Statistica Matematica A

Esercitazione # 8

1.Regressione lineare e ANOVA

Esercizio # 1.1

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 # 1.2

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?

Soluzioni

Soluzione es 1.1:

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 1.2:

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

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

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

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 $[-6.3059, -3.4505]$.

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

$$SS_{E_2} < 27.1627.$$