

Soluzione HW 3 (Regressione lineare semplice)

Modelli Statistici a.a. 2019/20

Esercizio 1 Sezione 4.8

Nell'equazione stimata della retta di regressione per l'esempio della molla, il valore 0.049 che cos'è? Opzioni: un parametro, una variabile aleatoria, il valore realizzato di una variabile aleatoria.

Si veda esempio 4.4 (Legge di Hooke) a pag. 109 delle Dispense. La retta di regressione dei minimi quadrati risulta $y = 439.006 + 0.049x$.

Il valore 0.049 è il valore realizzato di una variabile aleatoria. Lo stimatore dei MQ $\hat{\beta}$, infatti, è una variabile aleatoria, mentre la stima 0.049 è una sua realizzazione.

Esercizio 5 della Sezione 4.8

Nella equazione $Y_i = \alpha + \beta x_i + \varepsilon_i$, β che caratteristiche ha? Opzioni (potete sceglierne anche più di una): osservabile, non osservabile, un parametro, una variabile aleatoria.

β è un parametro della popolazione, quindi un valore costante, non noto.

Quesiti 9, 10, 11, 12 della Sezione 4.9.

9. *Vero o falso: la retta di regressione ha pendenza nulla se le variabili sono incorrelate. Giustificare.*

Vero. Infatti la pendenza della retta dei MQ è data da $\hat{\beta} = \text{cov}(X, Y)/V(X)$.

10. *Vero o falso: se il coefficiente di correlazione è -1 la retta di regressione ha pendenza negativa, con inclinazione di 45 gradi. Giustificare.*

Falso. Infatti la retta inclinata di 45 gradi ha pendenza pari a -1 , mentre $\hat{\beta} = r_{xy} \frac{\sigma_y}{\sigma_x}$. Quindi se $r_{xy} = -1$ la pendenza è negativa (il segno di $\hat{\beta}$ dipende dal segno di r_{xy}), ma NON è detto che $\hat{\beta}$ sia pari a -1 : dipenderà dal valore del rapporto $\frac{\sigma_y}{\sigma_x}$ tra le deviazioni standard di X e Y .

11. *Vero o falso: I valori adattati hanno una somma uguale a $n\bar{y}$. Giustificare.*

Vero. Osserviamo che $n\bar{y} = \sum_{i=1}^n y_i$. Ricordando che una delle proprietà dei MQ, derivabile dalla prima delle equazioni normali, è che $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$, si ha che $n\bar{y} = \sum_{i=1}^n \hat{y}_i$.

12. *Vero o falso: se le due varianze campionarie di X e Y sono uguali a 1, il coefficiente di regressione e il coefficiente di correlazione sono uguali.*

Vero. Infatti, se $V(X) = 1$ anche $\sigma_x = 1$, ne consegue che $\text{corr}(Y, X) = \frac{\text{Cov}(Y, X)}{\sigma_y \sigma_x} = \text{Cov}(Y, X)$ e $\hat{\beta} = \frac{\text{cov}(Y, X)}{V(X)} = \text{cov}(Y, X) \Rightarrow \hat{\beta} = \text{corr}(Y, X)$.

Esercizio test di ingresso Economia

Prendere i dati dei test d'ingresso a Economia del 2013 (dati in formato .csv) e adattare il modello di regressione lineare semplice per studiare la dipendenza del punteggio al test dal voto alla maturità. Commentare i risultati ottenuti, in particolare il significato e l'unità di misura dell'intercetta e della pendenza.

Il codice Stata è memorizzato nel file `HW3_test_economia.do`. Per stimare il modello si utilizza il comando `reg`.

Il modello spiega solo una piccola parte della variabilità osservata nei punteggi al test: $R^2 = dev(\hat{y})/dev(y) = 0.0510$. Ricordiamo che $R^2 = r_{xy}^2$ e osserviamo che la correlazione tra le due variabili è positiva ma bassa, $r_{xy} = 0.2259$.

Le stime dei MQ sono: $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = .0802793$ e $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 7.126854$, quindi, arrotondando il valore delle stime alla seconda cifra decimale, la retta regressione stimata è:

$$\hat{y} = 7.13 + 0.08x$$

Per valutare se esiste una relazione lineare significativa tra il punteggio al test e il voto di maturità, possiamo sottoporre a verifica l'ipotesi $H_0 : \beta = 0$. Questo test può essere fatto sia utilizzando la statistica test $T = \frac{\hat{\beta}}{SE}$, dove $T \sim t_1$, che la statistica test $F = \frac{SS_{Model}}{SS_{residual}/(n-2)}$, $F \sim F(1, n-2)$. Ricordiamo inoltre che $F = T^2$. Si ottiene $t = \frac{0.08}{0.0103} = 7.78$ e $F = \frac{897.39}{14.82} = 60.54$, con un $p - value < 0.01$ per entrambi i test. Siamo quindi portati a rifiutare l'ipotesi nulla: c'è evidenza empirica di una relazione lineare tra punteggio al test e voto di maturità (β è significativamente diverso da zero).

L'unità di misura dell'intercetta α è la stessa della variabile dipendente. Il valore stimato per l'intercetta non è in questo caso molto importante, infatti esprime una stima del valore atteso del punteggio quando il voto di maturità è pari a zero: $\hat{E}(Y|X = 0) = \hat{\alpha} = 7.13$.

L'unità di misura del coefficiente di regressione β è invece il rapporto tra l'unità di misura del punteggio al test e quella del voto di maturità, possiamo quindi dire che per ogni punto in più sul voto di maturità, il valore atteso del test cresce di $\hat{\beta} = .08$. Per valutare quanto sia rilevante questa variazione, considerando che il voto di maturità varia tra 60 e 100, possiamo calcolare il valore atteso condizionato in corrispondenza del voto di maturità minimo:

$$\hat{E}(Y|X = 60) = 7.13 + 0.08 \times 60 = 11.94$$

Questo valore atteso si incrementa di 0.08 per ogni punto in più alla maturità. Per esempio, considerando un incremento di 20 punti nel voto, il punteggio atteso al test aumenta di $(80 - 60) \times 0.08 = 1.61$ e quindi si ha: $\hat{E}(Y|X = 80) = 11.94 + 1.61 = 13.55$.