{10}\)

The Hausman test\(^{11}\) is based on the idea of comparing the two estimators and checking if their difference is statistically significant. 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 $X$ 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 ($\hat{\theta}$) is robust to some problem and the other one isn't ($\tilde{\theta}$), 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

$$H = \delta' \left[ \overline{V}(\delta) \right]^{-1} \delta; \quad (6.16)$$

where $\overline{V}(\delta)$ is a consistent estimator of $AV[\delta]$, and, under some standard regularity conditions, it can be proven that $H$ is asymptotically $\chi^2$ 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.

\(^{12}\) 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.
The problem is, how do you compute \( V(\hat{\delta}) \)? 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 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 

\[
H = \left[ \hat{\theta} - \tilde{\theta} \right]' \left( AV[\hat{\theta}] - AV[\tilde{\theta}] \right)^{-1} \left[ \hat{\theta} - \tilde{\theta} \right].
\]

Table 6.3: Hausman Test — special case

<table>
<thead>
<tr>
<th>if ( H_0 ) is true</th>
<th>if ( H_0 ) is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\theta} )</td>
<td>CAN and efficient</td>
</tr>
<tr>
<td>( \tilde{\theta} )</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>( \hat{\theta} )</td>
<td>CAN but not efficient</td>
</tr>
</tbody>
</table>

In our case, the two estimators to compare are \( \beta \) and \( \tilde{\beta} \), so \( \delta = \tilde{\beta} - \beta \). If we assume that OLS is efficient (which would be under homoskedasticity), then

\[
V[\delta] = V[\tilde{\beta}] - V[\hat{\beta}].
\]

Since under \( H_0 \) OLS is consistent, then \( \hat{\sigma}^2 \) is a consistent estimator of \( \sigma^2 \) and we can use the matrix

\[
\hat{\sigma}^2 \left[ (X'PWX)^{-1} - (X'X)^{-1} \right];
\]

therefore,

\[
H = \frac{(\hat{\beta} - \tilde{\beta})' \left[ (X'PWX)^{-1} - (X'X)^{-1} \right]^{-1} (\hat{\beta} - \tilde{\beta})}{\hat{\sigma}^2}. \tag{6.17}
\]

In practice, actual computation of the test is performed even more simply, via an auxiliary regression: consider the model

\[
y = X\beta + \hat{X}\gamma + \text{residuals}. \tag{6.18}
\]

where \( \hat{X} \equiv PX \). By the Frisch–Waugh theorem (see section 1.5)

\[
\hat{\gamma} = [\hat{X}'M_x\hat{X}]^{-1}\hat{X}'M_xy;
\]

now rewrite the two matrices on the right-hand side of the equation above as

\[
\hat{X}'M_x\hat{X} = \hat{X}'\hat{X} - \hat{X}'P_x\hat{X} = \hat{X}'\hat{X} - \hat{X}'X(X'X)^{-1}X'\hat{X} = \\
= (\hat{X}'\hat{X}) \left[ (X'X)^{-1} - (X'X)^{-1} \right] (\hat{X}'\hat{X}) \tag{6.19}
\]

\[
\hat{X}'M_xy = \hat{X}'y - \hat{X}'X(X'X)^{-1}X'y = (\hat{X}'\hat{X}) (\hat{\beta} - \tilde{\beta}) \tag{6.20}
\]

13The 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 $\hat{X}'X = X'P_WX = \hat{X}'\hat{X}$; therefore,

$$\hat{\gamma} = (\hat{X}'\hat{X})^{-1} \left[ (\hat{X}'\hat{X})^{-1} - (X'X)^{-1} \right]^{-1} (\hat{\beta} - \hat{\beta}). \quad (6.21)$$

A Wald test for $\gamma = 0$ is

$$W = \frac{\hat{\gamma}'[\hat{X}'M_X\hat{X}] \hat{\gamma}}{\hat{\sigma}^2},$$

so, after performing a few substitutions, you get

$$W = \frac{(\hat{\beta} - \hat{\beta})' \left[ (\hat{X}'\hat{X})^{-1} - (X'X)^{-1} \right]^{-1} (\hat{\beta} - \hat{\beta})}{\hat{\sigma}^2} = H.$$

Of course, the possibility of $X$ and $W$ having some columns in common complicates slightly the setup above, and $\hat{X}$ should only contain the projection on $W$ of the endogenous regressors, because if a regressor is also contained in $W$ 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 $W$).

**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 package 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 $\chi^2$ 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 $\hat{\beta}$ 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 $\hat{\beta}$ 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 $\hat{X} = P_WX$ and rewrite (6.14) as follows:

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'y,$$

(6.22)

The matrix $\hat{X}$ contains, in the $j$-th column, the fitted value of a regression of the $j$-th column of $X$ on $W$; the regression

$$x_i = \Pi w_i + u_i$$

(6.23)

is what we call the first stage regression. In the second stage, you just regress $y$ on $\hat{X}$: the OLS coefficient equals $\hat{\beta}$. Note: this is a numerically valid procedure for computing $\hat{\beta}$, but the standard errors you get are not valid for inference. This is because second stage residuals are $e = y - \hat{X}\hat{\beta}$, which is a different vector from the IV residuals $\tilde{e} = y - X\tilde{\beta}$. Consequently, the statistic $\frac{e'e}{n}$ does not provide a valid estimator of $\sigma^2$, which in turn makes the estimated covariance matrix invalid.

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 $X$ (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 $W$. 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).

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;

\(^{14}\)In fact, the word Henri Theil used when he invented this method was “rounds”, but subsequent literature has settled on “stages”.
CHAPTER 6. INSTRUMENTAL VARIABLES

OLS, using observations 1-1917
Dependent variable: educ

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>5.43051</td>
<td>0.491957</td>
<td>11.04</td>
</tr>
<tr>
<td>male</td>
<td>-0.137838</td>
<td>0.142828</td>
<td>-0.9651</td>
</tr>
<tr>
<td>wexp</td>
<td>-0.205149</td>
<td>0.012957</td>
<td>-15.83</td>
</tr>
<tr>
<td>age</td>
<td>0.192950</td>
<td>0.014661</td>
<td>13.16</td>
</tr>
<tr>
<td>peduc</td>
<td>0.386233</td>
<td>0.020728</td>
<td>18.63</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>0.943553</td>
<td>0.071103</td>
<td>13.27</td>
</tr>
<tr>
<td>male</td>
<td>0.182926</td>
<td>0.015573</td>
<td>11.75</td>
</tr>
<tr>
<td>wexp</td>
<td>0.00860475</td>
<td>0.000826</td>
<td>10.42</td>
</tr>
<tr>
<td>hat_educ</td>
<td>0.0792106</td>
<td>0.004671</td>
<td>16.96</td>
</tr>
</tbody>
</table>

SSR = 209.316, R-squared = 0.173936

Table 6.5: Wage equation on the SHIW dataset — second stage
2. the standard errors 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 \( \sigma^2 \); 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

\[
y_i = x_i \cdot \beta + \varepsilon_i
\]

(6.24)

\[
x_i = w_i \cdot \pi + u_i
\]

(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 \( u_i \) is exogenous, the only way \( x_i \) can be correlated with \( \varepsilon_i \) is if \( \text{Cov}[u_i, \varepsilon_i] \neq 0 \). It is easy to show that OLS will overestimate \( \beta \) 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

\[
y_i = x_i \cdot \beta + u_i \cdot \theta + \nu_i
\]

(6.26)

If we could observe \( u_i \), we wouldn’t need the IV estimator at all, because \( \nu_i \) is orthogonal to both \( x_i \) and \( u_i \), so OLS would be our tool of choice.

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; the possibility of using those 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 \( E \) the first-stage residuals and perform an OLS regression of \( y \) on \( X \) and \( E \) together:

\[
y = X\beta + E\theta + \nu;
\]

(6.27)

the estimates we get have a very nice interpretation.

\footnote{Lazy writers like myself love the sentence: “the proof is left to the reader as an exercise.”}
Let’s begin with \( \beta \): by the Frisch-Waugh theorem, the coefficient vector associated with \( X \) is
\[
(X'MEX)^{-1}X'MEY;
\]
now focus on the matrix \( X'ME \):
\[
X'ME = X' - X'MWX(X'MWX)^{-1}X'MW = X' - X'MW = X'PW;
\]
therefore, the coefficient vector becomes
\[
(X'MEX)^{-1}X'MEY = (X'PWX)^{-1}X'PY = \tilde{\beta}
\]
So the OLS estimate of the coefficients associated with \( X \) in equation (6.27) is exactly equal to the IV estimator \( \tilde{\beta} \). Warning: the reader should be aware that, just like the two-stage procedure, the control function approach does not yield correct standard errors for \( \tilde{\beta} \); 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 \( \theta \) in (6.27) is
\[
\hat{\theta} = [E'MX]^{-1}E'MXy;
\]
of course \( \hat{e} = MXy \) are the OLS residuals. Now use the definition of \( E \) as \( E = MWX \) again:
\[
E'MXy = E'\hat{e} = X'MWX\hat{e} = X'\hat{e} - X'PW\hat{e} = -X'PW\hat{e};
\]
where the last equality comes from the OLS residuals \( \hat{e} \) being orthogonal to \( X \); therefore
\[
E'MXy = -X'PWy + X'PWX\cdot \hat{\beta} = [X'PWX] (\hat{\beta} - \tilde{\beta}),
\]
and as a consequence \( \hat{\theta} \) is just
\[
\hat{\theta} = [E'MX]^{-1} [X'PWX] (\hat{\beta} - \tilde{\beta});
\]
since the matrix \([E'MX]^{-1} [X'PWX]\) is invertible, \( \hat{\theta} \) it can be zero if and only if \( \hat{\beta} = \tilde{\beta} \). Therefore, the hypothesis \( h_0: \theta = 0 \) is logically equivalent to the null hypothesis of the Hausman test, and the test can be performed a a simple zero

\(^{16}\)If you really want to know, 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.
restriction on $\hat{\theta}$. If (as often happens) $\theta$ is a scalar, the result of the Hausman test is immediately visible as the significance $t$-test associated with that coefficient.

A very nice features of the control function approach is that it 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

<table>
<thead>
<tr>
<th></th>
<th>coefficient</th>
<th>std. error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>0.943553</td>
<td>0.0647377</td>
<td>14.58</td>
<td>1.04e-45 ***</td>
</tr>
<tr>
<td>male</td>
<td>0.182926</td>
<td>0.0141790</td>
<td>12.90</td>
<td>1.42e-36 ***</td>
</tr>
<tr>
<td>wexp</td>
<td>0.00860475</td>
<td>0.000752195</td>
<td>11.44</td>
<td>2.34e-29 ***</td>
</tr>
<tr>
<td>educ</td>
<td>0.0792106</td>
<td>0.00425341</td>
<td>18.62</td>
<td>2.89e-71 ***</td>
</tr>
<tr>
<td>resid</td>
<td>-0.0341349</td>
<td>0.00481934</td>
<td>-7.083</td>
<td>1.98e-12 ***</td>
</tr>
</tbody>
</table>

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 (displayed in Table 6.4) under the name resid, you can run an OLS regression just like the one in Table 6.1, with the only difference that you add resid to the list of regressors. The results are in Table 6.6. Again, the coefficients for the original regressors are $\hat{\beta}$ and again, the standard errors are not to be trusted (they’re a bit smaller than the one 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)

\[ y_i = x_i^\star \beta + \epsilon_i \]
\[ x_i = x_i^\star + \eta_i \]
suppose you have a third observable variable $w_t$ such that

$$w_t = x_t^* + v_t$$

It can be proven that, if $\eta_t$ and $v_t$ are uncorrelated, then $w_t$ is a valid instrument and the statistic

$$\tilde{\beta} = \frac{\sum w_t y_t}{\sum w_t x_t}$$

is consistent for $\beta$.

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):

\[
q_t = \alpha_0 - \alpha_1 p_t + u_t \\
p_t = \beta_0 + \beta_1 q_t + v_t.
\]

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 $E[p_t] = \pi_1$; clearly, the parameter $\pi_1$ 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 $E[q_t] = \pi_0$.

Can we use these two parameters to estimate the structural ones? The answer is no: the relationship between the $(\pi_0, \pi_1)$ pair and the structural parameters is

$$\pi_0 = \frac{\alpha_0 - \beta_0 \alpha_1}{1 + \beta_1 \alpha_1}, \quad \pi_1 = \frac{\beta_0 + \beta_1 \alpha_0}{1 + \beta_1 \alpha_1},$$

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

$$\Gamma y_t = B x_t + \varepsilon_t; \quad (6.28)$$

the $y_t$ vector contains the $q$ endogenous variables, while $x_t$ is an $m$-vector holding the exogenous variables. The matrix $\Gamma$ (assumed non-singular) is $q \times q$ and $B$ is a $q \times m$ matrix. In the demand-supply example,

$$\Gamma = \begin{bmatrix} 1 & \alpha_1 \\ -\beta_1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}. $$

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.
6.7. ARE MY INSTRUMENTS OK?

By pre-multiplying (6.28) by $\Gamma^{-1}$, you get the so-called reduced form:

$$y_t = \Pi x_t + u_t,$$

where $\Pi = \Gamma^{-1} B$ and $u_t = \Gamma^{-1} \varepsilon_t$. In our example, the matrix $\Pi$ is a column vector, containing $\pi_0$ and $\pi_1$.

If $x_t$ is exogenous, then Cov[$x_t, \varepsilon_t$], so the correlation between $x_t$ and $u_t$ is zero; hence, OLS is a consistent estimator for the parameters of the reduced form. However, by postmultiplying (6.29) by $x_t'$ you get:

$$y_t x_t' = \Pi x_t x_t' + u_t x_t',$$

which implies $E[y_t x_t'] = \Pi E[x_t x_t']$. Ordinarily, this matrix should not contain zeros; if variables were centred in mean, it would be the covariance matrix between the vector $y_t$ and the vector $x_t$. Therefore, each element of $x_t$ is correlated with each $y_t$ despite being uncorrelated with $\varepsilon_t$. In other words, $x_t$ is both relevant and exogenous and, as such, is a perfect instrument.

In the example above, $x_t$ contains only the constant term, and the reduced form looks like:

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} \cdot 1 + u_t;$$

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:

$$q_t = a_0 - a_1 p_t + a_2 y_t + u_t \quad (6.30)$$

$$p_t = \beta_0 + \beta_1 q_t + \beta_2 m_t + v_t, \quad (6.31)$$

and the two new variables (both assumed exogenous) are:

| $y_t$ | per capita income at time $t$ |
| $m_t$ | price of raw materials at time $t$ |

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. It goes without
saying that we cannot just 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 little bit more subtle.

Exogeneity, in our context, means uncorrelatedness between the instruments \( w_i \) and the structural disturbances \( \varepsilon_i \). If the latter were observable, a test would be straightforward to construct: under \( H_0 \),

\[
\frac{1}{\sqrt{n}} W' \varepsilon \xrightarrow{d} N(0, \sigma^2 W' W)
\]

and it can be proven that, under \( H_0 \),

\[
\frac{\varepsilon' P_W \varepsilon}{\sigma^2} \xrightarrow{d} \chi^2_m.
\] (6.32)

where \( m \) is the size of \( w_i \). Unfortunately, \( \varepsilon \) 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 \( \chi^2 \) is not \( m \) (the number of instruments), but rather \( m - k \), where \( k \) is the number of elements in \( \beta \): in other terms, the over-identification rank (see section 6.2.1). In formulae,

\[
S = \frac{\tilde{\varepsilon}' P_W \tilde{\varepsilon}}{\sigma^2} \xrightarrow{d} \chi^2_{m-k}.
\] (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 \( \chi^2_m \) variable. Can this result be generalised?

Consider the conditional expectation \( E[y_i|w_i] \); assuming linearity, if \( w_i \) is exogenous we could estimate its parameters via OLS in a model like

\[
y_i = w_i' \pi + u_i; \tag{6.34}
\]

in the simultaneous system jargon, equation (6.34) would be the reduced-form equation for \( y_i \). To see how the parameters \( \pi \) relate to \( \beta \), write the first-stage equation (6.25) in matrix form as

\[
X = W\Pi + E
\]

\(^{17}\)Blitz proof: if \( m = k \), then \( \tilde{\beta} = (W'X)^{-1}W'y \). Therefore, \( P_W \tilde{\varepsilon} = P_W(y - X\tilde{\beta}) \); however, observe that \( P_W X \tilde{\beta} = W'(W'W)^{-1}W'X(W'X)^{-1}W'y = P_W y \). As a consequence, \( P_W \tilde{\varepsilon} = P_W y - P_W y = 0 \).
so that the $i$-th row of $X$ can be written as $x_i' = w_i' \Pi + e_i'$; now substitute in the structural equation:

$$y_i = x_i' \beta + \varepsilon_i = w_i' \Pi \beta + (\varepsilon_i + e_i' \beta);$$

(6.35) clearly, the two models (6.34) and (6.35) become equivalent only if $\pi = \Pi \beta$; in fact, the expression $\Pi \beta$ can be seen as a restricted version of $\pi$, where the constraint is that $\pi$ must be a linear combination of the columns of $\Pi$ (or, more concisely, that $\pi \in \text{Sp}(\Pi)$).

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 $\tilde{\varepsilon}$ are the residuals from the restricted model. All we have to do is apply the LM principle (see Section 3.6.1) 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, I will now prove that the number of degrees of freedom of the $\chi^2$ limit distribution is $m - k$: if the restriction is true, then $P_{\Pi} \pi = \pi$, or, equivalently, $M_{\Pi} \pi = 0$. Since the rank of $M_{\Pi}$ is $m - k$, we can write

$$M_{\Pi} \pi = UV' \pi = 0$$

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 $H_0: V' \pi = 0$, which is a system of $m - r$ constraints.

Now we can tackle point 3: the null hypothesis in the Sargan test is that the $m$ relationships implicit in the equation $\pi = \Pi \beta$ are non-contradictory; if they were, it would mean that at least one element of the vector $E[w_i \varepsilon_i]$ is non-zero, 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 $\beta$:

- a simple IV estimator using $w_{1,i}$ only (call it $\beta_1$);
- a simple IV estimator using $w_{2,i}$ only (call it $\beta_2$);
- the GIVE estimator using both $w_{1,i}$ and $w_{2,i}$ (call it $\beta_{12}$).
CHAPTER 6. INSTRUMENTAL VARIABLES

Suppose $\beta_1$ turns out to be positive (and very significant), and $\beta_2$ turns out to be negative (and very significant); clearly, there must be something wrong. At least one between $\beta_1$ and $\beta_2$ 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, $\beta_{12}$ would be certainly inconsistent, and we’d have to decide which one to keep between $\beta_1$ and $\beta_2$ (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 $\beta_{12}$ 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 = $P(\text{Chi-square}(1) > 0.138522) = 0.709755$

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 over-identification 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 $A = E[x_i w_i']$ 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 $\text{rk}(A) = m$ 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.\textsuperscript{18}

The problem is best exemplified by a little simulation study: consider the same model we used in section 6.5.1:

\begin{align*}
y_i &= x_i \cdot \beta + \epsilon_i \quad (6.36) \\
x_i &= w_i \cdot \pi + u_i \quad (6.37) \\
\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} &\sim N \left( \mathbf{0}, \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix} \right)
\end{align*}

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 $\epsilon_i$ and $u_i$ are correlated, then $x_i$ is endogenous; however, we have the variable $u_i$, which meets all the requirements for being a perfectly valid instrument: it is exogenous (uncorrelated with $\epsilon_i$) and relevant (correlated with $x_i$), as long as the parameter $\pi$ is nonzero.

However, if $\pi$ is a small number the correlation between $x_i$ and $w_i$ is very faint, so $u_i$ is weak, despite being relevant. In this experiment, the IV estimator is simply

$$\hat{\beta} = (W'X)^{-1}W'y = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i x_i} :$$

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 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.

\textsuperscript{18}Compared to the rest of the material contained in this book, inference under weak instruments is a fairly recent strand in econometric research. The obligatory references are Staiger and Stock (1997) and Stock and Yogo (2005), but a fairly accessible introductory treatment can be found in Hill et al. (2018), Chapter 10. Chapter 12 of Hansen (2019) is considerably more technical, but highly recommended.
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.

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 $\beta$ 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 $\pi$ 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 $\beta$;\footnote{I should add “provided it exists”; there are cases when the distribution of the IV estimator has no finite moments.}

- even more worryingly, the asymptotic approximations we use for our test statistics may be very misleading.

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 3.6.1) for the exclusion of the “true” instruments (that is, the elements of \( w_i \) not contained in \( x_i \)). 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:

<table>
<thead>
<tr>
<th>size</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>19.93</td>
<td>11.59</td>
<td>8.75</td>
<td>7.25</td>
</tr>
</tbody>
</table>

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)

\[ \hat{\beta} = (X'PWX)^{-1}X'PWY, \]

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

\[ y_i = x_i'\beta + \varepsilon_i, \]

we assume that:

1. \[ \frac{1}{n} \sum_{t=1}^{T} x_i w'_t = \frac{X'W}{n} \xrightarrow{P} A, \text{ where } \text{rk}(A) = k; \]
2. \[ \frac{1}{n} \sum_{t=1}^{T} w_t w'_t = \frac{WW}{n} \xrightarrow{P} B, \text{ where } B \text{ is invertible;} \]
3. \[ \frac{1}{n} \sum_{t=1}^{T} w_t u_t = \frac{W\varepsilon}{n} \xrightarrow{P} 0; \]

then \( \hat{\beta} = (X'PWX)^{-1}X'PWY \xrightarrow{P} \beta. \) The proof is a simple application of Slutsky’s theorem:

\[ \hat{\beta} = \beta + (X'PWX)^{-1}X'PW\varepsilon = \beta + \left( \frac{X'W}{n} \right)^{-1} \left( \frac{WW}{n} \right)^{-1} \left( \frac{W\varepsilon}{n} \right) \]

so that

\[ \hat{\beta} \xrightarrow{P} \beta + \left[ AB^{-1}A' \right]^{-1} AB^{-1} \cdot 0 = \beta. \quad (6.38) \]

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 \( w_1 \) that you assume independent of \( \varepsilon_i, \) you can use as instruments \( w_1, w_1^2, \log(w_1), \ldots \) (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 $[AB^{-1}A']$ to be invertible. This means that instruments $w_i$ must be relevant (see Section 6.2.2) for all the regressors $x_i$. If $[AB^{-1}A']$ 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 $\tilde{\beta}$: if we also assume

$$4. \frac{1}{\sqrt{n}} \sum_{t=1}^{T} w_t u_t \xrightarrow{d} N(0, Q);$$

then $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, \Sigma)$, where

$$\Sigma = [AB^{-1}A']^{-1} AB^{-1}QB^{-1}A' [AB^{-1}A']^{-1}.$$  

In the standard case $Q = \sigma^2 B$ (from the homoskedasticity assumption $E[\varepsilon \varepsilon' | W] = \sigma^2 I$), and therefore

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N\left(0, \sigma^2 [AB^{-1}A']^{-1}\right). \quad (6.39)$$

So the precision of the IV estimator is severely impaired any time the matrix $[AB^{-1}A']$ is close to being singular.

The last thing is proving that $\tilde{\sigma}^2$ is consistent: from

$$\tilde{\varepsilon} = X(\beta - \tilde{\beta}) + \varepsilon,$$

you get

$$\tilde{\varepsilon}' \tilde{\varepsilon} = (\beta - \tilde{\beta})' X'X (\beta - \tilde{\beta}) + 2 (\beta - \tilde{\beta})' X' \varepsilon + \varepsilon' \varepsilon$$

and therefore

$$\frac{1}{n} \tilde{\varepsilon}' \tilde{\varepsilon} = (\beta - \tilde{\beta})' \left( \frac{XX}{n} \right) (\beta - \tilde{\beta}) + 2 (\beta - \tilde{\beta})' \left( \frac{X' \varepsilon}{n} \right) + \left( \frac{\varepsilon' \varepsilon}{n} \right).$$

By taking probability limits,

$$\tilde{\sigma}^2 = \frac{1}{n} \tilde{\varepsilon}' \tilde{\varepsilon} \xrightarrow{P} 0' (Q) 0 + 2 \cdot 0' \lambda + \sigma^2 = \sigma^2,$$

where $Q = \text{plim} \left( \frac{XX}{n} \right)$ and $\lambda = \text{plim} \left( \frac{X' \varepsilon}{n} \right)$. Note that $\tilde{\sigma}^2$ is consistent even though $\lambda \neq 0$. 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 $\tilde{\sigma}^2(X'P_W X)^{-1}$ 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 X is exogenous can be given as follows: given the model \( y_i = x_i' \beta + \epsilon_i \), define the following quantities:

\[
Q = \text{plim} \left( \frac{X'X}{n} \right) \quad A = \text{plim} \left( \frac{W'X}{n} \right) \quad B = \text{plim} \left( \frac{W'W}{n} \right);
\]

under homoskedasticity, we have that \( \text{AV} [\hat{\beta}] = \sigma^2 Q^{-1} \) (see section 3.2.2) and \( \text{AV} [\tilde{\beta}] = \sigma^2 [A' B^{-1} A]^{-1} \) (see 6.A.1), where \( \sigma^2 = \text{V} [\epsilon_i] \). In order to prove that \( \text{AV} [\tilde{\beta}] - \text{AV} [\hat{\beta}] \) is positive semi-definite, we re-use two of the results on positive definite matrices that we employed in section 4.A.3:

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

\[
\Delta \equiv \sigma^2 \cdot \left[ \text{AV} [\tilde{\beta}]^{-1} - \text{AV} [\hat{\beta}]^{-1} \right] = Q - A' B^{-1} A;
\]

Since \( \sigma^2 > 0 \), it is sufficient to prove that \( \Delta \) is psd. Now define the vector \( z_i' = [x_i' \ w_i'] \) and consider the probability limit of its second moments:

\[
C = \text{plim} \left( \frac{Z'Z}{n} \right) = \begin{bmatrix} Q & A' \\ A & B \end{bmatrix}
\]

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 \( \theta \); 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 \( \hat{\theta}(\lambda) = \lambda \hat{\theta} + (1 - \lambda) \tilde{\theta} \), where \( \lambda \in \mathbb{R} \). Obviously, \( \hat{\theta}(\lambda) \) is also a consistent estimator for any \( \lambda \): 

\[
\hat{\theta}(\lambda) \xrightarrow{p} \lambda \theta + (1 - \lambda) \tilde{\theta} = \theta.
\]

Its asymptotic variance is

\[
\text{AV}[\hat{\theta}(\lambda)] = \begin{pmatrix} \lambda & 1 - \lambda \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} \lambda \\ 1 - \lambda \end{pmatrix} = \lambda^2 a + 2\lambda(1 - \lambda)b + (1 - \lambda)^2 c.
\]

If you choose \( \lambda \) so that \( \text{AV}[\hat{\theta}(\lambda)] \) is minimised, you get

\[
\lambda^* = \frac{c - b}{a - 2b + c};
\]

Now, if \( \hat{\theta} \) is efficient, the optimal value of \( \lambda^* \) must be 1, because \( \hat{\theta}(\lambda^*) \) 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

\[
\text{AV}\begin{pmatrix} \hat{\theta} \\ \tilde{\theta} \end{pmatrix} = \begin{pmatrix} a & a \\ a & c \end{pmatrix}.
\]

so

\[
\text{AV} [\hat{\theta} - \tilde{\theta}] = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} a & a \\ a & c \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = c - a = \text{AV} [\hat{\theta}] - \text{AV} [\tilde{\theta}].
\]

### 6.A.4 Hansl script for the weak instrument simulation study

```hansl
function matrix run_experiment (scalar pi)
    # function to generate the simulations
    scalar rho = 0.75       # correlation between 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()
        series u = rho * eps + sqrt(1 - rho^2) * normal()  # correlated with eps by construction
        series x = pi * w + u  # generate x via the first-stage 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

cnamest(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
Bibliography


