ECONOMETRIA COD. 6089
PRIMO PROBLEM SET

SOLUZIONI

Esercizio 1

Dimostrate che quando c'è una costante tra i regressori la media campionaria della y è uguale a quella della $\widehat{y}$ predetta.

Dalla derivazione del modello OLS sappiamo che:
$\widehat{\beta}_1 = \overline{y} - \widehat{\beta}_2 \overline{x}$
per definizione, poi:
$\widehat{y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x = \overline{y} - \widehat{\beta}_2 \overline{x} + \widehat{\beta}_2 x$
calcoliamo la media della y predetta
$\overline{\widehat{y}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{y}_i = \frac{1}{N} \sum_{i=1}^{N} \left(\overline{y} - \widehat{\beta}_2 \overline{x} + \widehat{\beta}_2 x \right) = \overline{y} - \widehat{\beta}_2 \overline{x} + \widehat{\beta}_2 \overline{x} = \overline{y}$

Oppure possiamo dimostrarlo.
$y_i = \widehat{y}_i + \widehat{\varepsilon}_i$
Dividiamo tutto per N
$\frac{y_i}{N} = \frac{\widehat{y}_i}{N} + \frac{\widehat{\varepsilon}_i}{N}$
Consideriamo le sommatorie:
$\sum \frac{y_i}{N} = \sum \frac{\widehat{y}_i}{N} + \sum \frac{\widehat{\varepsilon}_i}{N}$
Se c'è la costante $\sum \frac{\widehat{\varepsilon}_i}{N} = 0$
Allora,
$\sum \frac{y_i}{N} = \sum \frac{\widehat{y}_i}{N}$
$\overline{\widehat{y}} = \overline{y}$

Esercizio 2

Considerate il modello
$\mathbf{y} = \alpha + \boldsymbol{\beta} X + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim i.i.d \left(0, \sigma^2 I\right)$

a) Definite il coefficiente di determinazione R^2

R^2: misura della bontà esplicativa del modello
$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

b) Discutete intuitivamente il fatto che l'eliminazione di una variabile da un modello lineare peggiora sempre il valore di R^2.

Lo stimatore OLS corrisponde ad una minimizzazione della somma dei quadrati degli errori senza vincoli (R^2 è dato dal totale della varianza di Y meno la varianza residua dopo la minimizzazione divisa per la varianza totale di Y)

$R^2 = \frac{ESS}{TSS} = \mathit{1} - \frac{RSS}{TSS}$

Eliminare una variabile equivale a vincolare a zero il valore del rispettivo parametro. Il risultato di tale minimizzazione è certamente una somma dei quadrati degli errori maggiore.

Quindi,

$RSS\,(\text{vincolato}) > RSS\,(\text{non vincolato}) \Longrightarrow R^2\,(\text{vincolato}) < R^2\,(\text{non vincolato})\,.$

Attenzione: deve comunque rimanere la costante.

c) Dimostrate che $R^2 = corr^2\left(\mathbf{y}, \widehat{y}\right)$.

$R^2 = \frac{ESS}{TSS} = \mathit{1} - \frac{RSS}{TSS}$

misura della bontà esplicativa del modello.

$corr^2\left(y, \widehat{y}\right) = \frac{cov^2(y,\widehat{y})}{var(y)var(\widehat{y})} = \frac{cov^2(\widehat{y}+\widehat{e},\widehat{y})}{var(y)var(\widehat{y})} = \frac{var^2(\widehat{y})}{var(y)var(\widehat{y})} = \frac{var(\widehat{y})}{var(y)} = \frac{\sum\left(\widehat{y}-\overline{\widehat{y}}\right)^2}{\sum(y-\overline{y})^2} = \frac{ESS}{TSS} = R^2$

Esercizio 3

Dato il vettore y e la matrice X

$$\mathbf{y} = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} ; X = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix}$$

Calcolate:

a) il prodotto $X'X$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 1 & 5 & 2 & 4 \\ 5 & 4 & 6 & 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{bmatrix}$$

b) il prodotto $X'\mathbf{y}$

$$X'y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 1 & 5 & 2 & 4 \\ 5 & 4 & 6 & 4 & 6 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix}$$

c) lo stimatore dei minimi quadrati $\widehat{\boldsymbol{\beta}}$ per il modello: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (suggerimento: non invertite la matrice $X'X$, ma svolgete i calcoli usando il metodo di eliminazione)

Considerando i passaggi ai punti a) e b) possiamo scrivere:

$$\begin{array}{rcl} (X'X)\,\boldsymbol{\beta} & = & X'\mathbf{y} \\ \begin{bmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} & = & \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix} \end{array}$$

Invece di invertire $(X'X)$, *possiamo utilizzare il metodo di eliminazione per risolvere queste equazioni. Il primo passo consiste nel sottrarre tre volte la prima riga dalla seconda e cinque volte la prima riga dalla terza, ottenendo in tal modo:*

$$\begin{bmatrix} 5 & 15 & 25 \\ 0 & 10 & 6 \\ 0 & 6 & 4 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 20 \\ 16 \\ 9 \end{bmatrix}$$

Al passo successivo, moltiplicando la seconda riga per 0.6 e sottraendo dalla riga 3 si ricava:

$$\begin{bmatrix} 5 & 15 & 25 \\ 0 & 10 & 6 \\ 0 & 0 & 0.4 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 20 \\ 16 \\ -0.6 \end{bmatrix}$$

La terza equazione diventa: $0.4\beta_3 = -0.6$,*dove* $\beta_3 = -1.5$

Sostituendo il valore di β_3 *nella seconda equazione* $10\beta_2 + 6\beta_3 = 16$, *dove* $\beta_2 = 2.5$ *:*

Infine la prima equazione: $5\beta_1 + 15\beta_2 + 25 + \beta_3 = 20$, *dove* $\beta_1 = 4$

d) Scrivete la retta di regressione $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}}$

$$\widehat{\mathbf{y}} = 4 + 2.5X_2 - 1.5X_3$$

Esercizio 4

Considerate il seguente modello lineare:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

dove X è la matrice che contiene gli altri k regressori, compresa la costante.

Considera anche il seguente modello lineare dove tutti i regressori sono divisi per una costante a $\left(u \text{ è } \frac{\epsilon}{\alpha}\right)$:

$$\mathbf{y} = \frac{1}{a}X\tilde{\boldsymbol{\beta}} + \mathbf{u} \tag{2}$$

a) Mostrate che la relazione tra gli stimatori OLS dei modelli (1) e (2) è la seguente:

$$\widehat{\widetilde{\boldsymbol{\beta}}}_{OLS} = a\widehat{\boldsymbol{\beta}}_{OLS}$$

Considerate il seguente modello lineare dove tutti i regressori sono divisi per una costante a :

$$\mathbf{y}=\frac{1}{a}X\widetilde{\boldsymbol{\beta}}+\mathbf{u}$$

Calcoliamo lo stimatore OLS dei parametri $\widetilde{\beta}$*:*
prima definiamo $\omega = \frac{1}{a}X$

$$\widehat{\widetilde{\beta}} = (\omega'\omega)^{-1}\omega'\mathbf{y}$$

$$\widehat{\widetilde{\beta}} = \left(\frac{X'}{a}\frac{X}{a}\right)^{-1}\frac{X'}{a}\mathbf{y} = a(X'X)^{-1}X'\mathbf{y} = a\widehat{\boldsymbol{\beta}}$$

lo stimatore del modello (2) quindi coincide con quello calcolato per il modello (1) premoltiplicato per una costante a.

b)Mostrate che l'R^2 nei due modelli è lo stesso. Potrebbe essere negativo?

Consideriamo i residui del secondo modello:

$$\widehat{u} = \mathbf{y} - \omega\widehat{\widetilde{\boldsymbol{\beta}}} = \mathbf{y} - \frac{1}{a}Xa\widehat{\boldsymbol{\beta}} = \mathbf{y} - X\widehat{\beta} = \widehat{\epsilon}$$

i residui del secondo modello coincidono con i residui del primo modello, per cui la loro varianza è identica:

$$\widehat{\sigma}_u^2 = \widehat{\sigma}_e^2$$

anche il coefficiente di determinazione R^2*, che misura la bontà esplicativa del modello, non cambia.*

Potrebbe essere negativo? No, perchè è presente una costante (v è un vettore colonna dove tutti gli elementi sono uguali a 1).

c) Calcolate la varianza di $\widehat{\widetilde{\boldsymbol{\beta}}}_{OLS}$ se k=3, $a = 2$ e $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)$
$= \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$.

Considerate la varianza di $\widehat{\widetilde{\boldsymbol{\beta}}}$:

$$var\left(\widehat{\widehat{\beta}}\right)=\sigma^2\left(\omega'\omega\right)^{-1}=\sigma^2\left(\frac{X'}{a}\frac{X}{a}\right)^{-1}=a^2\sigma^2\left(X'X\right)^{-1}=a^2 var\left(\widehat{\beta}\right)$$

$$var\left(\widehat{\widehat{\beta}}\right)=2^2\left(\begin{array}{ccc}10&0&0\\0&10&0\\0&0&10\end{array}\right)=\left(\begin{array}{ccc}40&0&0\\0&40&0\\0&0&40\end{array}\right)$$

Esercizio 5

Sotto tutte le ipotesi del modello di regressione lineare, costruire la regione di rifiuto ($\alpha=0.05$), per l'ipotesi $H_0:\boldsymbol{\beta}_1=4\boldsymbol{\beta}_2$ contro $H_1:\boldsymbol{\beta}_1>4\boldsymbol{\beta}_2$ nel modello $\mathbf{y}=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1X_1+\boldsymbol{\beta}_2X_2+\boldsymbol{\epsilon}$.

Sappiamo che $\widehat{\beta}_1-4\widehat{\beta}_2\sim N\left(\boldsymbol{\beta}_1-4\boldsymbol{\beta}_2;var\left(\widehat{\beta}_1-4\widehat{\beta}_2\right)\right)$

dove $var\left(\widehat{\beta}_1-4\widehat{\beta}_2\right)=var\left(\widehat{\beta}_1\right)+16var\left(\widehat{\beta}_2\right)-8cov\left(\widehat{\beta}_1,\widehat{\beta}_2\right)$

se $\boldsymbol{\beta}_1-4\boldsymbol{\beta}_2=0:$

$t=\frac{\widehat{\beta}_1-4\widehat{\beta}_2}{\sqrt{\widehat{\sigma}^2_\varepsilon\left((X'X)^{-1}_{22}+16(X'X)^{-1}_{33}-8(X'X)^{-1}_{23}\right)}}\sim t_{(n-3)}$

dove $\widehat{\sigma}^2_\varepsilon$ *è lo stimatore non distorto della varianza dell'errore, e* $(X'X)^{-1}_{ij}$ *è l'elemento* ij *della matrice* $(X'X)^{-1}$

Rifiuterò quindi H0 con una dimensione d'errore (massima) di primo tipo pari all'α*%, se il rapporto t sopra definito sarà maggiore di* $t(n-3,a)$*, indicando con questo simbolo il quantile di una distribuzione t con n-3 gradi di libertà che lascia alla sua destra una probabilità pari ad* α*.*

Esercizio 6

a) Regredendo la variabile Y su regressori X_1 e X_2 otteniamo i seguenti risultati:

	coefficiente	std.error
X_1	0.50	0.50
X_2	0.80	0.40

Sappiamo inoltre che il numero di osservazioni N è 42, la somma dei quadrati dei residui (RSS_U) è 80, mentre la somma dei quadrati dei residui nell'ipotesi che i coefficienti di X_1 e X_2 siano pari a zero (RSS_R) è 120.

Calcolate il test-F.

Ricorriamo alla formula

$F=\frac{(RSS_R-RSS_U)/j}{RSS/N-k}=\frac{(120-80)/2}{80/40}=$ *10*

b) Dato il seguente output di STATA, commentate i valori dei test-t e del test-F. Cosa potete concludere riguardando alla regressione proposta?

```
reg   y x1 x2

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =  148.36
       Model |  8.6155e+10     2  4.3077e+10           Prob > F      =  0.0000
    Residual |  1.3646e+10    47   290348771           R-squared     =  0.8633
-------------+------------------------------           Adj R-squared =  0.8574
       Total |  9.9801e+10    49  2.0368e+09           Root MSE      =   17040

------------------------------------------------------------------------------
    y        |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       X1    |   3614.933    2304.47     1.57   0.123    -1021.063   8250.929
       X2    |  -3378.924    7218.92    -0.47   0.642    -17901.52   11143.67
     _cons   |   .1122415    .1419259   17.23   0.000     .1270837   .0073778
------------------------------------------------------------------------------
```

Come possiamo vedere dai p-value, i test-t sui coefficienti non rifiutano $H_0 : \beta_i = 0$*. Al contrario, invece, il test-F rifiuta l'ipotesi nulla* $H_0 : \beta_1 = \beta_2 = 0$*.*

Da ciò possiamo sospettare un problema di collinearità nei regressori (si veda anche l'ampiezza degli standard errors).

Esercizio 7

Regredendo la variabile Y sui regressori X_1 e X_2 si ottiene:

Regressore	Coefficiente	Std. Error
X_1	1.5	0.50
X_2	2	0.25

a) Definite il test-F per testare l'ipotesi $H_0 = \beta_1 = c_1; \beta_2 = c_2; ...; \beta_k = c_k$ contro l'ipotesi alternativa che almeno una delle componenti di H_0 sia falsa.

$F = \frac{(R\widehat{\beta}-c)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}-c)}{\widehat{\varepsilon}'\widehat{\varepsilon}} \frac{N-k}{j} \sim F(k, N-k)$

In questo caso, possiamo scrivere in maniera semplificata:

$\lambda = \frac{(\widehat{\beta}-c)'(X'X)(\widehat{\beta}-c)}{k\widehat{\sigma}^2} \sim F(k, N-k)$

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

$\widehat{\sigma}^2 = \frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{N-K}$

b) Costruite il test-F per l'ipotesi nulla che il coefficiente di X_1 sia nullo e che il coefficiente di X_2 sia pari a uno, contro l'ipotesi alternativa che almeno una restrizione non sia vera e fornite il valore della realizzazione della sua statistica test (supponiamo $X_1'X_2 = X_2'X_1 = 0$).

Consideriamo che :

$$(X'X) = \left(\begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix} \right) = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}$$

Sapendo che

$$var(\widehat{\beta}) = \widehat{\sigma}^2 \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} = \widehat{\sigma}^2 \frac{1}{\left(X_1'X_1\right)\left(X_2'X_2\right)} \begin{pmatrix} X_2'X_2 & 0 \\ 0 & X_1'X_1 \end{pmatrix}$$

$$\widehat{\sigma}^2 \begin{pmatrix} \frac{1}{X_1'X_1} & \\ & \frac{1}{X_2'X_2} \end{pmatrix} = \begin{pmatrix} 0.5^2 & 0 \\ 0 & 0.25^2 \end{pmatrix}$$

Da cui
$X_1'X_1 = 0.5^{-2}\widehat{\sigma}^2$
$X_2'X_2 = 0.25^{-2}\widehat{\sigma}^2$

$$\begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix} = \widehat{\sigma}^2 \begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix}$$

$\lambda = \frac{\widehat{\sigma}^2 \begin{pmatrix} 3/2 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 16 \end{pmatrix} \begin{pmatrix} 3/2 \\ 1 \end{pmatrix}}{2\widehat{\sigma}^2} = \frac{25}{2}$

c) Dato il valore del test-F calcolato al punto b) e supponendo che la somma dei quadrati nel modello non ristretto RSS_U sia pari a 10 e un campione di ampiezza 52, calcolate RSS_R.

$\lambda = \frac{(RSS_R - RSS_U)/j}{RSS_U/(N-k)}$

$\frac{25}{2} = \frac{(RSS_R - 10)/2}{10/(52-2)} \Longrightarrow RSS_U = 15$

Esercizio 8

Considerate il seguente modello:

$$\mathbf{y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x + \boldsymbol{\beta}_2 z + \boldsymbol{\epsilon}$$

Ecco la stima in STATA

```
reg y x z

      Source |       SS       df       MS              Number of obs =     953
-------------+------------------------------           F(  2,   950) =  966.71
       Model |  128389.471     2  64194.7355           Prob > F      =  0.0000
    Residual |  63085.3254   950  66.4056057           R-squared     =  0.6705
-------------+------------------------------           Adj R-squared =  0.6698
       Total |  191474.796   952  201.128988           Root MSE      =   8.149

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    7.08117   .2573069    27.52   0.000     6.576214    7.586125
           z |   .0078033   .0002896    26.94   0.000     .0072349    .0083717
       _cons |   7.476546   .7369036    10.15   0.000     6.030399    8.922693
------------------------------------------------------------------------------
```

a) Commentate i coefficienti di x e z.. Che effetto hanno sulla variabile y? Sono significativi?

Dalla tabella notiamo che il coefficiente di x è pari a 7.08117, questo significa che se x aumenta di 1, y aumenterà di 7.08117, tenendo fissa la variabile z. Il coefficiente è significativo, in quanto la statistica t calcolata è maggiore dell'1.96, se consideriamo come livello di significatività il 5%. Possiamo notare anche che il pvalue è pari a zero, quindi inferiore al 5%. Inoltre, notate che l'intervallo di confidenza riportato non include il valore zero.

Il coefficiente di z è pari a 0.0078033, questo significa che se z aumenta di 1, y aumenterà di 0.0078033, tenendo fissa la variabile x.Il coefficiente è significativo, in quanto la statistica t calcolata è maggiore dell'1.96, se consideriamo come livello di significatività il 5%. Possiamo notare anche che il pvalue è pari a zero, quindi inferiore al 5%. Inoltre, notate che l'intervallo di confidenza riportato non include il valore zero.

b) Costruite l'intervallo di confidenza al 95% di significatività per x. Ottenete lo stesso risultato della tabella?

(*Il valore critico di una normale nel caso di livello di confidenza del 95% è pari a 1.96*).

$H0 : \widehat{\beta} = 0 \ vs \ H1 : \widehat{\beta} \neq 0$

$t = \frac{\widehat{\beta}}{\sqrt{\widehat{\sigma}^2}}$

In questo caso, a qualsiasi livello di significatività, rifiutiamo l'ipotesi nulla, dato che il pvalue=0

L'intervallo di confidenza è il seguente:

$\widehat{\beta} - t_c \widehat{s.e.}(\widehat{\beta}) \leq \beta \leq \widehat{\beta} + t_c \widehat{s.e.}(\widehat{\beta})$

$Pr\left\{-t_c \leq \frac{\widehat{\beta}-\beta}{\widehat{s.e.}(\widehat{\beta})} \leq t_c\right\} = 1 - \alpha$

In questo caso, sostituendo i valori:

$7.08117 - 1.96 * 0.2573069 = 6.576$

7.08117 + 1.96 ∗ 0.2573069 = 7.586

c) Commentate la statistica F.

L'ipotesi nulla del test F riportato nella tabella è:

H0:$\beta_1 = \beta_2 = 0$

contro l'ipotesi alternativa H1: almeno uno tra β_1 *e* β_2 *sia diverso da zero.*

A qualsiasi livello di significatività, possiamo rifiutare l'ipotesi nulla.

d) Definite R^2 ed R^2_{adj}. Calcolateli, usando i valori della tabella di STATA e confrontate il risultato ottenuto con quello riportato in tabella.

$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{63085.3254}{191474.796} = 0.6705$

$R^2_{adj} = 1 - \frac{RSS/(N-k)}{TSS/(N-1)} = 1 - \frac{63085.3254/950}{191474.796/952} = 0.6698$

Esercizio 9

Consideriamo il seguente modello:

$$\mathrm{col}\,gpa = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 hesperc + \boldsymbol{\beta}_2 sat + \boldsymbol{\epsilon}$$

dove col *gpa* misura su una scala da 0 a 4 i voti degli studenti di un college americano, *hesperc* esprime in termini percentili la promozione alla scuola superiore (*hesperc*=5 significa che lo studente era nel 5% dei bravi della sua classe) e *sat* rappresenta il punteggio di un test attitudinale per l'ingresso in università, in cui vengono testate le abilità matematiche e verbali.

a) Spiegate perché ci aspettiamo una stima del coefficiente $\boldsymbol{\beta}_1 < 0$ e del coefficiente $\boldsymbol{\beta}_2 > 0$

La variabile hesperc rappresenta il percentile in cui lo studente si è piazzato al raggiungimento del diploma superiore, minore è il percentile, più bravo è lo studente. Più piccolo è hsperc , più bassa è la posizione in classifica nella propria classe di scuola superiore. A parità di condizioni più bassa è la posizione a scuola, più bassa è la media attesa dei voti all'università. Quindi ci aspettiamo un coefficiente negativo.

Invece, nel caso di sat ci aspettiamo che maggiore sia il punteggio raggiunto nel test, maggiore sarà la performance universitaria dello studente. Quindi ci aspettiamo un coefficiente positivo.

b) Ecco il modello stimato in STATA:

```
reg  colgpa  hsperc  sat

      Source |       SS       df       MS              Number of obs =    4137
-------------+------------------------------           F(  2,  4134) =  777.92
       Model |  490.606706     2  245.303353           Prob > F      =  0.0000
    Residual |  1303.58897  4134  .315333567           R-squared     =  0.2734
-------------+------------------------------           Adj R-squared =  0.2731
       Total |  1794.19567  4136  .433799728           Root MSE      =  .56155

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hsperc |  -.0135192   .0005495   -24.60   0.000    -.0145965    -.012442
         sat |   .0014762   .0000653    22.60   0.000     .0013482    .0016043
       _cons |   1.391757   .0715424    19.45   0.000     1.251495    1.532018
------------------------------------------------------------------------------
```

Ed ecco una descrizione delle tre variabili:

```
sum   colgpa hsperc sat

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      colgpa |      4137    2.652686    .6586347          0          4
      hsperc |      4137    19.23707    16.56873   .1666667         92
         sat |      4137    1030.331    139.4014        470       1540
```

Spiegate la significatività economica e statistica dei coefficienti di *hesperc* e *sat*.

Il coefficiente stimato della variabile sat è molto piccolo 0.0014762, ma guardando la t-statistic ed il pvalue associato, possiamo dire che la variabile sat è significativa nella regressione. Anche il segno assunto è quello atteso, in quanto maggiore è il punteggio al test sat, migliore sarà la performance all'università, anche se possiamo dire che l'impatto non è così rilevante.

c) Calcolate il valore di $\widehat{\mathrm{col}\,gpa}$, considerando che *hsperc*=20 e *sat*=1 050.

$$\widehat{\mathrm{col}\,gpa} = 1.391757 - 0.0135192 * 20 + 0.0014762 * 1050 = 2.6714$$

d) Supponiamo che due studentesse, Alice e Melissa, si siano diplomate piazzandosi con lo stesso percentile, ma Alice ha ottenuto un punteggio di 140 punti maggiore di Melissa al test SAT. Che differenze ci sono nel valore stimato di $\widehat{\mathrm{col}\,gpa}$?

La differenza tra Alice e Melissa è pari a 140 volte il coefficiente di sat, perché hsperc è uguale tra le due studentesse. Così il valore stimato di $\widehat{\mathrm{col}\,gpa}$ *per Alice è maggiore del* $0.0014762 * 140 = 0.20667$

e) Tenendo fissa la variabile *hsperc*, che differenza c'è nel punteggio del test SAT al fine di ottenere una variazione su *colgpa* pari a 0.5? Commentate.

Tenendo fissa la variabile hsperc, $\frac{\partial\,\mathrm{col}\,gpa}{\partial sat} = 0.0014762.$

Quindi se vogliamo trovare quanto deve aumentare il coefficiente di sat al fine di avere una variazione pari al 0.5 di colgpa,

$$\begin{aligned} 0.5 &= \Delta sat\,(0.0014762) \\ \Delta sat &= \frac{0.5}{0.0014762} = 338.71 \end{aligned}$$