

# Appunti di Econometria

## ARGOMENTO [1]: IL MODELLO DI REGRESSIONE LINEARE

Tommaso Nannicini – Università Bocconi

Settembre 2009

### 1 Antipasto: proprietà algebriche del metodo dei minimi quadrati

In questa parte del corso, introdurremo il nostro primo modello statistico-econometrico—il modello di regressione lineare—per stimare le relazioni tra determinate variabili nella popolazione di riferimento e per testare ipotesi su tali relazioni. Prima di addentrarci in questo campo, però, presentiamo alcune semplici proprietà di una tecnica che useremo ripetutamente: quella dei **minimi quadrati**. Si tratta di proprietà algebriche, piuttosto che statistiche, visto che per il momento non introdurremo nessuna ipotesi sulla natura dei nostri dati e sulla popolazione da cui sono stati estratti.

Assumiamo semplicemente di avere un insieme di dati sulla variabile  $y$  (per esempio, il salario) e sulle variabili  $x_2, \dots, x_k$  (per esempio, quelle elencate nell'introduzione), osservati sulle unità di osservazione  $i = 1, \dots, N$  (i lavoratori nel nostro campione). Senza preoccuparci dell'origine dei dati o della relazione generale tra queste variabili, ci proponiamo di approssimare  $y$  con una combinazione lineare delle variabili  $x_j$  ( $\beta_1 + \beta_2 x_2 + \dots + \beta_k x_k$ ). Giusto per descrivere i nostri dati e come la  $y$  varia insieme alle  $x_j$  nel campione. Per scegliere la migliore approssimazione lineare della  $y$ , ci rifacciamo appunto al metodo dei minimi quadrati, che minimizza la somma dei quadrati degli scarti tra la  $y$  osservata e quella approssimata.

#### Caso bivariato ( $k = 2$ ).

Risolvendo

$$\min_{\beta_1, \beta_2} \left[ S(\beta_1, \beta_2) = \sum_{i=1}^N (y_i - \beta_1 - \beta_2 x_i)^2 \right], \quad (1)$$

si ottengono le condizioni di primo ordine (dette anche equazioni normali del metodo dei minimi quadrati):

$$-2 \sum_{i=1}^N (y_i - \beta_1 - \beta_2 x_i) = 0, \quad (2)$$

$$-2 \sum_{i=1}^N (y_i - \beta_1 - \beta_2 x_i) x_i = 0. \quad (3)$$

Dalla (2):

$$\sum y_i - N\beta_1 - \beta_2 \sum x_i = 0$$

$$\hat{\beta}_1 = \frac{\sum y_i}{N} - \hat{\beta}_2 \frac{\sum x_i}{N} = \bar{y} - \hat{\beta}_2 \bar{x}. \quad (4)$$

Dalla (3):

$$\sum x_i y_i - \hat{\beta}_1 \sum x_i - \hat{\beta}_2 \sum x_i^2 = 0.$$

Introducendo la (4) e usando  $\sum x_i = N\bar{x}$ :

$$\begin{aligned} \sum x_i y_i - N\bar{x}\bar{y} + \hat{\beta}_2 N\bar{x}^2 - \hat{\beta}_2 \sum x_i^2 \\ \hat{\beta}_2 (\sum x_i^2 - N\bar{x}^2) = \sum x_i y_i - N\bar{x}\bar{y}. \end{aligned}$$

Da cui, ricordando che  $Var(x) = (\sum x_i^2/N) - \bar{x}^2$  e  $Cov(x, y) = (\sum x_i y_i/N) - \bar{x}\bar{y}$ , otteniamo:

$$\hat{\beta}_2 = \frac{Cov(x, y)}{Var(x)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}. \quad (5)$$

### Caso multivariato ( $k > 2$ ).

Per risolvere i minimi quadrati nel caso multivariato, introduciamo un po' di notazione matriciale. Definiamo  $X$  come la matrice  $[n \times k]$  con per riga le osservazioni di ogni unità  $i$  e per colonna l'insieme di osservazioni su ogni variabile  $j$ :

$$X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

$\mathbf{y}$  come il vettore  $[n \times 1]$  con le osservazioni sulla variabile  $y$ , e  $\beta$  come il vettore  $[k \times 1]$  contenente i  $k$  parametri da individuare<sup>1</sup>. Il problema dei minimi quadrati diventa quindi:

$$\min_{\beta} \left[ S(\beta) = (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) \right]. \quad (6)$$

Visto che:

$$S(\beta) = \mathbf{y}'\mathbf{y} + \beta'X'X\beta - \beta'X'\mathbf{y} - \mathbf{y}'X\beta = \mathbf{y}'\mathbf{y} + \beta'X'X\beta - 2\beta'X'\mathbf{y},$$

le condizioni di primo ordine danno:

$$-2X'\mathbf{y} + 2X'X\beta = \mathbf{0},$$

che rappresentano un sistema di  $k$  equazioni (le equazioni normali dei minimi quadrati). Risolvendo:

$$\hat{\beta} = (X'X)^{-1}X'\mathbf{y} \quad (7)$$

dove stiamo assumendo che la matrice  $(X'X)$  sia invertibile (questa è l'unica assunzione di cui abbiamo bisogno per il momento). Vediamo di seguito alcune proprietà algebriche importanti dei minimi quadrati.

<sup>1</sup>Di seguito, cercherò di attenermi a questa convenzione: lettere maiuscole per matrici, lettere minuscole in grassetto per vettori, lettere minuscole per scalari.

Definiamo  $\hat{y} = X\hat{\beta}$  come la  $y$  approssimata (o fittata) ed  $\mathbf{e} = \mathbf{y} - \hat{y}$  come il vettore  $[n \times 1]$  dei residui dell'approssimazione. Si definiscano altresì:

$$SQT = \sum_{i=1}^N (y_i - \bar{y})^2, \quad (8)$$

$$SQS = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2, \quad (9)$$

$$SQR = \sum_{i=1}^N (e_i - \bar{e})^2, \quad (10)$$

rispettivamente come la somma dei quadrati totale, la somma dei quadrati spiegata (dall'approssimazione lineare) e la somma dei quadrati residua.

### **Condizioni di ortogonalità.**

1.  $X'\mathbf{e} = \mathbf{0}$ . Infatti, dalle equazioni normali:  $X'\mathbf{y} - X'X\hat{\beta} = \mathbf{0} \Rightarrow X'(\mathbf{y} - X\hat{\beta}) = \mathbf{0} \Rightarrow X'\mathbf{e} = \mathbf{0}$ .
2.  $\hat{y}'\mathbf{e} = 0$ . Infatti:  $(X\hat{\beta})'\mathbf{e} = \hat{\beta}'X'\mathbf{e} = 0$ .

### **Implicazioni delle condizioni di ortogonalità.**

- a. Poiché abbiamo incluso una costante nella matrice  $X$ :  $\bar{e} = 0$ . Questo deriva direttamente dall'ortogonalità tra la prima colonna di  $X$  e il vettore dei residui, per cui  $\sum e_i = 0$ .
- b. Poiché abbiamo incluso una costante nella matrice  $X$ :  $\bar{\hat{y}} = \bar{y}$ . Questo deriva dall'implicazione (a) e dal fatto che  $y_i = \hat{y}_i + e_i$ , per cui  $\sum y_i = \sum \hat{y}_i + \sum e_i = \sum \hat{y}_i$ .
- c. In generale (costante o non costante):  $\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$ . Infatti:  $\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 2 \sum \hat{y}_i e_i$ , ma  $\sum \hat{y}_i e_i = 0$  per la seconda condizione di ortogonalità.
- d. Poiché abbiamo incluso una costante nella matrice  $X$ :  $SQT = SQS + SQR$ . Infatti, sfruttando i risultati precedenti:  $y_i = \hat{y}_i + e_i \Rightarrow y_i - \bar{y} = \hat{y}_i - \bar{y} + e_i \Rightarrow \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2 + 2 \sum e_i(\hat{y}_i - \bar{y}) \Rightarrow \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2$ .

Sulla base di questi risultati (e—teniamolo presente—grazie al fatto che abbiamo incluso una costante tra le  $x_j$ ), è utile definire il coefficiente di determinazione  $R^2$  come:

$$R^2 = \frac{SQS}{SQT} = 1 - \frac{SQR}{SQT}. \quad (11)$$

$R^2$  è compreso tra zero e uno, e fornisce una misura della qualità dell'approssimazione lineare rispetto ai dati del nostro campione. Il suo valore, infatti, cattura la frazione della variazione campionaria di  $y$  spiegata dalla variazione delle  $x_j$ . Si può dimostrare facilmente che (i) basta introdurre una  $x$  aggiuntiva per incrementare  $R^2$ , e (ii)  $R^2$  non è altro che il coefficiente di correlazione tra  $y$  e  $\hat{y}$  al quadrato (si veda il problem set 1).

## 2 Il modello statistico: ipotesi, stima, inferenza

Adesso, acquisita una certa familiarità con il metodo dei minimi quadrati per ottenere la migliore approssimazione lineare nel campione, facciamo un primo salto di qualità e introduciamo un modello statistico con lo scopo di stimare e fare **inferenza** sui veri parametri che legano la  $y$  alle  $x_j$  nella popolazione di riferimento. Di nuovo, non osserviamo l'intera popolazione, ma solo un campione  $i = 1, \dots, N$ . Questa volta, però, assumeremo l'esistenza di un modello valido per la popolazione, e su questa base useremo i  $\beta_j$  calcolati con il metodo dei minimi quadrati come **stimatori**. Uno stimatore è una regola che prende i dati del campione e restituisce una **stima** dei veri parametri che si ipotizza legano la  $y$  alle  $x_j$  nella popolazione. Sotto alcune assunzioni, ovviamente, riguardanti il modello statistico di riferimento: senza assunzioni, in econometria, non si va da nessuna parte. Di seguito, quindi, specificheremo, stimeremo e faremo inferenza con il cosiddetto modello di regressione lineare sia nel caso bivariato, sia nel caso multivariato.

### 2.1 Il modello di regressione lineare nel caso bivariato

Si ipotizzi l'esistenza di questo legame (lineare) tra la variabile  $y$  e la variabile  $x$  nella popolazione:

$$y = \beta_1 + \beta_2 x + \epsilon, \quad (12)$$

dove  $\beta_1$  e  $\beta_2$  sono due parametri (sconosciuti), e  $\epsilon$  rappresenta una variabile aleatoria definita termine di **errore**, che comprende sia elementi puramente casuali sia tutti i fattori esplicativi della  $y$  diversi da  $x$ . Per il momento, siamo pronti a fare le seguenti assunzioni sul nostro modello<sup>2</sup>.

- A1. La relazione tra  $y$  e  $x$  è **lineare**, come espresso nell'equazione (12).
- A2. C'è un po' di **variazione campionaria** nella  $x$ , per cui  $x_i \neq x_j$  per qualche  $i, j \in (1, \dots, N)$ .
- A3. Le osservazioni  $i = 1, \dots, N$  sono un **campione casuale** della popolazione. Questo implica che le  $\epsilon_i$  sono identicamente e indipendentemente distribuite, per cui  $Cov(\epsilon_i, \epsilon_j) = 0$  per  $i \neq j$ .
- A4. La media condizionata di  $\epsilon$  è nulla:  $E(\epsilon|x) = 0$  (**esogeneità** della  $x$ ).
- A5. La varianza condizionata di  $\epsilon$  è costante:  $Var(\epsilon|x) = \sigma^2$  (**omoschedasticità**).
- A6. La distribuzione di  $\epsilon$  è data da:  $\epsilon \sim N(0, \sigma^2)$  (**normalità**).

Adesso, siamo pronti per derivare alcune importanti proprietà degli stimatori dei minimi quadrati  $\hat{\beta}_1$  e  $\hat{\beta}_2$  (derivati nella sezione precedente), come stimatori dei veri parametri  $\beta_1$  e  $\beta_2$  nel modello (12). Gli stimatori dei minimi quadrati ordinari sono detti anche OLS (*Ordinary Least Squares*)<sup>3</sup>.

**Teorema 1.** *Se valgono le assunzioni da A1 ad A4,  $\hat{\beta}_1$  e  $\hat{\beta}_2$  sono stimatori corretti:  $E(\hat{\beta}_i) = \beta_i$ ,  $i = 1, 2$ .*

---

<sup>2</sup>Come discusso in classe, queste assunzioni sono molto forti e anche abbastanza implausibili in molte applicazioni empiriche, ma sono un buon punto di partenza per derivare le proprietà degli stimatori dei minimi quadrati. Inoltre, abbandonando alcune di queste assunzioni, potremo ancora usare questi stimatori con alcuni accorgimenti; abbandonandone altre, invece, avremo bisogno di introdurre stimatori diversi. Il nostro corso di introduzione all'econometria si muove grosso modo lungo queste coordinate.

<sup>3</sup>Tra parentesi, si noti che la scelta di  $\hat{\beta}_1$  e  $\hat{\beta}_2$  come stimatori può essere motivata sia dalla loro proprietà di minimizzare la somma dei residui al quadrato, sia dal fatto che le loro condizioni di ortogonalità sui residui  $e_i$  "assomigliano" da vicino alle condizioni che stiamo imponendo sugli errori  $\epsilon_i$  nella popolazione (**metodo dei momenti**). Chiusa parentesi.

*Dimostrazione:* Diamo la dimostrazione per il coefficiente di interesse  $\beta_2$ . Possiamo riscrivere il numeratore di  $\hat{\beta}_2$  come segue (si noti che  $\sum(x_i - \bar{x}) = 0$ ):

$$\begin{aligned} \sum(x_i - \bar{x})(y_i - \bar{y}) &= \sum(x_i - \bar{x})y_i = \sum(x_i - \bar{x})(\beta_1 + \beta_2 x_i + \epsilon_i) = \beta_1 \sum(x_i - \bar{x}) + \\ &+ \beta_2 \sum(x_i - \bar{x})x_i + \sum(x_i - \bar{x})\epsilon_i = \beta_2 \sum(x_i - \bar{x})^2 + \sum(x_i - \bar{x})\epsilon_i, \end{aligned}$$

dove abbiamo usato i fatti che:

$$\begin{aligned} \sum(x_i - \bar{x})(y_i - \bar{y}) &= \sum(x_i - \bar{x})y_i - \bar{y} \sum(x_i - \bar{x}) = \sum(x_i - \bar{x})y_i, \\ \sum(x_i - \bar{x})^2 &= \sum(x_i - \bar{x})x_i - \bar{x} \sum(x_i - \bar{x}) = \sum(x_i - \bar{x})x_i. \end{aligned}$$

Di conseguenza, riscritto il numeratore come sopra, abbiamo che:

$$E(\hat{\beta}_2) = E\left(\beta_2 \frac{\sum(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} + \frac{\sum(x_i - \bar{x})\epsilon_i}{\sum(x_i - \bar{x})^2}\right) = \beta_2 + \frac{\sum(x_i - \bar{x})E(\epsilon_i|x)}{\sum(x_i - \bar{x})^2} = \beta_2,$$

dove il penultimo passaggio segue dalla legge delle aspettative iterate e l'ultimo dall'assunzione A4.  $\square$

La proprietà di correttezza (o non distorsione) ci dice che, in media, il nostro stimatore centra il bersaglio: cioè, è uguale al vero parametro nella popolazione. Ma, in generale, non ci interessa soltanto centrare il bersaglio, vogliamo anche evitare di andarci troppo lontano quando sbagliamo. È quindi utile guardare alla varianza dello stimatore  $\hat{\beta}_2$ :

$$Var(\hat{\beta}_2) = \frac{Var(\sum(x_i - \bar{x})\epsilon_i)}{(\sum(x_i - \bar{x})^2)^2} = \frac{\sum(x_i - \bar{x})^2 \sigma^2}{(\sum(x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \quad (13)$$

dove il penultimo passaggio usa l'assunzione che le  $\epsilon$  siano indipendenti tra loro e con varianza costante (stiamo quindi usando l'ipotesi di omoschedasticità).

**Teorema 2.** *Se valgono le assunzioni da A1 ad A5,  $\hat{\beta}_2$  è lo stimatore con varianza minima nella classe degli stimatori lineari e corretti di  $\beta_2$  (teorema di Gauss-Markov).*

*Dimostrazione:* In via preliminare, si noti che lo stimatore OLS è una funzione lineare delle  $y_i$  con i seguenti pesi:  $\beta_2 = \sum w_i y_i$ , dove  $w_i = (x_i - \bar{x}) / \sum(x_i - \bar{x})^2$ . Si definisca quindi un generico stimatore lineare per  $\beta_2$ :  $b_2 = \sum c_i y_i$ . Abbiamo:

$$b_2 = \beta_1 \sum c_i + \beta_2 \sum c_i x_i + \sum c_i \epsilon_i.$$

Di conseguenza,  $E(b_2) = \beta_2$  solo se  $\sum c_i = 0$  e  $\sum c_i x_i = \sum c_i (x_i - \bar{x}) = 1$ . Imponiamo queste condizioni affinché  $b_2$  sia non solo lineare ma anche corretto. Ne segue che:

$$Var(b_2) = Var\left(\beta_2 + \sum c_i \epsilon_i\right) = \sigma^2 \sum c_i^2,$$

per via della condizione di omoschedasticità. Si noti che:

$$\sum c_i^2 = \sum (w_i + c_i - w_i)^2 = \sum w_i^2 + \sum (c_i - w_i)^2 + 2 \sum w_i (c_i - w_i) = \sum w_i^2 + \sum (c_i - w_i)^2,$$

dato che  $\sum w_i c_i = 1 / \sum(x_i - \bar{x})^2$  e  $\sum w_i^2 = 1 / \sum(x_i - \bar{x})^2$ . Ne segue che:

$$Var(b_2) = \sigma^2 \sum c_i^2 = \sigma^2 \sum w_i^2 + \sigma^2 \sum (c_i - w_i)^2 = Var(\hat{\beta}_2) + \sigma^2 \sum (c_i - w_i)^2.$$

Quindi abbiamo che  $Var(b_2) > Var(\hat{\beta}_2)$  a meno che  $c_i = w_i$  (come volevasi dimostrare).  $\square$

Questo risultato è spesso sintetizzato dicendo che  $\hat{\beta}_2$  è BLUE (*Best Linear Unbiased Estimator*). Non è solo quello che in media centra il bersaglio, ma anche quello che se ne allontana di meno quando sbaglia.<sup>4</sup>

Per fare inferenza e testare ipotesi sui veri coefficienti  $\beta_i$  nella popolazione dobbiamo introdurre assunzioni sulla distribuzione degli errori (a meno che non si voglia ricorrere a proprietà asintotiche, si veda il paragrafo 3). Quindi, possiamo usare l'assunzione A6 sulla normalità degli errori, per mostrare che:

$$\frac{\hat{\beta}_2 - \beta_2}{\sigma / \sqrt{\sum (x_i - \bar{x})^2}} \sim N(0, 1). \quad (14)$$

Visto che non conosciamo  $\sigma$ , dobbiamo stimarla. E si può dimostrare (lo faremo nel caso multivariato) che uno stimatore corretto è dato da:  $s^2 = \sum e_i^2 / (N - 2)$ . Per via del fatto che  $\sum e_i^2 / \sigma^2 \sim \chi_{N-2}^2$ , abbiamo<sup>5</sup>:

$$\frac{\hat{\beta}_2 - \beta_2}{s / \sqrt{\sum (x_i - \bar{x})^2}} \sim t_{N-2}. \quad (15)$$

Quindi, possiamo testare l'ipotesi nulla  $H_0 : \beta_2 = q$  (per esempio, l'ipotesi nulla di significatività statistica del coefficiente con  $q = 0$ ), sfruttando il fatto che la

$$t_{oss} = \frac{\hat{\beta}_2 - q}{SE(\hat{\beta}_2)}$$

è distribuita come una  $t$  con  $(N - 2)$  gradi di libertà se l'ipotesi nulla è vera. Sulla base di questo risultato, possiamo costruire tutte le procedure inferenziali che abbiamo visto e analizzato in classe: test a due code o a una coda con il confronto tra la  $t_{oss}$  e il valore critico della  $t$  (scelto a seconda della significatività del test)<sup>6</sup>; intervalli di confidenza (di nuovo, a seconda del livello di significatività prescelto);  $p$ -value.

## 2.2 Il modello di regressione lineare nel caso multivariato

Adesso, estendiamo i risultati appena visti al caso multivariato con più di un regressore, facendo riferimento alla notazione matriciale già introdotta. Si ipotizza l'esistenza di questo modello statistico per la popolazione:

$$\mathbf{y} = X\beta + \epsilon, \quad (16)$$

dove  $\epsilon$  è un vettore  $[N \times 1]$  di termini di errore. Vogliamo stimare il vettore di parametri  $\beta$  nella popolazione con lo stimatore dei minimi quadrati  $\hat{\beta} = (X'X)^{-1}X'y$ . Per derivare le proprietà di questi stimatori, al momento, siamo disposti a fare le seguenti assunzioni sul modello.

- A1. La relazione tra la  $\mathbf{y}$  e le  $X$  è **lineare**, come espresso nell'equazione (16).
- A2. Possiamo escludere l'esistenza di **multicollinearità perfetta**, cioè:  $\text{ranko}(X) = K < N$ .
- A3. Le osservazioni  $i = 1, \dots, N$  sono un **campione casuale** della popolazione.
- A4. La media condizionata di  $\epsilon$  è nulla:  $E(\epsilon | x_1, \dots, x_K) = 0$  (**esogeneità delle  $X$** ).

<sup>4</sup>Se siamo disposti ad aggiungere l'assunzione di normalità (A6), è possibile dimostrare che gli stimatori OLS sono il *Best Unbiased Estimator*: cioè, hanno varianza minima nella classe di tutti gli stimatori, non solo di quelli lineari.

<sup>5</sup>Si ripassino velocemente le definizioni di distribuzione normale,  $t$ ,  $\chi^2$  ed  $F$  (per esempio, nel formulario di statistica).

<sup>6</sup>Si ripassi anche il *trade-off* tra errore del tipo I ed errore del tipo II nella scelta della significatività del test.

A5. La varianza condizionata di  $\epsilon$  è costante:  $Var(\epsilon|x_1, \dots, x_K) = \sigma^2$  (**omoschedasticità**).

A6. La distribuzione di  $\epsilon$  è una **normale**.

Per via di A3 e A4, abbiamo che:  $E(\epsilon|X) = \mathbf{0}$  (**esogeneità forte**, cioè l'errore  $\epsilon_i$  è incorrelato non solo con le variabili esplicative dell'osservazione  $i$ , ma anche con quelle delle osservazioni  $j \neq i$ ). Per via di A3 e A5, la matrice di varianza e covarianza degli errori è:  $Var(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I$  (**sfericità degli errori**). Per via di A6, infine, abbiamo che:  $\epsilon \sim N(\mathbf{0}, \sigma^2 I)$ .

**Teorema 3.** Se valgono le assunzioni da A1 ad A4,  $\hat{\beta}$  è uno stimatore corretto di  $\beta$ , cioè:  $E(\hat{\beta}) = \beta$ .

*Dimostrazione:*

$$E[\hat{\beta}] = E[(X'X)^{-1}X'y] = E_{x,\epsilon}[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon] = E_x[\beta + (X'X)^{-1}X'E(\epsilon|X)] = \beta$$

dove il penultimo passaggio segue dalla legge delle aspettative iterate e l'ultimo da  $E(\epsilon|X) = \mathbf{0}$ .  $\square$

Di nuovo, oltre alla correttezza, ci interessa analizzare la varianza di  $\hat{\beta}$ . Tenendo presente che (vedi sopra)  $\hat{\beta} - \beta = (X'X)^{-1}X'\epsilon$ , abbiamo che:

$$\begin{aligned} Var[\hat{\beta}] &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] = (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} = \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

dove stiamo usando l'assunzione di omoschedasticità e assenza di autocorrelazione negli errori:  $Var(\epsilon) = E(\epsilon\epsilon') = \sigma^2 I$ . Si noti che la matrice di varianza e covarianza degli stimatori ha questa struttura:

$$Var[\hat{\beta}] = \begin{pmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) & \dots & Cov(\hat{\beta}_1, \hat{\beta}_k) \\ Cov(\hat{\beta}_2, \hat{\beta}_1) & Var(\hat{\beta}_2) & \dots & Cov(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_k, \hat{\beta}_1) & Cov(\hat{\beta}_k, \hat{\beta}_2) & \dots & Var(\hat{\beta}_k) \end{pmatrix} = \sigma^2(X'X)^{-1}.$$

**Teorema 4.** Se valgono le assunzioni da A1 ad A5, la varianza  $\sigma^2(X'X)^{-1}$  dello stimatore  $\hat{\beta}$  è minima nella classe degli stimatori lineari e corretti di  $\beta$  (teorema di Gauss-Markov).

*Dimostrazione:* Si definisca il generico stimatore lineare:  $\mathbf{b} = L\mathbf{y}$ . Si noti che:  $E[LX\beta + L\epsilon] = LX\beta$ . Affinché  $\mathbf{b}$  sia corretto, dobbiamo quindi imporre:  $LX = I_k$ . Inoltre, dato che a questo punto  $\mathbf{b} - \beta = LX\beta + L\epsilon - \beta = L\epsilon$ , abbiamo:

$$Var(\mathbf{b}) = E[(\mathbf{b} - \beta)(\mathbf{b} - \beta)'] = E[L\epsilon\epsilon'L'] = \sigma^2 LL'.$$

Si definisca la matrice:  $D = L - (X'X)^{-1}X'$ . È facile dimostrare che  $DX = 0$ . Si può anche dimostrare che  $DD'$  è una matrice semidefinita positiva<sup>7</sup>. Si noti che:

$$\begin{aligned} LL' &= [(X'X)^{-1}X' + D][(X'X)^{-1}X' + D]' = [(X'X)^{-1}X' + D][X(X'X)^{-1} + D'] = \\ &= (X'X)^{-1} + DX(X'X)^{-1} + (X'X)^{-1}X'D' + DD' = (X'X)^{-1} + DD'. \end{aligned}$$

Quindi:  $Var(\mathbf{b}) = \sigma^2(X'X)^{-1} + \sigma^2 DD' = Var(\hat{\beta}) + \sigma^2 DD' \Rightarrow Var(b_j) \geq Var(\hat{\beta}_j)$ .  $\square$

<sup>7</sup>Se  $\mathbf{x}'A\mathbf{x} \geq 0, \forall \mathbf{x} \neq \mathbf{0}$ , la matrice  $A$  è definita come semidefinita positiva.

Di nuovo, adesso che conosciamo le proprietà degli stimatori dei minimi quadrati in termini di correttezza ed efficienza, vogliamo fare inferenza sui veri parametri del modello ipotizzato per la popolazione. Per farlo, dobbiamo rendere operativa la varianza degli stimatori ottenendo una stima non distorta di  $\sigma^2$ . A tal fine, si noti che:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\beta}} = [I - X(X'X)^{-1}X']\mathbf{y} = M\mathbf{y}$$

dove  $M$  è una matrice produttrice di residui, nel senso che prende la variabile  $y$  e ci rende un vettore di residui dalla regressione di  $y$  sulle variabili contenute nelle colonne di  $X$ . È facile dimostrare (provate!) che:  $M$  è simmetrica ( $M = M'$ );  $M$  è idempotente ( $MM = M$ );  $M'M = M$  (grazie ai risultati precedenti);  $MX = 0$  (ortogonalità); ed  $\mathbf{e} = M\boldsymbol{\epsilon}$ . Questi risultati ci sono utili per dimostrare il seguente teorema.

**Teorema 5.** *Se valgono le assunzioni da A1 ad A5, la quantità  $s^2 = \mathbf{e}'\mathbf{e}/(N - K)$  è uno stimatore corretto della varianza dell'errore  $\sigma^2$ .*

*Dimostrazione:* Grazie al fatto che (i)  $M'M = M$  e che (ii) il valore atteso di uno scalare è pari al valore atteso della traccia, possiamo riscrivere:

$$E[\mathbf{e}'\mathbf{e}] = E[\boldsymbol{\epsilon}'M\boldsymbol{\epsilon}] = E[\text{tr}(\boldsymbol{\epsilon}'M\boldsymbol{\epsilon})] = E[\text{tr}(M\boldsymbol{\epsilon}\boldsymbol{\epsilon}')] = \sigma^2\text{tr}(M).$$

Ma la traccia di  $M$  è data da:

$$\text{tr}(M) = \text{tr}(I) - \text{tr}(X(X'X)^{-1}X') = N - \text{tr}((X'X)^{-1}X'X) = N - K.$$

Da cui otteniamo che  $E[\mathbf{e}'\mathbf{e}] = \sigma^2(N - K)$ . □

Se valgono le assunzioni da A1 ad A6 (inclusa la normalità degli errori, quindi),  $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1})$ ,  $\hat{\beta}_j \sim N(\beta_j, \sigma^2(X'X)^{-1}_{jj})$ ,  $\mathbf{e}'\mathbf{e}/\sigma^2 \sim \chi_{N-K}$ . Questi elementi ci servono come base delle seguenti procedure inferenziali di test delle ipotesi.

**(a) Test t su un singolo coefficiente.**

Per testare una singola ipotesi sul coefficiente di una generica variabile  $x_j$ ,  $H_0 : \beta_j = q$ , possiamo usare tutte le procedure analizzate nel caso bivariato, dato che—se l'ipotesi nulla è vera—la quantità osservata  $t_{oss} = (\hat{\beta}_j - q)/SE(\hat{\beta}_j)$  è distribuita come una variabile  $t_{N-K}$ .

**(b) Test t su una singola restrizione lineare.**

Per testare una singola ipotesi su una generica restrizione lineare,  $H_0 : \mathbf{r}'\boldsymbol{\beta} = q$ , dove  $\mathbf{r}$  è un vettore  $[K \times 1]$  opportunamente definito, possiamo usare il fatto che:  $t_{oss} = (\mathbf{r}'\hat{\boldsymbol{\beta}} - q)/SE(\mathbf{r}'\hat{\boldsymbol{\beta}}) \sim t_{N-K}$ . Su questa base, possiamo usare le procedure già viste per un test t. L'unico problema è il calcolo dello SE al denominatore della  $t_{oss}$ . Si tenga presente che spesso, per aggirare questo problema, si può riscrivere il modello in una forma che contiene direttamente  $H_0$  come ipotesi sulla significatività statistica di un singolo coefficiente (si vedano gli esempi fatti in classe o quelli sul corrispondente capitolo del Wooldridge).

**(c) Test F sulla significatività congiunta di alcuni coefficienti.**

Adesso, si assuma di volere sottoporre a test delle ipotesi:  $H_0 : \beta_{K-J+1} = \dots = \beta_K = 0$ , cioè che tutti i coefficienti di  $J$  variabili esplicative sono congiuntamente uguali a zero nella popolazione. Definiamo come modello completo ( $M_c$ ) quello che include tutti i  $K$  regressori e come modello ristretto ( $M_r$ ) quello

che include soltanto  $K - J$  regressori (escludendo le  $J$  variabili oggetto di ipotesi, quindi). Si può dimostrare che, se vale l'ipotesi nulla:

$$F = \frac{(SQR_r - SQR_c)/J}{SQR_c/(N - K)} \sim F_{J, N-K}.$$

Questo ci permette di applicare le normali procedure di test usando i valori critici della distribuzione  $F$  (che è sempre maggiore di zero). Quindi, se  $F_{oss} > F_{crit}$ , possiamo rifiutare l'ipotesi nulla.

Sfruttando il fatto che  $SQR_c = (1 - R_c^2)SQT$  e  $SQR_r = (1 - R_r^2)SQT$ , possiamo riscrivere la statistica per il test  $F$  come:

$$F = \frac{(R_c^2 - R_r^2)/J}{(1 - R_c^2)/(N - K)} \sim F_{J, N-K}.$$

Questa statistica, nel caso classico in cui si vuole testare la significatività congiunta delle  $(K - 1)$  variabili esplicative, diventa semplicemente:

$$F = \frac{R_c^2/(K - 1)}{(1 - R_c^2)/(N - K)} \sim F_{K-1, N-K}$$

visto che in questo caso il modello ristretto contiene soltanto una costante e quindi  $R_r^2 = 0$ .

Si noti che il test  $F$  sulla significatività congiunta di un insieme di variabili esplicative ci dice una cosa del tutto diversa dal quanto indicato dall' $R^2$ . E che l' $R^2$  non ci dice niente sulla bontà del modello o sull'effetto delle variabili esplicative sulla variabile spiegata. L' $R^2$  è una misura della bontà dell'approssimazione lineare: se il nostro modello è giusto, l' $R^2$  quantifica la capacità del modello di predire nel campione.

Si ricordi, infine, che un valore elevato della statistica  $F$  (che ci porta a rifiutare l'ipotesi nulla che i coefficienti non sono congiuntamente significativi) e valori bassi delle statistiche  $t$  sui singoli coefficienti (che non ci permettono di rifiutare l'ipotesi nulla di non significatività di ogni coefficiente) potrebbero essere spiegati da un problema di **multicollinearità imperfetta** nei nostri dati. A riguardo, si vedano gli esempi fatti in classe o quelli sul corrispondente capitolo del Wooldridge.

#### (d) Test $F$ su restrizioni lineari multiple.

Vediamo adesso la formulazione generale che ci permette di testare  $J$  restrizioni lineari allo stesso tempo. L'ipotesi nulla sarà:  $H_0 : R\beta = \mathbf{q}$ , dove  $R$  e  $\mathbf{q}$  sono una matrice  $[J \times K]$  e un vettore  $[J \times 1]$  opportunamente definiti. Sotto  $H_0$ :

$$R\hat{\beta} - \mathbf{q} \sim N(\mathbf{0}, \sigma^2 R(X'X)^{-1}R').$$

Ne segue che:

$$(R\hat{\beta} - \mathbf{q})'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - \mathbf{q}) \sim \chi_J^2.$$

Sfruttando il fatto che  $\mathbf{e}'\mathbf{e}/\sigma^2 \sim \chi_{N-K}$ , possiamo calcolare<sup>8</sup>:

$$\frac{(R\hat{\beta} - \mathbf{q})'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - \mathbf{q})}{\mathbf{e}'\mathbf{e}} \cdot \frac{N - K}{J} \sim F_{J, N-K}.$$

E applicare le normali procedure di test delle ipotesi.

<sup>8</sup>In generale, si noti che, se  $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$ , allora:  $\mathbf{x}'\Sigma\mathbf{x} \sim \chi_n^2$ .

### 3 Cenni sulle proprietà asintotiche degli stimatori OLS

Finora, abbiamo fatto vedere che gli stimatori OLS sono corretti (sotto le assunzioni da A1 ad A4), efficienti (cioè, con varianza minima, sotto le assunzioni da A1 ad A5), e possono essere usati per una serie di test delle ipotesi sui veri parametri della popolazione (sotto le assunzioni da A1 ad A6). Nel resto del corso, vedremo che cosa succede a questi stimatori se abbandoniamo alcune di queste ipotesi. In alcuni casi, potremo ancora utilizzarli, a patto di tenere presente l'opportunità di alcuni accorgimenti. In altri casi, invece, le loro proprietà deterioreranno a tal punto che dovremo rimpiazzarli con nuovi stimatori.

Cominciamo col chiederci che cosa avviene se rimuoviamo l'ipotesi di normalità degli errori (A6). Un'analisi approfondita di questo tema richiederebbe un livello di trattazione che va al di là degli obiettivi di questo corso. Abbandonare A6 richiede infatti di guardare a nuove proprietà degli stimatori, dette **proprietà asintotiche**, perché valide per  $N$  che tende all'infinito. Certo, anche se pare che Chuck Norris abbia contato fino all'infinito per ben due volte, il concetto di infinito non è immediatamente operativo per noi comuni mortali. A tal fine, basti dire che interpreteremo queste proprietà come valide soltanto in **grandi campioni**.

**Consistenza.** Se  $Pr(|W_n - \theta| > \epsilon) \rightarrow 0$  con  $n \rightarrow \infty$ , allora diciamo che  $W_n$  è uno stimatore consistente di  $\theta$ , ovvero:  $plim(W_n) = \theta$ . Ciò significa che la distribuzione dello stimatore si concentra via via sul vero parametro al crescere del campione. Si può dimostrare che lo stimatore  $\hat{\beta}$  è consistente (per  $\beta$ ) sotto le assunzioni da A1 ad A4. In verità, la consistenza di  $\hat{\beta}$  richiede condizioni meno stringenti di quelle che abbiamo imposto finora. Infatti, ricordando una delle formulazioni dello stimatore OLS nel caso bivariato, grazie alla **legge dei grandi numeri**, si può vedere che:

$$plim(\hat{\beta}_2) = plim\left(\beta_2 + \frac{\sum(x_i - \bar{x})\epsilon_i}{\sum(x_i - \bar{x})^2}\right) = \beta_2 + \frac{Cov(x, \epsilon)}{Var(x)},$$

dove  $Cov(x, \epsilon)$  e  $Var(x)$  si riferiscono alla popolazione. È quindi sufficiente assumere che  $x$  ed  $\epsilon$  siano incorrelate (cioè,  $Cov(x, \epsilon) = 0$  o  $E(x\epsilon) = 0$ ) per dimostrare la consistenza dello stimatore dei minimi quadrati. È facile vedere come lo stesso valga nel caso multivariato.<sup>9</sup>

**Normalità asintotica.** In aggiunta, guardando alle proprietà di  $\hat{\beta}$  per  $N$  che tende all'infinito, si può anche dimostrare che

$$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim^a N(0, 1),$$

che implica la normalità asintotica di questa statistica (si noti che è equivalente parlare di una  $t$  di Student, dato che  $N$  tende all'infinito). Di conseguenza, possiamo di nuovo usare tutte le procedure inferenziali presentate nel paragrafo precedente, anche se l'assunzione A6 non vale. L'unico accorgimento richiesto è di sapere che queste procedure (senza A6) sono valide solo su basi asintotiche, cioè solo in grandi campioni.

---

<sup>9</sup>Si noti che la precedente assunzione A4 di media condizionata nulla era più forte dell'ipotesi di assenza di correlazione, visto che la prima implica la seconda ma non viceversa. Infatti,  $E[\epsilon|x] = 0$  implica  $E[\epsilon g(x)] = 0$  per ogni funzione  $g(\cdot)$ .