Proprietà dei coefficienti di regressione

Psicometria 1 - Lezione 15
Lucidi presentati a lezione

AA 2000/2001 dott. Corrado Caudek

I coefficienti di regressione possono essere considerati delle *variabili aleatorie* in quanto assumono valori diversi in campioni diversi.

Poniamoci ora il problema di determinare le proprietà della *distribuzione campionaria* di A e B calcolati con il metodo dei minimi quadrati.

Valore atteso di b

$$B = \frac{S_{XY}}{S_X^2} = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2}$$

$$= \frac{\sum \left(X_i - \overline{X}\right)Y_i - \overline{Y}\sum \left(X_i - \overline{X}\right)}{\sum \left(X_i - \overline{X}\right)^2}$$

$$= \frac{\sum \left(X_i - \overline{X}\right)Y_i}{\sum \left(X_i - \overline{X}\right)^2}$$

$$E(B) = E\left(\frac{\sum (X_i - \overline{X}) Y_i}{\sum (X_i - \overline{X})^2} \right)$$

$$= \frac{\sum (X_i - \overline{X}) E(Y_i)}{\sum (X_i - \overline{X})^2}$$

$$= \frac{\sum (X_i - \overline{X})(\boldsymbol{a} + \boldsymbol{b} X_i)}{\sum (X_i - \overline{X})^2}$$

$$= \boldsymbol{a} \frac{\sum \left(X_i - \overline{X} \right)}{\sum \left(X_i - \overline{X} \right)^2} + \boldsymbol{b} \frac{\sum \left(X_i - \overline{X} \right) X_i}{\sum \left(X_i - \overline{X} \right)^2}$$

$$= 0 + \boldsymbol{b} \frac{\sum \left(X_i - \overline{X} \right) X_i}{\sum \left(X_i - \overline{X} \right)^2}$$

$$= \boldsymbol{b} \frac{\sum X_i^2 - \overline{X} \sum X_i}{\sum \left(X_i - \overline{X} \right)^2}$$

$$= \boldsymbol{b} \frac{\sum X_i^2 - \overline{X} n \overline{X}}{\sum \left(X_i - \overline{X}\right)^2} = \boldsymbol{b} \frac{\sum X_i^2 - n\overline{X}^2}{\sum \left(X_i - \overline{X}\right)^2} = \boldsymbol{b}$$

dato che

$$\sum \left(X_i - \overline{X}\right)^2 = \sum X_i^2 + \sum \overline{X}^2 - 2\overline{X} \sum X_i$$

$$= \sum X_i^2 + n\overline{X}^2 - 2\overline{X} n \overline{X}$$

$$= \sum X_i^2 + n\overline{X}^2 - 2n\overline{X}^2 = \sum X_i^2 - n\overline{X}^2$$

Abbiamo così dimostrato che B è uno stimatore privo di errore sistematico di $\boldsymbol{b}$: $E(B)=\boldsymbol{b}$.

In modo analogo si può mostrare che $E(A)=\boldsymbol{a}$.

Varianza di *b*

In base alle assunzioni del modello della regressione bivariata le osservazioni $Y_1, Y_2, \ldots, Y_n$ sono indipendenti e quindi

$$V(B) = V\left(\frac{\sum (X_i - \overline{X}) Y_i}{\sum (X_i - \overline{X})^2} \right)$$

$$= \left(\frac{1}{\sum (X_i - \overline{X})^2} \right)^2 \sum V\left((X_i - \overline{X}) Y_i \right)$$

$$= \left(\frac{1}{\sum \left(X_i - \overline{X} \right)^2} \right)^2 \sum \left(X_i - \overline{X} \right)^2 V\left(Y_i \right)$$

Dato che $V(Y_i) = \boldsymbol{s}_{\boldsymbol{e}}^2$ per $i = 1, 2, \ldots, n$, la varianza di B risulta dunque essere uguale a

$$V(B) = \frac{\boldsymbol{s}_{\boldsymbol{e}}^2}{\sum \left(X_i - \overline{X} \right)^2}$$

$$V(B) = \frac{\boldsymbol{s}_{\boldsymbol{e}}^{2}}{\sum \left(X_i - \overline{X}\right)^2}$$

La varianza di *B* è dunque piccola quando:

(i) la varianza dei residui è piccola

(ii) (*ii*) la varianza di *X* è grande.

In modo analogo è possibile trovare la varianza della distribuzione campionaria di *A*:

$$V(A) = \frac{\boldsymbol{s}_{\boldsymbol{e}}^{2} \sum X_i^2}{n \sum \left(X_i - \overline{X} \right)^2}$$

Dato che A e B sono funzioni lineari di Y_i, se le osservazioni Y_i sono distribuite normalmente, allora anche A e B lo saranno.

$$B \sim N\left(\boldsymbol{b}, \frac{\boldsymbol{s}_{\boldsymbol{e}}^{2}}{\sum\left(X_i - \overline{X}\right)^2}\right)$$

$$A \sim N\left(\boldsymbol{a}, \frac{\boldsymbol{s}_{\boldsymbol{e}}^{2}\sum X_i^2}{n\sum\left(X_i - \overline{X}\right)^2}\right)$$

Verifica di ipotesi relative ai coefficienti di regressione

Secondo problema dell'analisi della regressione: come si esegue un test statistico a proposito dei coefficienti di regressione?

Supponiamo, che l'ipotesi nulla sia:

$$H_0 : \boldsymbol{b} = \boldsymbol{b}_0$$

dove $\boldsymbol{b}_0$ è una costante (spesso $\boldsymbol{b}_0 = 0$).

Se la varianza degli errori nella popolazione $\boldsymbol{s}_{\boldsymbol{e}}^{2}$ fosse conosciuta, l'ipotesi nulla potrebbe essere sottoposta a verifica usando la statistica

$$z = \frac{B - \boldsymbol{b}_0}{\sqrt{\dfrac{\boldsymbol{s}_{\boldsymbol{e}}^{2}}{\sum \left(X_i - \overline{X}\right)^2}}}$$

con $z \sim N(0,1)$.

Solitamente, però, il parametro σ_e^2 non è conosciuto e deve essere stimata sulla base delle osservazioni del campione.

Può essere dimostrato che uno stimatore privo di errore sistematico di σ_e^2 è fornito da

$$\frac{SQ_{ERR}}{n-2} = \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

laddove il valore 2 presente al denominatore corrisponde al numero di parametri stimati dal modello (α, β) per stimare il valore atteso di Y.

Se la varianza degli errori σ_e^2 viene stimata sulla base dei dati del campione, dunque, l'ipotesi nulla $H_0: \beta = \beta_0$ può essere sottoposta a verifica usando la statistica

$$t = \frac{B - \beta_0}{\sqrt{\frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{(n-2)\sum \left(X_j - \overline{X}\right)^2}}}$$

distribuita come t con (n - 2) gradi di libertà.

Si noti che il valore della statistica cresce quando:

(*i*)la varianza dei residui è piccola,

(*ii*) la varianza di X è grande,

(*iii*) le dimensioni del campione sono grandi.

Una procedura simile a quella sopra descritta viene usata per le inferenze riguardanti l'intercetta.

Coefficiente di determinazione

Il terzo problema dell'analisi della regressione è quello di stabilire quanto precisamente la retta di regressione si approssima ai dati.

La risposta a questa domanda ci viene data dal coefficiente di determinazione.

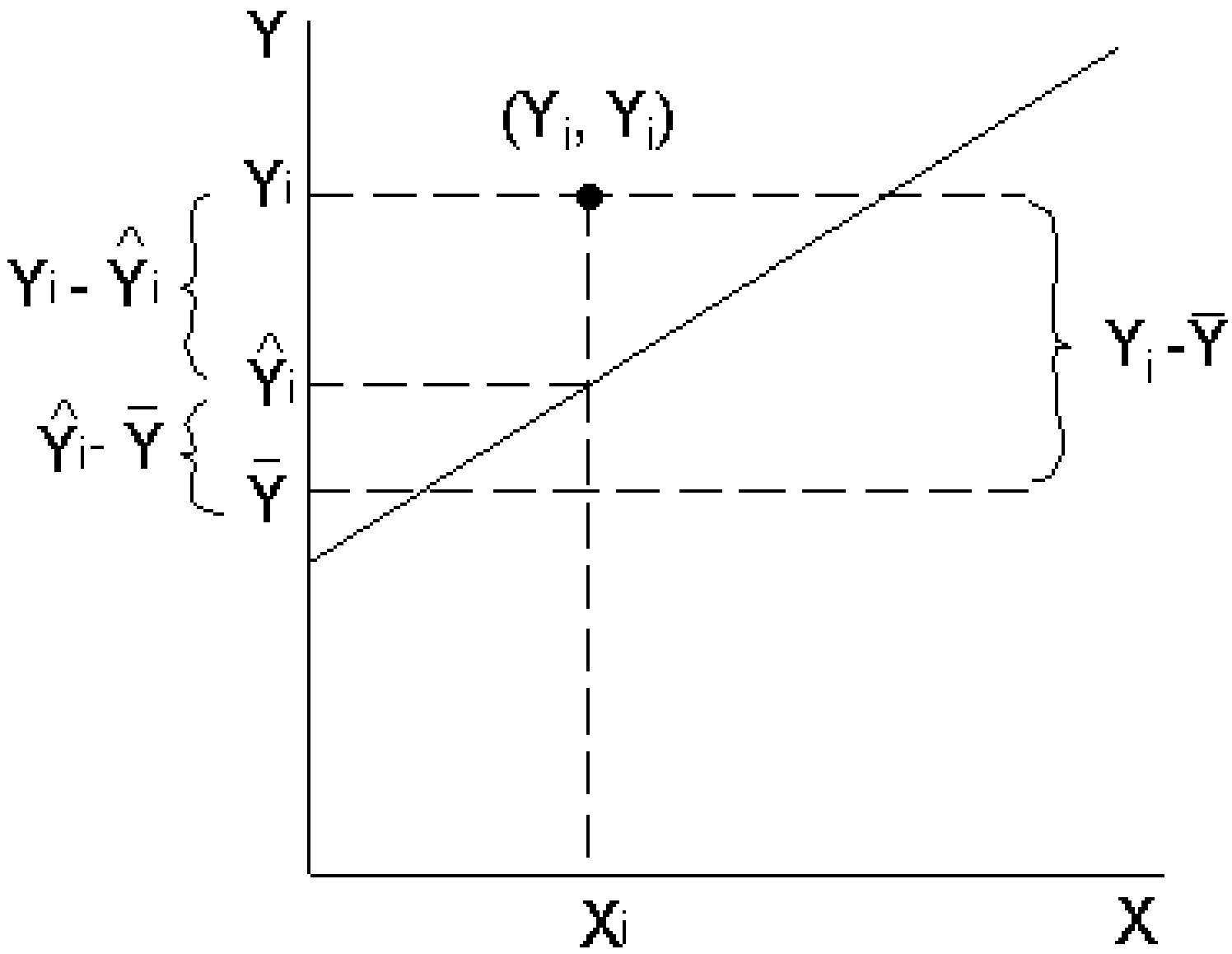
Iniziamo con il definire la *somma dei quadrati totale* (SQ_{TOT}) , ovvero la somma dei quadrati degli scostamenti di ciascuna osservazione Y_i dalla media $\overline{Y}$:

$$SQ_{TOT} = \sum \left(Y_i - \overline{Y} \right)^2$$

Lo scostamento di ciascun punteggio Y_i dalla media $\overline{Y}$ può essere scisso in due componenti:

$$Y_i - \overline{Y} = \left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

La prima componente è uguale allo scostamento tra il punteggio predetto $\hat{Y}_i$ e il punteggio medio $\overline{Y}$;
la seconda componente è uguale allo scostamento tra il punteggio osservato Y_i e il punteggio predetto $\hat{Y}_i$.


Y
(Y_i, Y_i)
Y_i
$Y_i - \hat{Y}_i$
$\hat{Y}_i$
$\hat{Y}_i - \bar{Y}$
$\bar{Y}$
$Y_i - \bar{Y}$
X_i
X

$$SQ_{TOT} = \sum \left(\left(\hat{Y}_i - \bar{Y} \right) + \left(Y_i - \hat{Y}_i \right) \right)^2$$

$$= \sum \left(\hat{Y}_i - \bar{Y} \right)^2 + \sum \left(Y_i - \hat{Y}_i \right)^2 + 2 \sum \left(\hat{Y}_i - \bar{Y} \right) \left(Y_i - \hat{Y}_i \right)$$

L’ultimo termine dell’equazione precedente è uguale a zero.

$$\sum \left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right) =$$

$$= \sum \left(\hat{Y}_i Y_i - \hat{Y}_i^2 - \bar{Y} Y_i + \bar{Y}\hat{Y}_i\right)$$

$$= \sum \left(\hat{Y}_i Y_i - \hat{Y}_i^2\right) - \sum \left(\bar{Y} Y_i - \bar{Y}\hat{Y}_i\right)$$

$$= \sum \left(\hat{Y}_i \left(Y_i - \hat{Y}_i\right)\right) - \bar{Y} \sum \left(Y_i - \hat{Y}_i\right)$$

$$= \sum \left(\hat{Y}_i E_i \right) - \bar{Y} \sum E_i$$

$$= \sum \left(\hat{Y}_i E_i \right)$$

$$= \sum \left(A + BX_i \right) E_i$$

$$= A \sum E_i + B \sum E_i X_i$$

$$= B\sum E_i X_i = 0$$

dato che

$$\sum X_i E_i = \sum X_i \left(Y_i - A - BX_i\right)$$

$$= \sum X_i Y_i - A\sum X_i - B\sum X_i^2 = 0$$

L'espressione precedente è uguale a zero in quanto *i coefficienti di regressione sono stati calcolati ponendo questa espressione uguale a zero*.

In conclusione, la somma totale dei quadrati può essere espressa come la somma di due componenti, una componente "spiegata" dalla regressione e una componente "non spiegata" di errore:

$$\sum \left(Y_i - \overline{Y}\right)^2 = \sum \left(\hat{Y}_i - \overline{Y}\right)^2 + \sum \left(Y_i - \hat{Y}_i\right)^2$$

$$SQ_{TOT} = SQ_{REG} + SQ_{ERR}$$

Queste due componenti consentono di definire l'indice r^2_{XY} che misura la dispersione delle osservazioni attorno alla retta di regressione:

$$r^2_{XY} \equiv \frac{SQ_{REG}}{SQ_{TOT}}$$

L'indice r_{XY}^2 è chiamato *coefficiente di determinazione* e rappresenta la porzione della variazione in Y spiegata dalla regressione su X.

Se c'è una perfetta relazione lineare tra X e Y, tutte le osservazioni cadono sulla retta di regressione. In queste circostanze $SQ_{REG} = SQ_{TOT}$, nessun errore viene commesso nella predizione di Y a partire da X ($SQ_{ERR} = 0$) e $r_{XY}^2 = 1$.

Se non c'è relazione lineare tra X e Y, la dispersione delle osservazioni attorno alla retta di regressione è massima e la retta di regressione ha pendenza 0.

In queste circostanze, $SQ_{REG} = 0$, $SQ_{ERR} = SQ_{TOT}$ e $r_{XY}^2 = 0$.

Tra questi estremi, la grandezza di r_{XY}^2 indica il grado di dispersione delle osservazioni del campione attorno alla retta di regressione, ovvero *la porzione della varianza di Y che è predicibile in base alla relazione lineare con X*.

Coefficiente di correlazione e coefficiente di regressione

L’esame del coefficiente di correlazione nell’ambito del modello della regressione bivariata consente di attribuire al coefficiente di correlazione la sua interpretazione più semplice.

Il coefficiente di correlazione è definito come la media dei prodotti delle variabili standardizzate X e Y:

$$r_{XY} \equiv \frac{\sum z_{X_i} z_{Y_i}}{n}$$

$$r_{XY} = \frac{\sum \left(X_i - \bar{X} \right) \sum \left(Y_i - \bar{Y} \right)}{n S_X S_Y}$$

$$= \frac{S_{XY}}{S_X S_Y}$$

il coefficiente di correlazione può essere concepito come una covarianza standardizzata, ovvero come il rapporto tra la covarianza S_{XY} e le deviazioni standard S_X e S_Y.

In precedenza abbiamo definito il coefficiente B come:

$$B = \frac{S_{XY}}{S_X^2}$$

Dato che $S_{XY} = r_{XY} S_X S_Y$, possiamo dunque scrivere:

$$B = \frac{r_{XY} S_X S_Y}{S_X^2} = r_{XY} \frac{S_Y}{S_X}$$

Che cosa significa questo risultato se le variabili X e Y vengono standardizzate?

Il modello della regressione bivariata è

$$Y_i = A + BX_i + E_i$$

Dato che $\bar{Y} = A + B\bar{X}$, possiamo scrivere

$$Y_i = \left(\bar{Y} - B\bar{X}\right) + BX_i + E_i$$

$$Y_i - \bar{Y} = -B\bar{X} + BX_i + E_i$$

$$Y_i - \bar{Y} = B\left(X_i - \bar{X}\right) + E_i$$

$$Y_i - \bar{Y} = r_{XY}\frac{S_Y}{S_X}\left(X_i - \bar{X}\right) + E_i$$

$$\frac{Y_i - \bar{Y}}{S_Y} = r_{XY}\frac{\left(X_i - \bar{X}\right)}{S_X} + \frac{E_i}{S_Y}$$

$$z_{Y_i} = r_{XY} z_{X_i} + E_i^*$$

E’ dunque possibile concludere che, standardizzando le variabili X e Y, il coefficiente di correlazione diventa uguale al coefficiente di regressione.

Quando il coefficiente di correlazione è uguale a 0, z_{Y_i} non può essere in nessun modo predetto conoscendo z_{X_i} dato che la retta di regressione è piatta.

Quando il coefficiente di correlazione è uguale a 1, z_{Y_i} è completamente predetto da z_{X_i} dato che tutte le osservazioni giacciono sulla retta di regressione.

Coefficiente determinazione e coefficiente di correlazione

Dimostreremo ora che il coefficiente di determinazione è uguale al quadrato del coefficiente di correlazione.

I valori standardizzati z_{X_i} e z_{Y_i} sono uguali a

$$z_{X_i} = \frac{X_i - \bar{X}}{S_X} \qquad z_{Y_i} = \frac{Y_i - \bar{Y}}{S_Y}$$

e, per definizione, hanno varianza unitaria.

I punteggi predetti standardizzati hanno una varianza uguale a

$$S^2_{z_{\hat{Y}}} = \frac{\sum \left(z_{\hat{Y}_i} \right)^2}{n} - \left(\frac{\sum z_{\hat{Y}_i}}{n} \right)^2$$

Il secondo termine dell'equazione precedente si riduce a zero

$$\frac{1}{n}\sum z_{\hat{Y}_i} = \frac{1}{n} r_{XY} \sum z_{X_i} = 0$$

in quanto $\sum z_{X_i} = 0$.

Il primo termine è invece uguale al quadrato del coefficiente di correlazione

$$\sum \frac{\left(z_{\hat{Y}_i}\right)^2}{n} = \sum \frac{r_{XY}^2 z_{X_i}^2}{n} = r_{XY}^2$$

dato che

$$\sum \frac{z_{X_i}^2}{n} = \frac{\sum \left(\frac{X_i - \overline{X}}{S_X} \right)^2}{n} = \frac{1}{S_X^2} \frac{\sum \left(X_i - \overline{X}\right)^2}{n} = 1$$

In conclusione, la varianza dei punteggi predetti standardizzati è quindi uguale al quadrato del coefficiente di correlazione:

$$S^2_{z_{\hat{Y}}} = r^2_{XY}$$

Dato che la varianza della variabile dipendente standardizzata è uguale a 1.0, il rapporto tra $S^2_{z_{\hat{Y}}}$ e $S^2_{z_Y}$ risulterà anch'esso essere uguale al quadrato del coefficiente di correlazione:

$$\frac{S^2_{z_{\hat{Y}}}}{S^2_{z_Y}} = r^2_{XY}$$

Risulta così dimostrato che il coefficiente di determinazione è uguale coefficiente di correlazione innalzato al quadrato:

$$\frac{S^2_{z_{\hat{Y}}}}{S^2_{z_Y}} = \frac{\sum \frac{\left(\hat{Y}_i - \bar{Y}\right)^2}{n}}{\sum \frac{\left(Y_i - \bar{Y}\right)^2}{n}} = \frac{SQ_{REG}}{SQ_{TOT}} = r^2_{XY}$$

```
% REGRESSIONE SEMPLICE - PSICOMETRIA 1, Marzo 2001

% dati
data=[1 2 12
      1 3 21
      1 5 25
      1 7 39
      1 9 41];

data

% vettore con un vettore unitario e la variabile X
x=data(1:5,2);

% vettore (20x2) che contiene la variabile Y
y=data(1:5,3);

% vettore unitario
uno=data(1:5,1);

% calcolo della media di X
xmed=(x'*uno)/5;
xmed=(2+3+5+7+9)/5;

% calcolo della media di Y
ymed=(y'*uno)/5;
ymed=(12+21+25+39+41)/5;

% varianza di X
varx=(1/5)*(x-xmed)'*(x-xmed);
varx=(1/5)*( (2-xmed)^2+(3-xmed)^2+(5-xmed)^2+(7-xmed)^2+(9-xmed)^2);

% varianza di Y
vary=(1/5)*(y-ymed)'*(y-ymed);

% covarianza XY
covarxy=(1/5)*(x-xmed*uno)'*(y-ymed*uno);
covarxy=(1/5)*( (2-xmed)*(12-ymed)+(3-xmed)*(21-ymed)+(5-xmed)*(25-ymed)+(7-xmed)*(39-ymed)+(9-xmed)*(41-ymed));
```

```
% calcolo del coefficiente di regressione B
b = covarxy/varx;

% calcolo del coefficiente di regressione A
a = ymed - b*xmed;

% calcolo dei punteggi predetti
ypred=a+b*x;

% vettore (5x1) degli errori
e=y-ypred;

% la somma degli errori deve essere uguale a zero
somma_err=e'*uno;


% INFERENZA RELATIVA AL COEFFICIENTE B

% stima della varianza degli errori
var_err=e'*e/(5-2);
var_err=1/(5-2)*( (-2.2927)^2+(2.5488)^2+(-1.7683)^2+(3.9146)^2+(-2.4024)^2 );

% varianza di B
var_b=1/((x-xmed)'*(x-xmed))*var_err;
var_b=1/( (2-xmed)^2+(3-xmed)^2+(5-xmed)^2+(7-xmed)^2+(9-xmed)^2)*var_err;

% calcolo di t
t=b/sqrt(var_b);
```

```
% SOMME DEI QUADRATI E COEFFICIENTE DI DETERMINAZIONE

% calcolo della somma dei quadrati totale
sq_tot=(y-ymed)'*(y-ymed);

% calcolo della somma dei quadrati della regressione
sq_pred=(ypred - ymed)'*(ypred - ymed);

% calcolo della somma dei quadrati dei residui
sq_res=e'*e;

% calcolo del coefficiente di determinazione r^2
r_sq=sq_pred/sq_tot;

% CALCOLO DEL COEFFICIENTE DI CORRELAZIONE

% calcolo di zx
zx=(x-xmed)/sqrt(varx);

% calcolo di zy
zy=(y-ymed)/sqrt(vary);

% calcolo di rxy
rxy = (zx'*zy)/5;

% calcolo dei coefficienti di regressione con un'altra formula
x1=data(1:5,1:2);
coeff=inv(x1'*x1)*x1'*y;

% calcolo dei coefficienti di regressione per i dati standardizzati
zx1=[1 -1.2494
     1 -0.8590
     1 -0.0781
     1  0.7028
     1  1.4837];

newcoeff=inv(zx1'*zx1)*zx1'*zy;

% calcolo del coefficiente di determinazione a partire dal coefficiente di correlazione
r_sq=rxy^2;
```