# Gli OLS come statistica descrittiva

Cos'è una statistica descrittiva? È una funzione dei dati che fornisce una sintesi su un particolare aspetto dei dati che a noi interessa; naturalmente, è auspicabile che questa sintesi sia quanto più informativa possibile. L'idea che motiva l'uso delle statistiche descrittive è grosso modo questa: vogliamo studiare un fenomeno, ed abbiamo dei dati; questi dati, però, sono "tanti", e non abbiamo tempo/voglia/modo di guardarli tutti. Cerchiamo allora una funzione di questi dati che, una volta calcolata, ci dica quel che vogliamo sapere, senza appesantirci con dettagli non necessari.

L'esempio più ovvio di statistica descrittiva è la media aritmetica, che ogni studente sa calcolare, se non altro per l'attenzione maniacale che riserva al proprio libretto. Dato un vettore colonna $\mathbf{y}$ di dimensione $T$, la media aritmetica non è che

$$\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} y_t = \frac{1}{T}\boldsymbol{\iota}'\mathbf{y} \tag{1}$$

La notazione con la sommatoria sarà probabilmente più familiare ai più; io, però, userò di più la seconda per la sua maggiore concisione. Per convenzione, indichiamo con $\boldsymbol{\iota}$ un vettore colonna i cui elementi sono tutti pari a 1.

Vediamo come possiamo motivare l'uso della media aritmetica. Come ho già detto, noi vorremmo poter usare una statistica descrittiva, che provvisoriamente chiamerò $\beta$, come sintesi dell'informazione contenuta nell'intero campione. Se ci mettiamo nell'ottica di usare $\beta$ — che, a questo stadio del ragionamento, non è ancora la media aritmetica — come "Bignami" del campione completo, è naturale chiedersi quanta e quale sia l'informazione che perdiamo. Vediamo: se di un campione conoscessimo solo $\beta$, cosa potremmo dire su ogni singolo elemento del campione? In assenza di altre informazioni, la cosa più sensata che possiamo dire è che, per un $t$ generico, $y_t$ sarà "più o meno" uguale a $\beta$. Se dello studente Pinco Pallino sappiamo solo che ha la media del 23, alla domanda "Quanto ha preso P.P. in Storia Economica?", risponderemmo "Boh? Avrà preso ventitré.". Se poi venisse fuori che P.P. ha effettivamente preso 23, tutto bene. Se invece ha preso 30, l'abbiamo sottovalutato, e possiamo misurare la discrepanza in 7 punti.

Nella situazione ideale, in cui l'uso di $\beta$ come sintesi dei dati non provoca perdita di informazione, la discrepanza è 0 per ogni elemento del campione (Pinco Pallino ha un libretto di tutti 23). Nella situazione non ideale, si può pensare di misurare la bontà di $\beta$ tramite la dimensione degli errori, che in gergo si chiamano **residui**. Se questo criterio, che quindi è una funzione di $\beta$, è basato sulla somma dei quadrati dei residui (così da valutare equanimemente residui in difetto e in eccesso), allora parliamo di criterio dei **minimi quadrati**. L'idea è, a questo punto, di scegliere come statistica descrittiva qualla funzione dei dati che rende minimo tale criterio.

Il criterio può essere scritto come

$$C(\beta) = \sum_{t=1}^{T} (y_t - \beta)^2$$

e per trovare il minimo rispetto a $\beta$ non facciamo altro che derivare $C$ rispetto a $\beta$;

$$C'(\beta) = \sum_{t=1}^{T} \frac{\mathrm{d}\,(y_t - \beta)^2}{\mathrm{d}\beta} = -2 \sum_{t=1}^{T} (y_t - \beta)$$

Nel punto di minimo la derivata dev'essere 0, per cui

$$\sum_{t=1}^{T} (y_t - \beta) = 0$$

che implica

$$T\beta = \sum_{t=1}^{T} y_t$$

e quindi $\beta = \bar{Y}$. In notazione matriciale si faceva ancora prima:

$$C(\beta) = (\mathbf{y} - \boldsymbol{\iota}\beta)'(\mathbf{y} - \boldsymbol{\iota}\beta)$$

la derivata è

$$C'(\beta) = -2\boldsymbol{\iota}'(\mathbf{y} - \boldsymbol{\iota}\beta) = 0$$

da cui

$$\beta = (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\mathbf{y} = \bar{Y}$$

Il lettore è invitato a controllare che $\boldsymbol{\iota}'\boldsymbol{\iota} = T$. La funzione $C$ è normalmente indicata con la sigla **SSR**, dall'inglese *Sum of Squared Residuals*.

Proviamo ora a generalizzare questo ragionamento al caso in cui abbiamo, oltre ai dati contenuti nel vettore $\mathbf{y}$, anche altri dati (detti **regressori**) che fanno riferimento alle stesse unità, che possiamo raccogliere in una matrice

$\mathbf{X}$. Ad esempio noi sappiamo, per ogni esame che Pinco Pallino ha dato, non solo quanto ha preso, ma anche in quanti giorni l'ha preparato e la percentuale delle lezioni che ha frequentato; questi dati per il $t$-esimo esame stanno in un vettore $\mathbf{x}_t'$. A questo punto, la nostra sintesi deve essere una regola che ci dia un valore 'emblematico' di $y_t$ in funzione di $\mathbf{x}_t'$.

In linea di principio, questa funzione (detta funzione di **regressione**) può avere molte forme. Se però la funzione è lineare, allora il problema ha una soluzione semplice ed elegante. Se il residuo che vogliamo minimizzare è

$$e_t(\beta) = y_t - \mathbf{x}_t'\boldsymbol{\beta}$$

allora il vettore dei residui può essere scritto

$$\mathbf{e}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \tag{2}$$

cosicché la funzione criterio da minimizzare sarà

$$C(\boldsymbol{\beta}) = \mathbf{e}(\boldsymbol{\beta})'\mathbf{e}(\boldsymbol{\beta})$$

Poiché la derivata di $\mathbf{e}(\boldsymbol{\beta})$ non è che $-\mathbf{X}$, la condizione di primo ordine sarà semplicemente

$$\mathbf{X}'\mathbf{e}(\boldsymbol{\beta}) = \mathbf{0} \tag{3}$$

Il senso di questa equazione è il seguente: il vettore $\boldsymbol{\beta}$, se esiste, deve avere la proprietà di far sì che i residui siano ortogonali ai regressori.

Mettendo assieme la (2) con la (3) si ottiene un sistema di equazioni note come **equazioni normali**:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \tag{4}$$

dalle quali si ricava l'espressione per $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{5}$$

sempreché la matrice $\mathbf{X}'\mathbf{X}$ sia invertibile. Si noti che la media aritmetica può essere ottenuta come caso particolare ponendo $\mathbf{X} = \boldsymbol{\iota}$.

**Esempio 1** *Supponiamo che*

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

*Il lettore è invitato a controllare che*

$$\boldsymbol{\beta} = \begin{bmatrix} 3.5 \\ -0.5 \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} -1 \\ 0.5 \\ 0.5 \end{bmatrix}$$

*e la validità della (3).*

I coefficienti $\boldsymbol{\beta}$ ottenuti dalla (5) hanno il nome di **coefficienti OLS**, dall'inglese *Ordinary Least Squares*, ossia minimi quadrati ordinari[1].

Vorrei sottolineare che non abbiamo mai, fino ad ora, tirato in ballo alcuna affermazione di tipo probabilistico. Ciò di cui stiamo parlando è solo ed esclusivamente una statistica descrittiva, che ha la proprietà di fornire una sintesi (ottimale da un certo punto di vista) dei dati.

A questo punto, è il caso di esplorare una serie di caratteristiche della statistica $\boldsymbol{\beta}$ e di altre grandezze da essa derivate. In primo luogo, introduciamo una grandezza che, per mancanza di un termine migliore, ci rassegniamo a chiamare **y fittato** (dall'inglese *fitted*).

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} = \mathbf{y} - \mathbf{e} \tag{6}$$

L'elemento $t$-esimo di $\hat{\mathbf{y}}$, cioè $\hat{y}_t = \mathbf{x}_t'\boldsymbol{\beta}$, è il valore che, sulla base della sintesi dei dati contenuta in $\boldsymbol{\beta}$, ci aspetteremmo per $y_t$. Come dire, il 23 dell'esempio precedente, salvo il fatto che ora questo valore deriva non solo dalla conoscenza della media di Pinco Pallino, ma anche dall'ulteriore informazione che Pinco Pallino ha preparato Storia Economica in due settimane dopo averla frequentata religiosamente. A seconda del valore degli elementi di $\boldsymbol{\beta}$, il voto che ci attenderemmo potrà essere, a questo punto, 18, 24.4, o 29, o che so io.

**Esempio 2** *Coi dati dell'esempio precedente, abbiamo*

$$\hat{\mathbf{y}} = \begin{bmatrix} 3 \\ 2.5 \\ 3.5 \end{bmatrix}$$

Si noti che, data la definizione di $\boldsymbol{\beta}$, la (6) implica

$$\mathbf{e}'\hat{\mathbf{y}} = 0$$

in forza della (3). Questo ha la conseguenza immediata che

$$\mathbf{y}'\mathbf{y} = (\hat{\mathbf{y}} + \mathbf{e})'(\hat{\mathbf{y}} + \mathbf{e}) = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}$$

[1] Il senso dell'aggettivo "ordinari" diventerà chiaro più in là, quando incontreremo coefficienti di minimi quadrati non "ordinari".

Tutti gli elementi dell'espressione precedente sono positivi, perché somme di quadrati, cosicché deve valere la seguente espressione:

$$0 \leq \mathbf{e}'\mathbf{e} \leq \mathbf{y}'\mathbf{y}$$

che si può interpretare semplicemente come il fatto che la funzione SSR, che abbiamo assunto come criterio, può andare da un caso ideale ($\mathbf{e}'\mathbf{e} = 0$) a un caso che peggiore non si potrebbe ($\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y}$). È naturale, quindi, costruire un indice che ci dica a quale punto dell'intervallo fra i due estremi ci troviamo. Questo indice, che si chiama **indice $\mathbf{R^2}$**, è definito come

$$R^2 = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y}} \tag{7}$$

L'indice $R^2$ è quindi sempre compreso fra 0 e 1, e vale 1 nel caso ideale e 0 nel caso peggiore.

I vettori $\hat{\mathbf{y}}$ e $\mathbf{e}$ possono essere anche definiti usando le seguenti matrici:

$$\mathbf{P_X} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \tag{8}$$

$$\mathbf{M_X} = I - \mathbf{P_X} \tag{9}$$

cosicché

$$\hat{\mathbf{y}} = \mathbf{P_X}\mathbf{y} \tag{10}$$

$$\mathbf{e} = \mathbf{M_X}\mathbf{y} \tag{11}$$

Queste matrici, dette **matrici di proiezione** per motivi geometrici sui quali non mi dilungo, sono simmetriche e idempotenti, e quindi singolari. La loro principale caratteristica è che

$$\mathbf{P_X}\mathbf{X} = \mathbf{X}$$

ciò che implica $\mathbf{M_X}\mathbf{X} = 0$, e quindi $\mathbf{P_X}\mathbf{M_X} = \mathbf{M_X}\mathbf{P_X} = 0$.

Sebbene queste matrici siano del tutto inutili dal punto di vista computazionale (sono matrici $T \times T$), è importante familiarizzarsi con questi operatori, poiché consentono di scrivere molti risultati in forma compatta ed elegante, cosa di grande aiuto nelle dimostrazioni. Ad esempio, per dimostrare che $\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}$ basta scrivere

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\left(\mathbf{P_X} + \mathbf{M_X}\right)\mathbf{y} = \mathbf{y}'\mathbf{P_X}\mathbf{y} + \mathbf{y}'\mathbf{M_X}\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}$$

L'uso delle matrici di proiezione è fondamentale, ad esempio, nell'illustrare una caratteristica dell'indice $R^2$. L'indice $R^2$, così come l'abbiamo

definito nell'equazione (7) presenta la sgradevole caratteristica di non essere invariante ad una traslazione dell'unità di misura scelta per la $\mathbf{y}$. Poiché quest'ultima è spesso arbitraria, chiaramente non ha molto senso usare un criterio che non sia invariante. Mi spiego meglio: supponiamo di misurare $\mathbf{y}$ su una scala diversa (ad esempio, gradi Fahrenheit anziché Celsius), così da ottenere un vettore $\mathbf{z}$ definito come

$$\mathbf{z} = a\boldsymbol{\iota} + b\mathbf{y}$$

dove $a$ e $b$ sono costanti note. Ragionevolezza vuole che, se $\hat{\mathbf{y}}$ è una approssimazione di $\mathbf{y}$, la corrispondente approssimazione di $\mathbf{z}$ sia

$$\hat{\mathbf{z}} = a\boldsymbol{\iota} + b\hat{\mathbf{y}}$$

Se $\mathbf{P_X}\boldsymbol{\iota} = \boldsymbol{\iota}$, e quindi se l'intercetta fa parte di $\mathbf{X}$, si ha che l'approssimazione di $\mathbf{z}$ soddisfa perfettamente questa proprietà:

$$\hat{\mathbf{z}} = \mathbf{P_X}\mathbf{z} = a\mathbf{P_X}\boldsymbol{\iota} + b\mathbf{P_X}\mathbf{y} = a\boldsymbol{\iota} + b\hat{\mathbf{y}}$$

Inoltre,

$$\mathbf{M_X}\mathbf{z} = a\mathbf{M_X}\boldsymbol{\iota} + b\mathbf{M_X}\mathbf{y} = b\mathbf{e}$$

e quindi i residui della regressione di $\mathbf{z}$ su $\mathbf{X}$ sono gli stessi della regressione di $\mathbf{y}$ su $\mathbf{X}$, solo moltiplicati per $b$. Considerando la (7), si avrebbe che l'indice $R^2$ sarebbe immutato se $\mathbf{z}'\mathbf{z}$ fosse pari a $b^2\mathbf{y}'\mathbf{y}$. Così non è se $a \neq 0$; infatti:

$$\mathbf{z}'\mathbf{z} = a^2T + b^2\mathbf{y}'\mathbf{y} + 2ab\boldsymbol{\iota}'\mathbf{y} \neq b^2\mathbf{y}'\mathbf{y}$$

Questo problemino conduce ad usare più spesso il c.d. $R^2$ **centrato**, che è invariante a trasformazioni di questo tipo[2], e che è definito come:

$$R_c^2 = 1 - \frac{\mathbf{y}'\mathbf{M_X}\mathbf{y}}{\mathbf{y}'\mathbf{M}_\iota\mathbf{y}} \tag{12}$$

che naturalmente ha senso solo se $\mathbf{P_X}\boldsymbol{\iota} = \boldsymbol{\iota}$. Il denominatore è quello che in statistica si chiama devianza, cioè la somma dei quadrati degli scarti di $\mathbf{y}$ dalla propria media aritmetica, ossia la somma dei quadrati dei residui della regressione di $\mathbf{y}$ su $\boldsymbol{\iota}$. Sono considerazioni di questo tipo che fanno sì che praticamente ogni regressione includa l'intercetta (o una sua trasformata), così da rendere vera $\mathbf{P_X}\boldsymbol{\iota} = \boldsymbol{\iota}$. Aggiungo che l'utilizzo nella pratica dell'indice $R_c^2$ è così diffuso che di solito, quando si parla di indice $R^2$ è alla versione centrata, e non all'$R^2$ vero e proprio, che ci si riferisce.

[2]Dimostrare per esercizio.

Un'altra cosa che si vede molto bene usando le matrici di proiezione è il **teorema di Frisch-Waugh**: supponiamo di dividere le colonne di $\mathbf{X}$ in due gruppi, che chiamiamo $\mathbf{Z}$ e $\mathbf{W}$. Naturalmente, viene diviso di conseguenza anche il vettore $\boldsymbol{\beta}$, così che possiamo scrivere

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

L'applicazione della (5) produce la seguente espressione:

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

Si possono, a questo punto, ricavare $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ in funzione di $\mathbf{Z}$, $\mathbf{W}$ e $\mathbf{y}$ andando a vedere che forma ha l'inversa della matrice $\mathbf{X}'\mathbf{X}$; la cosa presenta anche un certo interesse didattico, ma c'è un modo più conciso ed elegante di recuperare il risultato che ci interessa. Consideriamo che

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

Premoltiplicando questa espressione per $\mathbf{M_W}$ si ha

$$\mathbf{M_W y} = \mathbf{M_W Z}\boldsymbol{\beta}_1 + \mathbf{e}$$

perché $\mathbf{M_W W} = 0$ e $\mathbf{M_W e} = \mathbf{e}$. Premoltiplicando ancora per $\mathbf{Z}'$ otteniamo

$$\mathbf{Z}'\mathbf{M_W y} = \mathbf{Z}'\mathbf{M_W Z}\boldsymbol{\beta}_1$$

perché $\mathbf{Z}'\mathbf{e} = 0$. Di conseguenza,

$$\boldsymbol{\beta}_1 = \left(\mathbf{Z}'\mathbf{M_W Z}\right)^{-1} \mathbf{Z}'\mathbf{M_W y} \tag{13}$$

Per ragioni di simmetria, è del tutto ovvio che risulta

$$\boldsymbol{\beta}_2 = \left(\mathbf{W}'\mathbf{M_Z W}\right)^{-1} \mathbf{W}'\mathbf{M_Z y}$$

Si noti che la (13) potrebbe anche essere scritta

$$\boldsymbol{\beta}_1 = \left[(\mathbf{Z}'\mathbf{M_W})(\mathbf{M_W Z})\right]^{-1} (\mathbf{Z}'\mathbf{M_W})(\mathbf{M_W y})$$

e quindi $\boldsymbol{\beta}_1$ è il vettore dei coefficienti della regressione che approssima i residui di $\mathbf{y}$ rispetto a $\mathbf{W}$ sui residui di $\mathbf{Z}$ rispetto a $\mathbf{W}$. Cosa ci dice questo risultato? Ci dice che i coefficienti relativi ad un gruppo di regressori misurano la risposta di $\hat{\mathbf{y}}$ **al netto** degli altri. L'esempio che si fa in genere è: l'inclusione del vettore $\boldsymbol{\iota}$ fra i regressori fa sì che i coefficienti associati agli altri regressori (chiamiamoli $\mathbf{Z}$) sono quelli che si otterebbero facendo la regressione degli scarti dalla media di $\mathbf{y}$ sugli scarti di $\mathbf{Z}$ dalla propria media. Dimostrazione: immediata, ponendo $\mathbf{W} = \boldsymbol{\iota}$.