HW4: soluzione degli esercizi assegnati
Modelli Statistici a.a. 2019/20

November 2, 2019

Esercizio 7, capitolo 4

Supponiamo di aver raccolto dati sul prezzo X (in euro) dell'hamburger e sulla quantità Y venduta in due giorni, in $n = 12$ punti vendita. I dati vengono sintetizzati con le seguenti misure:

$$\sum x_i y_i = 32709.2,\ \sum x_i = 35.9,\ \sum y_i = 11090,\ \sum x_i^2 = 108.13,\ \sum y_i^2 = 10584612$$

Stimare la retta di regressione e interpretarla.

Per ottenere la retta di regressione calcoliamo prima di tutto le seguenti quantità:

- prezzo medio dell'hamburger: $\overline{x} = \sum x_i/n = 2.99$
- numero medio di hamburger venduti: $\overline{y} = \sum y_i/n = 924.17$
- devianza di X: $dev(X) = S_{XX} = \sum x_i^2 - n\overline{x}^2 = 0.7292$
- devianza di Y: $dev(y) = S_{YY} = \sum y_i^2 - n\overline{y}^2 = 335603.67$
- codevianza: $codev(x, y) = S_{XY} = \sum x_i y_i - n\overline{y}\,\overline{x} = -468.38$

Le stime dei MQ di α e β sono: $\hat{\beta} = \frac{S_{XY}}{S_{XX}} = -642.35$ e $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 2845.88$. La retta di regressione stimata è quindi:

$$\hat{y} = 2845.88 - 642.35x$$

- $\hat{\beta}$ è la pendenza della retta di regressione che, come atteso, è negativa: se il prezzo sale il numero di hamburgher venduti si riduce. In particolare, se il prezzo dell'hamburgher cresce di un euro ci si aspetta che il numero di hamburgher venduti si riduca in media di 642 unità.
- $\hat{\alpha}$ è l'intercetta della retta di regressione e corrisponde al numero atteso di hamburgher venduti quando il prezzo è zero. In questo caso non c'è un'interpretazione del valore di α.

Per trovare la devianza residua osserviamo che $dev(\hat{\varepsilon}) = dev(y) - dev(\hat{y})$ e $dev(\hat{y}) = \hat{\beta}^2 dev(x)$, quindi $dev(\hat{\varepsilon}) = dev(y) - \hat{\beta}^2 dev(x) = 34735.63$.

La stima della varianza di errore è $S^2 = dev(\hat{\varepsilon})/(n-2) = 3473.56$, da cui $S = 58.94$. Quindi l'errore standard del coefficiente di regressione è $SE = \sqrt{S^2/dev(x)} = 69.02$.

L'intervallo di confidenza al $(1-c)\%$ per la pendenza è dato da $\hat{\beta} \pm t_{c/2,(n-2)} SE$.

Posto per esempio $c = 0.05$ si ha $-642.35 \pm 2.23 \times 69.02$, cioè $[-796.14, -488.57]$.

Esercizio 9 cap.4

In un articolo del 1857 il fisico scozzese James Forbes ha presentato i risultati di una serie di esperimenti che aveva realizzato in varie località delle Alpi e in Scozia per studiare la relazione tra pressione atmosferica e punto di ebollizione dell'acqua. Forbes era interessato a misurare la pressione con un termometro anzichè con uno strumento fragile e costoso (a quel tempo) come il barometro. I dati osservati sono riportati nel file forbes.csv. La pressione `Pressure` è misurata in pollici, mentre la temperatura `Temp` è misurata in gradi Fahrenheit. Trasformare le due variabili in mm e in gradi Celsius usando la relazione $C = 5/9(F-32)$. Adattare un modello di regressione lineare semplice per studiare la relazione tra le due variabili. Interpretare i risultati.

I dati per questo esercizio sono sono nel file `forbes.csv`. Il codice Stata si trova nel file `ese9_cap4.do`. I risultati sono riportati in Tabella 1.

Table 1: Modello di regressione lineare $pressione = \alpha + \beta tempC + \varepsilon$.

Source	SS	df	MS	Number of obs	=	17
				$F(1,15)$	=	2677.13
Model	93628.9735	1	93628.9735	Prob > F	=	0.0000
Residual	524.6053	15	34.9737	R-squared	=	0.9944
				Adj R-squared	=	0.9941
Total	94153.5788	16	5884.5987	Root MSE	=	5.9139

pressione	Coef.	Std. Err.	t	P¿t	[95% Conf.	Interval]
tempC	23.907	0.4621	51.74	0.000	22.922	24.892
_cons	-1634.015	43.9057	-37.22	0.000	-1727.598	-1540.432

La retta dei MQ è $\hat{y}_i = -1634.015 + 23.907 x_i$. Il valore di $\hat{alpha} = -1634.015$ è espresso in mm, e rappresenta il valore atteso stimato della pressione ad una temperatura di zero gradi Centigradi. Poichè l'acqua bolle a temperature di minimo 90 gradi, questo valore non è interpretabile. Il valore stimato del coefficiente associato alla temperatura $\hat{\beta} = 23.907$ è espresso in $mm/°C$ e rappresenta la variazione attesa nella pressione quando la temperatura aumenta di 1 grado. Poichè il test t per l'ipotesi $H_0 : \beta = 0$ è significativo (osservare valore della statistica test e del p-value in Tabella 1, concludiamo che c'è una relazione tra pressione e temperatura. Il test sulla significatività di β può essere fatto anche utilizzando la statistica F

che confronta la devianza di errore del modello nullo con la devianza del modello stimato: $F = \frac{[dev(e_0) - dev(e)/1}{dev(e)/(n-2)} = \frac{dev(\hat{y})}{dev(e)/(n-2)} \sim F_{(1,n-2)}$. Il valore della statistica test, riportato in Tabella 1, è $F = 2677.13$. Si noti anche che risulta $t^2 = F$.

Per valutare l'adattamento del modello osserviamo che il valore di $R^2 = 0.9944$ è piuttosto elevato, e la retta di regressione sembra adattarsi bene ai dati osservati (Figura 2).

Tuttavia, dal grafico dei residui $e_i = y_i - \hat{y}_i$ rispetto alla temperatura riportato in Fig. 2, osserviamo che l'andamento non è casuale, ma sembra esserci una relazione quadratica tra i residui e la temperatura; dal grafico si rileva inoltre la presenza di una osservazione con un residuo particolarmente elevato (outlier).

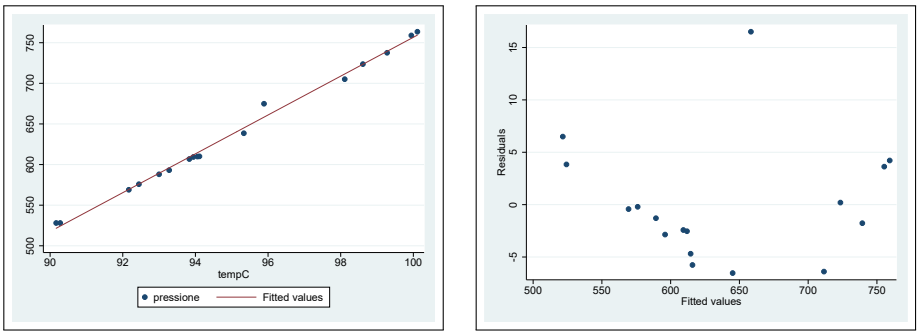

Table 2: Retta dei MQ e plot dei residui

Per migliorare l'adattamento del modello si può eliminare l'outlier e stimare un modello con anche un termine al quadrato per la temperatura.

Esercizio 3 paragrafo 7.8

(Freedman, 2009). Considerate il modello di regressione multipla $Y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i$ con ε_i indipendenti e normali $n(0, \sigma^2)$. Supponiamo che la teoria sottostante al problema ci suggerisca di sottoporre a test l'ipotesi $\beta = 0$. Si adatta il modello ai dati per $i = 1, \ldots, 53$ col metodo dei minimi quadrati ottenendo $\hat{\beta} = 3.79$ con errore standard $SE = 1.88$. Rispondere vero o falso alle domande seguenti e giustificare.

a. La statistica test per sottoporre a verifica $\beta = 0$ è $t = 2.02$.

   VERO. Infatti $t = 3.79/1.88 = 2.0159574 \simeq 2.02$.

b. $\hat{\beta}$ è significativo.

   VERO. Infatti il p-value osservato per l'alternativa bilaterale $H_1 : \beta \neq 0$ è $2P(t_{n-3} > 2.02) = 2 \times 0.024 = 0.048 < 0.05$.

c. $\hat{\beta}$ è altamente significativo.

   FALSO. Il test si dice altamente significativo se il p-value è minore di 0.01.

d. La probabilità che β sia diverso da 0 è circa il 95%.

FALSO. Il parametro β è una costante incognita, a cui non è associata una distribuzione di probabilità.

e. La probabilita che β sia uguale a zero è circa il 5%.

FALSO. Per lo stesso motivo del punto precedente.

f. Se il modello è giusto e $\beta = 0$ vi è circa il 5% di probabilità di ottenere $\mid \hat{\beta}/SE \mid > 2$.

VERO. Infatti, poichè sotto $H_0 : \beta = 0$ si ha $\hat{\beta}/SE \sim t_{n-3}$, possiamo calcolare $P(|\hat{\beta}/SE| > 2) = 2P(t_{n-3} > 2) = 2[1 - P(t_{n-3} \leq 2)] = 2 \times 0.025 = 0.05$.

g. Se il modello è giusto e $\beta = 0$ vi e circa il 95% di probabilità di ottenere $\mid \hat{\beta}/SE \mid < 2$.

VERO. Per quanto visto al punto precedente.

h. Il test dimostra che il modello teorico è giusto.

FALSO. Il test consente di decidere, in base all'evidenza empirica, se rifiutare l'ipotesi che $\beta = 0$, assumendo che il modello teorico sia giusto.

i. Il test assume che il modello teorico sia giusto.

VERO. Infatti, se si suppone adeguata la versione forte del modello di regressione, cioé se gli errori sono normali, si dimostra che gli stimatori dei minimi quadrati sono normali, da cui si ottiene che la distribuzione campionaria del rapporto standardizzato $\hat{\beta}/SE$ è t di Student.

l. Se il modello è giusto, il test fornisce evidenza che $\beta = 0$.

FALSO. In questo caso il test fornisce evidenza empirica contro l'ipotesi $\beta = 0$ al livello di significatività del 5%.

Esercizio 5, paragrafo 7.8

Date 40 osservazioni sulle variabili Y, X, Z si adatta un modello di regressione lineare multipla normale considerando Y come risposta e ottenendo

$$\begin{array}{rlll} Y = 936.4+ & 120.9x & +163.0z \\ & (20.46) & (72.81) \end{array}$$

dove i valori tra parentesi sono gli errori standard. La devianza residua è $dev(\mathbf{e}) = 31370.04$.

Si adatta poi il sotto-modello seguente, prendendo solo X come esplicativa ottenendo

$$\begin{array}{rll} Y = 27974.7 & +112.7x \\ & (115.3) \end{array}$$

con devianza residua $dev(e_0) = 79714.12$.

(a) Determinare un intervallo di confidenza al livello del 95% per il coefficiente di regressione $\beta_{YZ.X}$

L'intervallo di confidenza al 95% per $\beta_{YZ.X}$ è dato da $\hat{\beta}_{YZ.X} \pm t_{0.975,(40-3)}SE$. Poichè $t_{0.975,37} = 2.026$ si ha $163 \pm 2.026 \times 21.587$, cioè $[119.261, 206.739]$.

(utilizzano $SE = 72.81$ l'intervallo risulta $[15.473, 310.527]$.)

(b) *Stimare la varianza degli errori nel primo modello.*

La stima della varianza di errore è $S^2 = dev(\hat{\varepsilon})/(n-3) = 31370.04/(40-3) = 847.839$, da cui $S = 29.118$.

(c) *Sottoporre a verifica l'ipotesi nulla di uguaglianza a zero del coefficiente regressione $\beta_{YZ.X}$ al livello dell'1% usando il test t di Student e il test F. Verificate che $F = t^2$.*

La statistica test t per l'ipotesi $H_0 : \beta_{YZ.X} = 0$ è data da $\widehat{\beta}_{YZ.X}/SE = 7.55$, mentre il valore critico per il test a due code è $t*_{0.005,37} = 2.43$ e il $p - value = 5.29803E - 09$. Osservando che $p - value < 0.001$, o che $t < t*_{0.005,37}$, concludiamo che il test risulta significativo. (utilizzando $SE = 72.81$ la statistica test risulta 2.24, con $p - value = 0.03128$).

Il test sulla significatività di $\widehat{\beta}_{YZ.X}$ può essere fatto anche utilizzando la statistica F che confronta la devianza di errore del modello ridotto (senza Z) con la devianza del modello completo (con Z): $F = \frac{[dev(e_0)-dev(e)]/1}{dev(e)/(n-3)} = \frac{[dev(e_0)-dev(e)]/1}{S^2} \sim F_{(1,n-k)}$. Il valore della statistica test è dunque $F = \frac{48344.08}{847.839} = 57.02$. Possiamo confrontare questo valore con il quantile superiore della distribuzione teorica $F_{(1,37)} = 7.373$, che porta a rifiutare l'ipotesi nulla.

Si noti che risulta $t^2 = (7.55)^2 = 57.02 = F$.