

ECONOMETRIA: UNA INTRODUZIONE

(con codice in R: <http://www.r-project.org>)

Domenico Suppa*

20/07/2011

Media e varianza

La media campionaria è definita dalla formula:¹

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

Se sono noti media θ e varianza σ^2 della popolazione (e, quindi, delle v.c. X_i componenti il campione e somiglianti alla popolazione), si possono calcolare la media e la varianza della v.c. media campionaria:

$$E(\bar{X}) = \frac{1}{N} \sum_{i=1}^N E(X_i) = \frac{1}{N} N\theta = \theta$$

$$Var(\bar{X}) = \frac{1}{N^2} \sum_{i=1}^N Var(X_i) = \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N}$$

Se la popolazione ha una distribuzione normale, per la proprietà riproduttiva delle v.c. normali ed indipendenti sottoposte ad una combinazione lineare (come la media), anche la media campionaria avrà una distribuzione normale $N(\theta, \frac{\sigma^2}{n})$. Nel caso meno restrittivo che la popolazione si distribuisca come una qualunque v.c. con varianza finita, la media campionaria sarà data comunque dalla combinazione lineare di v.c. *indipendenti ed identicamente distribuite*. In tal caso, applicando

*Università di Napoli "Federico II"

¹Per quanto segue v. Piccolo D., 1998, *Statistica*, Il Mulino, Bologna, pp. 507 e ss.

il teorema limite centrale di Lindeberg e Lévy, per $n \rightarrow \infty$, la distribuzione della media campionaria sarà ben approssimata da una distribuzione normale con media uguale alla media della popolazione e varianza pari a quella della popolazione divisa per la numerosità del campione.

Riguardo, in generale, alla varianza di una qualsiasi v.c. X , si ha:

$$\sigma^2 = \text{Var}(X) = E[(X - \bar{X})^2] = E(X^2 - 2X\bar{X} + \bar{X}^2) = E(X^2) - \bar{X}^2$$

e quindi:

$$E(X^2) = \text{Var}(X) + \bar{X}^2$$

La varianza campionaria è:

$$\sigma^2 = \text{Var}(X_i) = E(X_i^2) - \bar{X}^2$$

pertanto, il valore atteso della varianza campionaria è:

$$\begin{aligned} E[\text{Var}(X_i)] &= E[E(X_i^2) - \bar{X}^2] = E(X_i^2) - E(\bar{X}^2) = \\ &= \text{Var}(X_i) + \bar{X}^2 - [\text{Var}(\bar{X}) + \bar{X}^2] = \sigma^2 - \frac{\sigma^2}{N} = \sigma^2 \frac{N-1}{N} \end{aligned}$$

da cui segue la formula della varianza campionaria non distorta:

$$S^2 = \frac{n}{n-1} \text{Var}(X_i) = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1} = \sigma^2$$

Si può dimostrare che se la popolazione dalla quale si ottiene il campione ha una distribuzione normale, allora la media campionaria e la varianza campionaria sono v.c. tra loro indipendenti.

Ricordando, inoltre, che la somma dei quadrati di v.c. normali standardizzate, tra loro indipendenti, si distribuisce come una v.c. Chi-quadrato, si può dedurre che:

$$\begin{aligned} S^2 &\sim \frac{\sigma^2}{n-1} \chi_{(n-1)}^2 \\ \frac{n-1}{\sigma^2} S^2 &\sim \chi_{(n-1)}^2 \end{aligned}$$

Il modello di regressione lineare

L'econometria si pone gli obiettivi di **spiegare** e **prevedere**; a questi due obiettivi sono volte le metodologie che seguono.

Ipotesi che caratterizzano il modello *statistico* lineare di regressione:

1. Corretta specificazione lineare del modello: $y_i = X_i\beta + \varepsilon_i \quad \forall i = 1, \dots, N$
(dove N è la dimensione del campione).
 - Linearità nei parametri β .
 - Nessuna variabile omessa.
 - Nessuna variabile esterna.
 - Gli stessi coefficienti per ogni osservazione.
2. Il campione è costituito da osservazioni, $(y_i, X_i) \sim i.i.d.$, *indipendenti ed identicamente distribuite*, (l'attenzione è posta sulla variabile endogena y_i).
In casi speciali le osservazioni possono essere *indipendenti ma non identicamente distribuite: i.n.i.d.*
3. Le *variabili esplicative* X_i possono essere **stocastiche**.
4. La matrice $X'X$ ha rango pieno (è invertibile, altrimenti non vi sono sufficienti informazioni per determinare univocamente i parametri).
5. La struttura degli errori rispetta le seguenti ipotesi:
 - La media degli errori è zero: $E(\varepsilon_i|X_i) = 0$
 - *omoschedasticità* (dove σ^2 è la varianza degli errori):
 $E(\varepsilon\varepsilon'|X) = \sigma^2 I_N$
oppure
 - *eteroschedasticità*: $E(\varepsilon\varepsilon'|X) = \text{diag}(\sigma_i^2) \quad i = 1, \dots, N$
in entrambi gli ultimi due punti i residui non devono essere autocorrelati.

Nel caso siano presenti tanto l'eteroschedasticità che l'autocorrelazione dei residui, la varianza degli stimatori dei minimi quadrati di $\hat{\beta}$ assume una forma più generale. Tenendo presente che dalla condizione di ortogonalità (implicita nel metodo dei minimi quadrati) $E[(X'\varepsilon)] = 0$:

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

ed essendo $y = X\beta + \varepsilon$, abbiamo:

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

per cui $E[\hat{\beta}] = \beta$ ed inoltre:

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

Quindi, passando ai valori medi, la matrice di varianze-covarianze dei coefficienti di regressione contiene il termine (una matrice $N \times N$) $\Omega = E[\varepsilon\varepsilon']$ da stimare. L'espressione $X'\Omega X$ viene, generalmente, denominata *sandwich* e consente di ottenere la stima *robusta* degli errori standard dei parametri e delle loro correlazioni. Quando gli errori sono omoschedastici ed incorrelati abbiamo: $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

Regressione lineare

Per studiare la regressione può essere conveniente simulare il DGP (data generating process) e poi inferire dai dati le caratteristiche del DGP che sono, in realtà, già note. Ciò permette di verificare i risultati e di constatarne la coerenza rispetto alle prescrizioni teoriche del modello di regressione lineare.

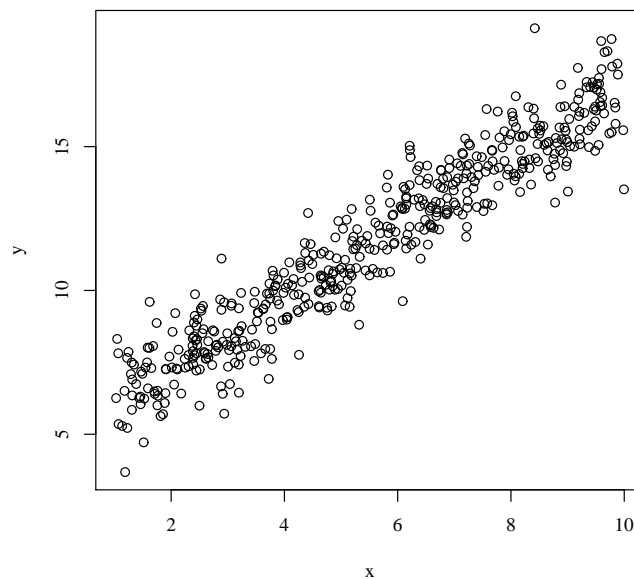


Figura 1: Scatter plot (x_i, y_i)

Poniamo che le osservazioni (y_i, x_i) siano generate da una relazione lineare con errori che si distribuiscono come una v.c. normale standardizzata $\varepsilon \sim N(0, 1)$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

dove $\beta_0 = 5$ e $\beta_1 = 1.2$ [vedi R1 e Fig. 1]. Uno dei test utilizzati per verificare la corretta specificazione funzionale del modello di regressione è il test di Ramsey (**reset test**).² Se il modello è stato specificato correttamente non si può rifiutare l'ipotesi H_0 : «la forma funzionale è ben specificata» dato il valore del p -value (che deve risultare maggiore di 0.05 per accettare H_0).

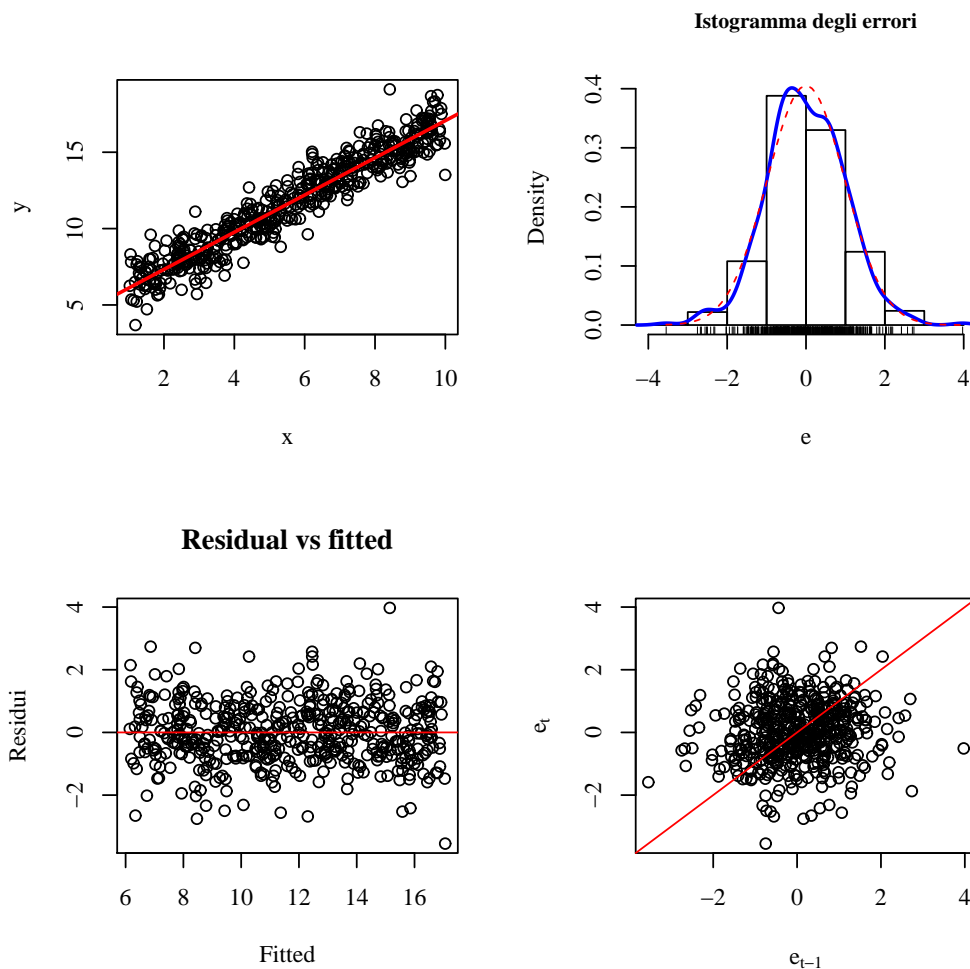


Figura 2: Grafici della regressione e dei residui

Inoltre, in base alle ipotesi sul modello di regressione, i residui dovrebbero distribuirsi in modo normale (vedi grafico in alto a destra della Fig. 2). L'analisi

²Per i test citati è necessario caricare alcune librerie - ad. es. Imtest - come specificato negli esempi.

grafica consente di avere un'idea della distribuzione degli errori e l'ipotesi di normalità può essere verificata applicando il test di **Shapiro-Wilk** sulla normalità dei residui (preventivamente standardizzati con la trasformazione $\frac{\hat{\epsilon}_i - \bar{\epsilon}}{\sigma_{\epsilon}}$), in tal caso l'ipotesi H_0 : «i residui sono distribuiti come una v.c. normale» può essere accettata solo se il p -value è sufficientemente elevato. Con risultati simili, dal grafico della

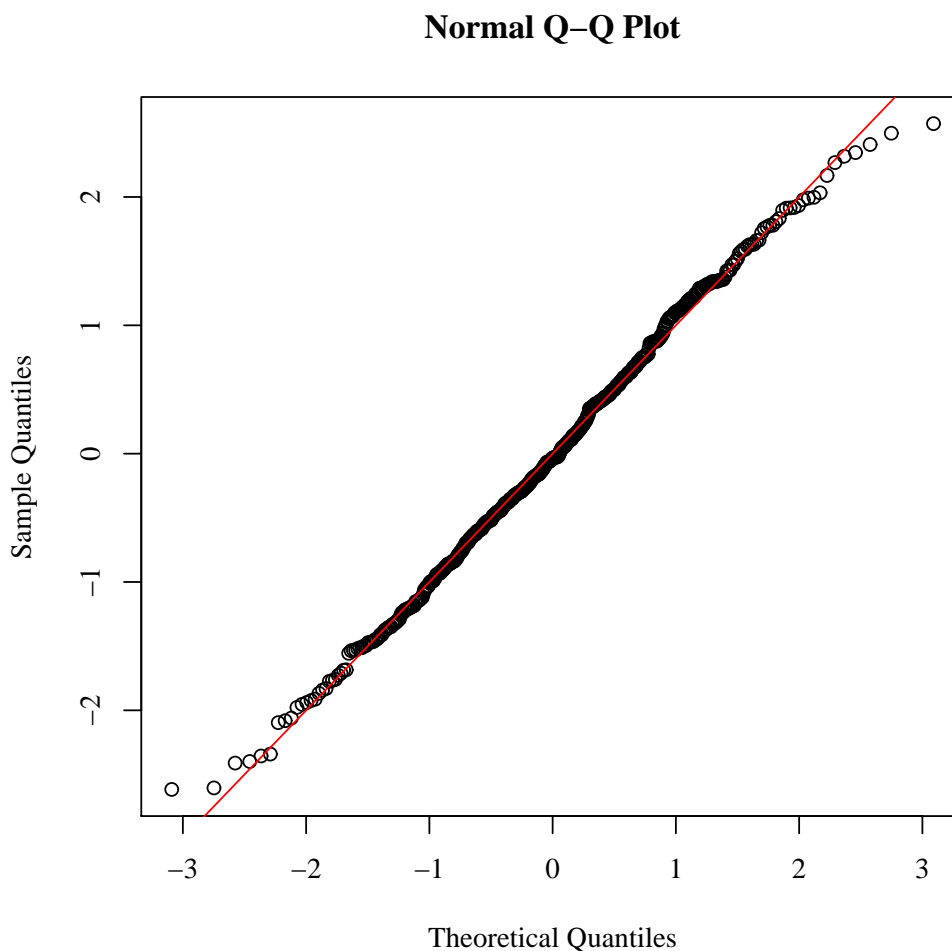


Figura 3: Adattamento dei residui alla Normale

Fig. 3 (*Normal Q-Q Plot*) si può verificare se gli errori standardizzati si distribuiscono come la normale standardizzata: i punti dovrebbero disporsi lungo la retta a 45° , se è verificata l'ipotesi di normalità. Riguardo alla corretta specificazione del modello di regressione lineare, è utile esaminare il grafico dei residui rispetto ai valori stimati della variabile dipendente: i punti dovrebbero disporsi orizzontal-

mente rispetto alle ascisse e simmetricamente rispetto alla linea orizzontale con intercetta zero (v. grafico in basso a sinistra nella Fig. 2). Lo stesso tipo di grafico può essere tracciato rispetto ad ogni variabile esplicativa: se la disposizione dei residui non è casuale ciò indica che il modello vero non è lineare rispetto a quel particolare regressore.

Per testare l'eteroschedasticità dei residui si può utilizzare il test di **Breusch-Pagan** per l'ipotesi H_0 : «la varianza degli errori è costante» (omoschedasticità) che verrà accettata se il p -value è elevato.

Soprattutto per i dati in serie storica viene utilizzato il test **Durbin-Watson** per verificare che non vi sia autocorrelazione nei residui sotto l'ipotesi H_0 : «i residui sono incorrelati» che può essere accettata se il valore del p -value è abbastanza alto. Nello stesso contesto seriale è possibile utilizzare la *funzione di autocorrelazione totale (acf)* e la *funzione di autocorrelazione parziale (pacf)* per inferire sul processo ARMA che potrebbe aver generato i residui.

In tutti i software statistici l'output del modello di regressione prevede il calcolo dell'*indice di determinazione* $R^2 = 1 - \frac{Var(e)}{Var(y)}$ e della sua versione *corretta* (non distorta): $R_c^2 = 1 - \frac{(N-1)Var(e)}{(N-k-1)Var(y)}$ (dove k è il numero delle variabili esplicative). L'indice R^2 (come la sua versione corretta) indica quanta parte della variabilità complessiva (*TSS: Total Sum of Squares*) della variabile dipendente y è spiegata dalla regressione ed è, quindi, basato sulla decomposizione della varianza: la variabilità totale è uguale per definizione alla somma della variabilità della regressione (*ESS: Explained Sum of Squares* della stima \hat{y}) e della variabilità degli errori (*RSS: Residual Sum of Squares*).

Un test globale su tutti i parametri del modello di regressione può essere condotto con il test **F di Fisher** che confronta la variabilità della stima \hat{y} con la variabilità dei residui \hat{e} . L'ipotesi nulla è H_0 : «le variabili esplicative non spiegano la variabile dipendente» (tutti i coefficienti del modello non sono significativamente diversi da zero). Per rifiutare tale ipotesi (ed attribuire qualche validità al modello) è necessario che la $F_{teorica}$, ottenuta in corrispondenza di k e $n - k - 1$ gradi di libertà, sia significativamente diversa da (maggiore di) zero (e cioè che il p -value sia sufficientemente piccolo).

Il test sui singoli coefficienti $\hat{\beta}_j$ formula l'ipotesi H_0 : «il coefficiente $\hat{\beta}_j$ è, in media, pari a zero» e confronta la statistica test $\hat{t}_j = \frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}$ con la $t_{teorica}$ che si distribuisce come la variabile casuale **t di Student** ed è ben approssimata dalla normale standardizzata al crescere della numerosità delle osservazioni. La $t_{calcolata}$ deve essere significativamente maggiore della $t_{teorica}$ e, quindi, il p -value deve essere abbastanza piccolo per poter rifiutare H_0 . Utili informazioni al riguardo possono essere dedotte anche dagli *intervalli di confidenza* che scaturiscono da questo test (ad es. la posizione dello zero).

Codice in R

In sequenza, cercando di rispettare l'ordine degli argomenti trattati nel testo:

R1

```
#Il DGP dell'esempio
n<-500                # numerosità del campione
e<-rnorm(n,0,1)      # genera una v.c. normale
x<-runif(n,1,10)     # genera la variabile esplicativa x
y=5+1.2*x+e          # genera la variabile dipendente y
plot(x,y)            # il grafico
cor.test(x,y,method="pearson") # test sulla correlazione
```

R2

```
# La stima del modello
library(lmtest)      # carica la libreria lmtest
r<-lm(y ~ x)        # stima del modello lineare
summary(r)          # mostra i risultati della stima
confint(r)          # intervalli di confidenza dei coefficienti
names(r)            # mostra le informazioni contenute in r
e<-residuals(r)     # salva i residui stimati in e
resettest(r)        # reset test sulla regressione
t.test(e)           # test sulla media dei residui
shapiro.test(
  (e-mean(e))/sd(e)) # test di normalità dei residui
```

R3: analisi grafica

```
# Regressione e distribuzione errori
par(mfrow=c(2,2))    # divide il grafico
plot(x,y)            # nube dei punti
abline(r, col="red", lwd=2) # linea di regressione
hist(e,freq=F, ylim=c(0,.45),
      main="Istogramma",cex.main=.92) # istogramma
rug(e,lwd=.25,side=1) # evidenziare la concentrazione
lines(density(e), lwd=2,
      col="blue")    # densità della distribuzione
z<-seq(min(e), max(e),.1) # genera il supporto
lines(z,dnorm(z, mean=mean(e), sd=sd(e)),
      col="red", lty=2) # curva della normale standardizzata
plot(fitted(r), e, ylab="Residui",
```

```

      xlab="Fitted",
      main=
        "Residual vs fitted") # grafico errori verso y stimato
abline(h=0,
       col="red") # linea sulla media (zero) dei residui
plot(e[-n], e[-1], xlab=expression(e[t-1]),
     ylab=
       expression(e[t])) # grafico della correlazione dei residui
abline(0,1,col="red") # bisettrice
par(mfrow=c(1,1))      # ripristina la finestra del grafico
qqnorm((e-mean(e))/sd(e))
  abline(0,1, col="red")      # Normal Q-Q plot
plot(x, e, ylab="Residui", xlab="x", main="Residual vs x");
  abline(h=0, col="red")      # grafico errori rispetto ad x
pairs(data.frame(x, fitted(r), r$residuals)) # grafico pairs

```

R4: Test sulla regressione e stima robusta

```

# Altri test sulla regressione
vcov(r)      # matrice var-covarianze dei coefficienti
bptest(r)    # Breusch-Pagan test (omoschedasticità residui)
dwtest(r)    # test di Durbin-Watson (autocorrelazione residui)
library(lmtest) # carica la libreria lmtest
coeftest(r)   # test sui coefficienti
coeftest(r, df=Inf,
         vcov=vcovHAC(r)) # test robusto sui coefficienti
library(sandwich) # carica la libreria sandwich
print(vcovHC(r)) # varcov dei parametri con eteroschedasticità
print(vcovHAC(r)) # con eteroschedasticità ed autocorrelazione
library(robust) # libreria per la stima robusta
lmRob(y ~ x)    # regressione robusta
summary(lmRob(y ~ x)) # riepilogo dei risultati della stima

```

Ulteriori comandi in R:

Per salvare il grafico corrente nel file scatterxy in formato eps:

```

par(family="Times") # da porre all'inizio dello script
dev.copy(postscript, file="scatterxy.eps", height=6, width=6,
         horizontal=F, onefile=F, fonts="Times")
dev.off()

```