Matematica II: Calcolo delle Probabilità e Statistica Matematica

ELT A-Z

Docente: dott. F. Zucca

Esercitazione # 8

Gli esercizi contrassegnati con (*) sono tratti da Eserc. # 10 - Statistica Matematica A 2002-2003- Prof. Secchi

1 Regressione lineare semplice

Esercizio 1 Per quattro Ingegneri si sono osservati gli anni trascorsi dalla Laurea X e il reddito annuale lordo Y misurato in euro. I risultati sono riassunti dalla seguente tabella:

Ingegnere	1	2	3	4
X	1	4	5	10
Y	15751,94	19625,36	21949,42	41316,55

1. Si determini la retta di regressione di Y su X
2. Se $Z = X^2$, la retta di regressione di Y su Z spiega una maggiore proporzione di variabilità di Y rispetto alla retta determinata al punto precedente?
3. Si stimi il reddito annuale lordo di un Ingegnere laureato da 3 anni

Esercizio 2 I valori assunti da due grandezze X ed Y in 20 diversi casi forniscono per entrambe le grandezze media campionaria nulla e varianze campionarie $s_X^2 = 9$ e $s_Y^2 = 4$ rispettivamente.

1. Si determini la retta di regressione di Y su X sapendo che essa passa per il punto $(3, 1)$
2. Si calcoli il coefficiente r_{xy}^2
3. Si determini la retta di regressione di X su Y

Esercizio 3 (*)In una tabella sono raccolti i seguenti dati

X	2	3	4.5	7
Y	10	10	15	20

1. Calcolare la retta di regressione di Y su X
2. Calcolare il coefficiente di correlazione di X ed Y
3. Calcolare il coefficiente di correlazione di X e $Z = \ln Y$. Questo modello è migliore del precedente?

Esercizio 4 (*)

Si vuole studiare la relazione tra le variabili Peso delle madri (X) e Peso dei figli (Y), entrambe misurate in Kg. Le osservazioni (x_i, y_i) per $i = 1, ..., 12$ sono tali che $\sum_{i=1}^{12} x_i = 800$, $\sum_{i=1}^{12} y_i = 811$, $\sum_{i=1}^{12} x_i^2 = 53418$, $\sum_{i=1}^{12} y_i^2 = 54849$, $\sum_{i=1}^{12} x_i y_i = 54107$.

1. Determinare la retta di regressione di Y su X.
2. Proporre uno stimatore non distorto per la varianza σ^2 delle componenti di errore ε_i del modello di regressione. Supposte valide le ipotesi gaussiane:
3. verificare l'ipotesi nulla $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$ al livello del 2%.
4. verificare se la retta di regressione passa per l'origine al livello del 2%.

Esercizio 5 (*)

Nel corso di uno studio naturalistico sono stati raccolti dati su 8 esemplari di una certa specie di albero: sono state misurate le altezze x_i (in metri) e i pesi y_i (in Kg.). Nell'intento di evidenziare una relazione tra le grandezze si ipotizza un legame lineare:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

sotto le ipotesi gaussiane (cioè $\varepsilon \sim N\left(0; \sigma^2\right)$). Dopo l'esecuzione della regressione, si sono trovati i seguenti valori per le stime dei parametri e la somma dei residui al quadrato:

$$\hat{\beta}_0 = 10.2;\ \hat{\beta}_1 = 25.3;\ SS_E = 172.5$$

con stime degli errori standard di $\hat{\beta}_0$ e $\hat{\beta}_1$ date da

$$se\left(\hat{\beta}_0\right) = 5.7859;\ se\left(\hat{\beta}_1\right) = 0.8088$$

1. Calcolare un intervallo di confidenza al livello 80% per β_0 e per β_1
2. verificare se la retta di regressione passa per l'origine al livello 20% ed al livello 10%
3. calcolare gli intervalli di confidenza al 90% per la previsione media e per la previsione del peso di un albero alto 10 metri.

Esercizio 6 (*)

In una certa comunità si registrano mensilmente il consumo di gelati X misurato in Kg. e il numero di casi di allergia al polline Y. I dati raccolti nell'ultimo anno forniscono le seguenti informazioni:

$$\begin{aligned} \bar{X} &= 120; \quad S_X = 100 \\ \bar{Y} &= 22; \quad S_Y = 5 \end{aligned}$$

Il coefficiente di correlazione vale 0.93

1. Si determini la retta di regressione di Y su X e se ne disegni il grafico
2. Si stimi (puntualmente) il numero di casi di allergia in un mese in cui il consumo di gelati è pari a 300 Kg. La previsione ottenuta è buona?

2 Regressione lineare multipla

Esercizio 7 (*)

Da un'elaborazione preliminare sui seguenti dati:

x_i	y_i
2.1	0.8518
3.2	2.3551
-1.2	-8.7368
-3.4	-11.2042
2.3	0.8329
2.4	-1.1961
1.7	2.3834
-0.9	-9.3468
-0.8	-6.1546
1.9	-1.9388

risultano le seguenti deviazioni standard campionarie $\sigma_X = 2.1427$, $\sigma_Y = 5.1807$ ed il coefficiente di correlazione $\rho_{xy} = 0.9421$. Si esegue una regressione lineare supponendo

$$Y = \beta_0 + \beta_1 \cdot x + \epsilon$$

e che valgano le ipotesi gaussiane.

1. Stimare i coefficienti β_0 e β_1.
2. Calcolare un intervallo di confidenza per β_0 al 95%.
3. Si valuti l'opportunità di aggiungere il nuovo regressore x^2

   $$Y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \epsilon$$

   con un test al 5% sapendo che per il modello in questione $SS_E = 25.639$.
4. Si considera il seguente modello

   $$Y = \beta_0 + \beta_1 \cdot x^3 + \epsilon;$$

   quanto dovrebbe valere la somma dei quadrati residua affinché sia preferibile questo modello a quello di regressione semplice?

Esercizio 8 (*)

Ad un gruppo di 4 uomini vengono misurati peso X, altezza Y e circonferenza toracica Z; il risultato è riassunto nella seguente tabella:

X (in Kg)	Y (in cm)	Z (in cm)
93	185	99
78	183	103
76	178	95
77	174	92

Eseguendo una regressione di X rispetto ad Y e Z si ottiene

$$SS_E = \sum_{i=1}^{n} (\widehat{x}_i - x_i)^2 = 22.56232409,$$
$$SS_R = \sum_{i=1}^{n} (\widehat{x}_i - \overline{x})^2 = 171.4376759,$$
$$SS_T = \sum_{i=1}^{n} (x_i - \overline{x})^2 = 194.$$

1. Eseguire un test per la significatività della regressione al 10%.
2. Determinare il P-value del test.

Esercizio 9

Una compagnia di sondaggi ha raccolto i seguenti dati circa il prezzo dei voli aerei (Y), la distanza chilometrica tra l'aeroporto di partenza e di arrivo (X1) e il numero di posti disponibili all'atto dell'acquisto (X2):

Y	X_1	X_2
160	600	10
240	1000	30
180	860	45
800	3000	25

1. Stimare i coefficienti della regressione multipla
2. Calcolare il coefficiente di detrminazione multipla corretto $R^2_{adjusted}$
3. Stimare la varianza dell'errore

3 Svolgimenti

Soluzione Es.1

1. Ipotizziamo che valga un modello del tipo $Y = \beta_0 + \beta_1 x$. Per determinare la retta di regressione occorre calcolare le stime $\hat{\beta}_0$ e $\hat{\beta}_1$ di β_0 e β_1 rispettivamente. È noto che $\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$ e $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$; utilizzando queste formule e calcolando

   $$s_{xy} = 1.2395 \cdot 10^5, \quad s_{xx} = 42 \quad s_{yy} = 3.8949 \cdot 10^8$$

   si ottiene $\hat{\beta}_0 = 9904,87$ e $\hat{\beta}_1 = 2951,19$. L'espressione della retta di regressione (stimata) è dunque $\hat{y} = 9904,87 + 2951,19x$.

2. Per confrontare due modelli di regressione utilizziamo il coefficiente di correlazione; è migliore il modello che presenta il coefficiente di correlazione più alto. Il coeff. per la regressione di Y su Z è dato dala relazione

   $$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} = 0,9691$$

   mentre, calcolando

   $$s_{yz} = 1.5083 \cdot 10^6, \quad s_{zz} = 5841,$$

   quello di Z su X è dato da

   $$r_{zy} = \frac{s_{zy}}{\sqrt{s_{zz} s_{yy}}} = 1.$$

   Il fatto che $r_{zy} = 1$ indica non solo che il secondo modello è preferibile al primo (ha un coeff. di correlaz. maggiore), ma anche che la relazione $Y = \hat{\gamma}_0 + \hat{\gamma}_1 z = 15493,69 + 258,23z$ è esatta.

3. Utilizzando dunque il secondo modello prevediamo che un Ingegnere laureato da 3 anni guadagni $Y = 15493,69 + 258,23 * 3^2 = 17817,76$ euro.

Soluzione Es. 2

1. Detti $\hat{\beta}_0$ e $\hat{\beta}_1$ i coefficienti della retta di regressione, si deve avere che $1 = \hat{\beta}_0 + \hat{\beta}_1 3$.. Questa equazione non è sufficiente da sola a determinare $\hat{\beta}_0$ e $\hat{\beta}_1$. Per il modo in cui viene costruita la retta dei minimi quadrati deve anche accadere che $\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$. Dal testo sappiamo che $\overline{x} = 0$ e $\overline{y} = 0$.. Risolvendo il sistema otteniamo $\hat{\beta}_0 = 0$ e $\hat{\beta}_1 = \frac{1}{3}$. Dunque la retta di regressione (stimata) ha equazione $y = \frac{1}{3}x$.

2. Notiamo che r^2_{xy} si può scrivere come $r^2_{xy} = \frac{cov(X,Y)^2}{var(X)var(Y)}$. Non conoscendo i dati, non possiamo calcolare direttamente $cov(X, Y)$. Possiamo però ricavarla osservando che $\hat{\beta}_1 = \frac{cov(X,Y)}{var(X)}$ (la varianza campionaria di X è nota dal testo). Si ha dunque $cov(X, Y) = \hat{\beta}_1 var(X) = \frac{1}{3} * 9 = 3$. Da ciò ricaviamo infine $r^2_{xy} = \frac{9}{9*4} = \frac{1}{4}$

3. Chiamiamo $\hat{\alpha}_0$ ed $\hat{\alpha}_1$ i coefficienti della retta di regressione di X su Y. Utilizzando quanto visto nei punti precedenti possiamo determinare $\hat{\alpha}_1$ come

$$\hat{\alpha}_1 = \frac{cov(Y, X)}{var(Y)} = \frac{cov(X, Y)}{var(Y)} = \frac{3}{4}.$$

Inoltre, dalla relazione $\overline{y} = \hat{\alpha}_0 + \hat{\alpha}_1 \overline{x}$ ricaviamo $\hat{\alpha}_0 = \overline{x} - \hat{\alpha}_1 \overline{y} = 0$. La retta di regressione di X su Y ha quindi equazione $\hat{x} = \frac{3}{4}y$

Soluzione Es.3

a) Calcoliamo innanzitutto le medie campionarie di X e Y.

$$\begin{aligned} \bar{x} &= \frac{2+3+4.5+7}{4} = 4.125 \\ \bar{y} &= ... = 13.75 \end{aligned}$$

Calcoliamo ora i coefficienti della retta di regressione di Y su X:

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = ... = 2.1586 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 4.8458 \end{aligned}$$

L'espressione della retta di regressione (stimata) è dunque $\hat{y} = 4.8458 + 2.1586x$.

b) Siccome $s_{xx} = 14.1875$, $s_{yy} = 68.75$ e $s_{xy} = 30.625$, il coefficiente di correlazione di Y e X è dato da

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} = 0.98059$$

c) Per confrontare due modelli di regressione utilizziamo il coefficiente di correlazione; è migliore il modello che presenta il coefficiente di correlazione più alto.

Facendo i conti, troviamo che $\bar{z} = 2.58$ e il coefficiente di correlazione di Z e X è dato da

$$r_{zx} = \frac{s_{zx}}{\sqrt{s_{zz} s_{xx}}} = 0.972.$$

Siccome $r_{xz} < r_{xy}$ deduciamo che è migliore il modello in a) di quello in c).

Soluzione Es.4

1. Per calcolare la retta di regressione incominciamo a ricavare s_{xx} e s_{xy} dai nostri dati.

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_i x_i = 66.\bar{6} \\ \bar{y} &= \frac{1}{n}\sum_i y_i = 67.6 \\ s_{xx} &= \sum_i x_i^2 - n\bar{x}^2 = 85.73 \\ s_{xy} &= \sum_i x_i y_i - n\bar{x}\bar{y} = 40.\bar{3} \end{aligned}$$

quindi

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = 0.476 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = 35 \end{aligned}$$

2. Uno stimatore non distorto della varianza degli errori della regressione è dato da

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2} = ... \simeq 2$$

3. Vogliamo testare l'ipotesi

$$H_0 : \beta_1 = 0 = \beta_{1,0}$$

contro

$$H_1 : \beta_1 \neq 0$$

Utilizziamo quindi la statistica test

$$\frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t_{n-2}$$

accettando l'ipotesi nulla al livello di significatività α se $-t_{\alpha/2;n-2} \leq \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \leq t_{\alpha/2;n-2}$.
Siccome $\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/s_{xx}}} = 3.116 > t_{\alpha/2;n-2} = 2.764$, allora rifiutiamo H_0 al livello del 2%.

4. Vogliamo testare l'ipotesi

$$H_0 : \beta_0 = 0 = \beta_{0,0}$$

contro

$$H_1 : \beta_0 \neq 0$$

Utilizziamo quindi la statistica test

$$\frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} \sim t_{n-2}$$

accettando l'ipotesi nulla al livello di significatività α se $-t_{\alpha/2;n-2} \leq \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} \leq t_{\alpha/2;n-2}$.
Siccome $\frac{\bar{x}^2}{S_{xx}} = 51.84$, $\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] = 103.852$, $\frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right]}} = 3.434 > t_{\alpha/2;n-2} = 2.764$, allora rifiutiamo H_0 al livello del 2%.

Soluzione Es.5

1. Gli intervalli di confidenza per i parametri β_0 e β_1 della regressione sono della forma:

$$\begin{aligned}\hat{\beta}_0 - t_{\alpha/2;n-2} se\left(\hat{\beta}_0\right) &\leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2;n-2} se\left(\hat{\beta}_0\right)\\ \hat{\beta}_1 - t_{\alpha/2;n-2} se\left(\hat{\beta}_1\right) &\leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2;n-2} se\left(\hat{\beta}_1\right)\end{aligned}$$

   Avendo quindi $\alpha/2 = 0.1$; $n = 8$; $t_{\alpha/2;6} = 1.440$, otteniamo quindi:

$$\begin{aligned}1.87 &\leq \beta_0 \leq 18.53\\ 23.65 &\leq \beta_1 \leq 26.95\end{aligned}$$

2. Vogliamo testare l'ipotesi

$$H_0 : \beta_0 = 0 = \beta_{0,0}$$

   contro

$$H_1 : \beta_0 \neq 0$$

   Siccome $\beta_{0,0} = 0$ non appartiene all'intervallo di confidenza trovato sopra, rifiutiamo H_0 al livello di significatività 20%. Mentre l'intervallo di confidenza al livello 90% è, sapendo che in questo caso $t_{\alpha/2,n-2}1.9432$,

$$[-1.043, 21.413];$$

   in questo caso $\beta_0 = 0$ appartiene all'intervallo, pertanto l'ipotesi nulla è accettata.

3. Dai dati ricaviamo immediatamente $\widehat{\sigma}^2 = SS_E/(n-2) \approx 21.5625$. Ricordando che

$$\begin{aligned}se(\widehat{\beta}_0) &= \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}\\ se(\widehat{\beta}_1) &= \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\end{aligned}$$

   da cui $s_{xx} \approx 39.9622$ e

$$\overline{x}^2 = \left(\frac{se(\widehat{\beta}_0)^2}{\widehat{\sigma}^2} - \frac{1}{n}\right) S_{xx} \approx 47.0549$$

   ma, essendo per la natura del problema $\overline{x} \geq 0$, si ha infine $\overline{x} \approx= 6.8597$. Infine

$$\begin{aligned}\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right) &\approx 9.1464\\ \widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right) &\approx 30.7089\end{aligned}$$

   da cui, calcolando come al punto precedente $t_{\alpha/2,n-2} = 1.9432$ (con $\alpha = 0.1$), i due intervalli sono, rispettivamente,

$$[260.1757, 266.2243] \qquad [257.6584, 268.7416].$$

Soluzione Es.6

1. Per calcolare la retta di regressione incominciamo a ricavare s_{xx} e s_{xy} dai nostri dati. Siccome $S_X^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ e analogamente per S_Y^2, otteniamo:

$$\begin{aligned} s_{xx} &= (n-1)\, s_X^2 = 1100 \\ s_{yy} &= ... = 55 \end{aligned}$$

Siccome

$$\begin{aligned} \hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}} = r_{xy}\frac{s_{yy}}{s_{xx}} = 0.0465 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 16.42 \end{aligned}$$

2. Dalla retta di regressione otteniamo

$$\hat{y}\,(300) = 16.42 + 0.0465 \cdot 300 = 30.37,$$

quindi il numero stimato di casi di allergia in un mese in cui il consumo di gelati è pari a 300 Kg. è 30.37

Soluzione Es.7

Anche in questo esercizio i dati sono ridondanti in quanto tutti i coefficienti che si servono possono essere ricavati dai dati in tabella.

1. Poniamo

$$\begin{aligned} s_{xx} &= \sum_{i=1}^{n} (x_i - \overline{x})^2, \\ s_{yy} &= \sum_{i=1}^{n} (y_i - \overline{y})^2, \\ s_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}), \end{aligned}$$

dove $\overline{x} = 0.73$ e $\overline{y} = -3.2154$; allora

$$\begin{aligned} s_{xx} &= \sigma_x^2 (n-1) = 9 \cdot 2.1427^2 = 41.3205, \\ s_{yy} &= \sigma_y^2 (n-1) = 9 \cdot 5.1807^2 = 241.5569, \\ s_{xy} &= \rho_{xy}\sqrt{s_{xx} s_{yy}} = 0.9421 \cdot \sqrt{41.3205 \cdot 241.5569} = 94.1216. \end{aligned}$$

Da cui facilmente

$$\begin{aligned} \widehat{\beta}_1 &= \frac{s_{xy}}{s_{xx}} = 2.2778, \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} = -3.2154 - 2.2778 \cdot 0.73 = -4.8782. \end{aligned}$$

2. Considerando la matrice

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{k,1} \\ 1 & x_{1,2} & \cdots & x_{k,2} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1,n} & \cdots & x_{k,n} \end{pmatrix}$$

si calcola facilmente

$$(X^t \cdot X)^{-1}_{1,1} = \frac{1}{n} + \frac{\overline{x}^2}{s_{xx}} = \frac{1}{10} + \frac{0.73^2}{41.3205} = 0.1129.$$

Ricordiamo che l'errore quadratico medio (o media quadratica dei residui) è

$$\widehat{\sigma}^2 = MS_E := \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

(in questo caso $n = 10$ e $p = k + 1 = 2$). Nel caso di regressione semplice

$$\widehat{\sigma}^2 = \left(SS_T - \widehat{\beta}_1 s_{xy}\right)\frac{1}{n-2} = \left(s_{yy} - \frac{s_{xy}^2}{s_{xx}}\right)\frac{1}{n-2} = 3.3953.$$

Allora l'intervallo di confidenza a livello α è $[\widehat{\beta}_o^-, \widehat{\beta}_o^+]$ dove

$$\widehat{\beta}_o^{\pm} := \widehat{\beta}_0 \pm se(\widehat{\beta}_o) \cdot t_{\frac{\alpha}{2},n-p}$$

e

$$se(\widehat{\beta}_o) := \sqrt{\widehat{\sigma}^2 \cdot (X^t \cdot X)_{1,1}} = \sqrt{3.3953 \cdot 0.1129} = 1.4277,$$

quindi, essendo $t_{0.025,8} = 2.306$, si ha che l'intervallo cercato è $[-8.1705, -1.5859]$.

3. Nel modello con regressore aggiunto, si tratta di testare l'ipotesi nulla

$$\mathrm{H}_0 \ : \ \beta_2 = 0.$$

Si calcola

$$\begin{pmatrix} 10 & 7.3 & 46.65 \\ 7.3 & 46.65 & 37.519 \\ 46.65 & 37.519 & 343.6245 \end{pmatrix}$$

da cui

$$\begin{aligned} \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} &= (X^t \cdot X)^{-1} \cdot X^t \cdot Y \\ &= \begin{pmatrix} 0.2815 & -0.0146 & -0.0366 \\ -0.0146 & 0.0243 & -0.0007 \\ -0.0366 & -0.0007 & 0.008 \end{pmatrix} \cdot \begin{pmatrix} -32.1541 \\ 70.6524 \\ -128.3333 \end{pmatrix} \\ &= \begin{pmatrix} -5.3827 \\ 2.2687 \\ 0.1096 \end{pmatrix}. \end{aligned}$$

Pertanto

$$\widehat{\sigma}^2 = \frac{SS_E}{n-p} = \frac{25.639}{10-3} = 3.6627,$$

e quindi

$$se(\widehat{\beta}_2) = \sqrt{\widehat{\sigma}^2 (X^t \cdot X)^{-1}_{3,3}} = 35.4767.$$

Lo stimatore che si utilizza è

$$T_0 := \frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)} = \frac{0.1096}{35.4767} = 0.0031$$

con regione di accettazione

$$|T_0| \leq t_{\alpha/2,n-p} \equiv t_{0.025,7} = 2.3646.$$

L'ipotesi nulla è pertanto accettata e non è opportuno aggiungere il regressore x^2.

4. Il secondo modello, con k_2 regressori, risulta migliore del primo, con k_1 regressori, se e solo se
$$\frac{SS_{E_2}}{n-k_2-1} < \frac{SS_{E_1}}{n-k_1-1}.$$
Nel caso in questione si ha $n = 10$, $k_1 = k_2 = 1$ ed $SS_{E_1} = 27.1627$ per cui li modello con regressore x^3 sarà migliore di quello con x se e solo se
$$SS_{E_2} < 27.1627.$$

Soluzione Es.8

Osserviamo che i dati del problema sono ridondanti e che gli "errori quadratici" siricavano banalmente dai dati in tabella. D'altro canto conosciuti gli "errori quadratici", non siamo più interessati ai dati della tabella se non per conoscere l'ampiezza del campione $n = 4$ ed il numero di regressori $k = 2$.

1. Si consideri l'ipotesi nulla
$$\mathrm{H}_0 \ : \ \beta_0 = \beta_1 = \beta_2 = 0.$$
Lo stimatore è
$$F_0 := \frac{SS_R/k}{SS_E/(n-p)}$$
(dove $p = k + 1$) con la regione di accettazione a livello α
$$f_0 < f_{\alpha,k,n-p}.$$
Eseguendo i calcoli si ottiene
$$f_0 = \frac{171.4376759/2}{22.56232409/(4-3)} = \frac{85.7188}{22.56232409} = 3.7992$$
e $f_{0.1,2,1} = 49.5$. Pertanto accetto H_0 al livello del 5% pertanto la regressione non è significativa.

2. Se utilizziamo le tabelle dei quantili otteniamo
$$f_{0.25,2,1} = 7.5 > 3.7992$$
da cui $\overline{\alpha} > 0.25$. Utilizzando un calcolatore si ottiene
$$\overline{\alpha} = f^{-1}_{\cdot,2,1}(3.7992) = 1 - F_{f_{2,1}}(3.7992) = 0.341.$$

Soluzione Es.9

1. La stima dei coefficienti della regressione: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ si ottengono dal seguente sistema di equazioni:
$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_1^i + \hat{\beta}_2 \sum_i x_2^i = \sum_i y^i \\ \hat{\beta}_0 \sum_i x_1^i + \hat{\beta}_1 \sum_i \left(x_1^i\right)^2 + \hat{\beta}_2 \sum_i x_2^i x_1^i = \sum_i x_1^i y^i \\ \hat{\beta}_0 \sum_i x_2^i + \hat{\beta}_1 \sum_i x_1^i x_2^i + \hat{\beta}_2 \sum_i \left(x_2^i\right)^2 = \sum_i x_2^i y^i \end{cases}$$
Siccome
$$\begin{aligned} \sum_i x_1^i &= 5460; \ \sum_i x_2^i = 110; \ \sum_i y^i = 1380 \\ \sum_i x_1^i x_2^i &= 149700; \ \sum_i \left(x_1^i\right)^2 = 11099600; \ \sum_i \left(x_2^i\right)^2 = 3650 \\ \sum_i x_1^i y^i &= 2891800; \ \sum_i x_2^i y^i = 36900 \end{aligned}$$

allora otteniamo

$$\begin{aligned}\hat{\beta}_0 &= 3.5\\ \hat{\beta}_1 &= 0.28\\ \hat{\beta}_2 &= -1.48\end{aligned}$$

ottenendo quindi $\hat{y} = 3.5 + 0.28x_1 - 1.48x_2$

2.
$$R^2_{adjusted} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

dove

$$\begin{aligned}SS_E &= \text{somma quadrati errori} = \sum_i (y_i - \hat{y}_i)^2\\ SS_T &= \sum_i (y_i - \bar{y})^2\\ p &= \text{numero parametri regressione}\end{aligned}$$

Siccome $p = 3$ $(\beta_0, \beta_1, \beta_2)$, $\bar{y} = 345$, $\hat{y}_1 = 156.7$, $\hat{y}_2 = 239.1$, $\hat{y}_3 = 177.7$, $\hat{y}_4 = 806.5$ allora $SS_T = 279500$ e $SS_E = 59.24$ e di conseguenza:

$$R^2_{adjusted} = 0.99936$$

3. La stima della varianza dell'errore è data da:

$$\hat{\sigma}^2 = \frac{SS_E}{n-p} = \frac{59.24}{1} = 59.24$$