Politecnico di Milano - Scuola di Ingegneria Industriale e dell'Informazione

I appello di Statistica per Ingegneria Fisica
11 Luglio 2019

Cognome, Nome e Numero di matricola:

Problema 1. Da tempo immemorabile, la villa del signor Burns è sempre stata illuminata a candele. Solo da poco, su consiglio del fido segretario Smithers, Burns si è finalmente deciso ad abbandonare le candele e a installare al loro posto un nuovo e più economico impianto a luci LED. È ovvio, però, che ora tocca a Smithers occuparsi dell'installazione...

Smithers ha dunque acquistato un lotto di 50 lampadine a LED da 4 Watt ciascuna. Dopo aver letto meglio le istruzioni sulla scatola, però, egli si è reso conto che questo valore è solo il consumo nominale delle lampadine. In realtà, per ottenere il loro consumo effettivo, ai 4 Watt bisogna ancora aggiungere un errore statistico X (anch'esso in Watt) che è una variabile aleatoria assolutamente continua con densità

$$f(x) = \begin{cases} \frac{c}{94}(16 - x^2) & \text{se } x \in [-1\,,\,1]\,, \\ 0 & \text{altrimenti}\,. \end{cases}$$

Sfortunatamente, nelle istruzioni il valore della costante c al numeratore risulta illeggibile.

(a) Aiutate voi il povero Smithers a trovare il valore di c per cui la funzione f è una densità di probabilità. Per tale valore, tracciate un grafico qualitativo di f.

D'ora in poi, usate il valore di c che avete trovato al punto precedente.

(b) Calcolate le probabilità $\mathbb{P}(X > 0)$ e $\mathbb{P}(|X| < 2)$.

(c) Calcolate la media e la varianza dell'errore X.

(d) Calcolate la media e la varianza della variabile aleatoria $P = 4 + X$, che è il consumo effettivo di una lampadina presa a caso tra quelle comprate da Smithers.

(e) Smithers ha intenzione di illuminare la villa di Burns installandoci tutte le 50 lampadine acquistate. Supponendo che i consumi effettivi delle lampadine siano indipendenti tra loro, calcolate la probabilità che l'intero impianto installato da Smithers consumi più di 195 Watt.

(f) L'impianto è alimentato a 220 Volt. Per l'effetto Joule, la resistenza elettrica di una lampadina a caso è la variabile aleatoria

$$R = \frac{220^2}{P} = \frac{48400}{4 + X} \qquad (\text{espressa in } \Omega)\,.$$

Calcolate il valore atteso di R. Se non riuscite a calcolarlo in modo esatto, fatelo almeno in modo approssimato.

Risultati.

(a) Per essere la densità di probabilità di una v.a. assolutamente continua, la funzione f deve soddisfare le due seguenti condizioni:

- *positività*:

$$f(x) \geq 0\ \forall x \in \mathbb{R} \quad \Leftrightarrow \quad \frac{c}{94}(16 - x^2) \geq 0\ \forall x \in [-1\,,\,1] \quad \Leftrightarrow \quad c \geq 0$$

in quanto $16 - x^2 \geq 0$ se $x \in [-1\,,\,1]$;

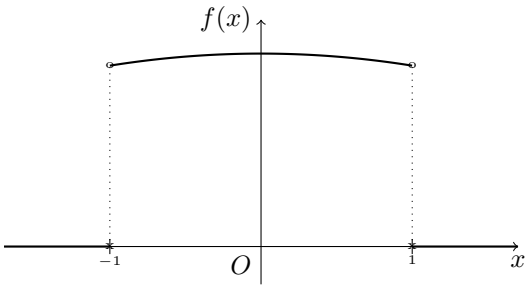
- *normalizzazione*:

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1 \quad \Leftrightarrow \quad \int_{-1}^{1} \frac{c}{94}(16-x^2)\mathrm{d}x = 1 \quad \Leftrightarrow \quad \frac{c}{94}\left[16x - \frac{x^3}{3}\right]_{x=-1}^{x=1} = 1 \quad \Leftrightarrow \quad \frac{c}{3} = 1$$
$$\Leftrightarrow \quad c = 3\,.$$

Vediamo dunque che $c = 3$ è l'unico valore che soddisfa entrambe le condizioni. Per tale valore abbiamo

$$f(x) = \begin{cases} \dfrac{3}{94}(16-x^2) & \text{se } x \in [-1\,,\,1] \\ 0 & \text{altrimenti} \end{cases}$$

il cui grafico è



(b) Per calcolare le probabilità richieste, bisognerebbe calcolare i due integrali

$$\mathbb{P}\,(X > 0) = \int_0^{+\infty} f(x)\mathrm{d}x\,, \qquad\qquad \mathbb{P}\,(|X| < 2) = \mathbb{P}\,(-2 < X < 2) = \int_{-2}^{2} f(x)\mathrm{d}x$$

Tuttavia, il primo integrale si calcola immediatamente osservando che la densità f è simmetrica rispetto all'asse delle y, e dunque $\mathbb{P}\,(X > 0) = 1/2$. Per calcolare il secondo, invece, basta osservare che f è tutta concentrata nell'intervallo $[-1\,,\,1] \subset [-2\,,\,2]$, e dunque $\mathbb{P}\,(-2 < X < 2) = 1$.

(c) La media

$$\mathbb{E}\,[X] = \int_{-\infty}^{+\infty} f(x)\mathrm{d}x$$

si può di nuovo calcolare senza svolgere i conti dell'integrale, osservando anche questa volta che f è simmetrica rispetto all'asse delle y, e dunque $\mathbb{E}\,[X] = 0$. Per la varianza, invece, abbiamo

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}\,[X]^2 = \mathbb{E}\left[X^2\right] = \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x$$
$$= \int_{-1}^{1} \frac{3}{94}x^2(16-x^2)\mathrm{d}x = \frac{3}{94}\left[\frac{16x^3}{3} - \frac{x^5}{5}\right]_{x=-1}^{x=1} = \frac{77}{235}\,.$$

(d) Per le usuali proprietà della media e della varianza, abbiamo

$$\mathbb{E}\,[P] = \mathbb{E}\,[4+X] = 4 + \mathbb{E}\,[X] = 4$$
$$\mathrm{Var}(P) = \mathrm{Var}(4+X) = \mathrm{Var}(X) = \frac{77}{235}\,.$$

(e) Se P_i è la potenza consumata dall'i-esima lampadina, con $i = 1, 2, \ldots, 50$, allora il consumo totale è la v.a. $S = P_1 + P_2 + \ldots + P_{50}$. Si richiede di calcolare la probabilità

$$\begin{aligned}\mathbb{P}(S > 195) &= \mathbb{P}\left(\underbrace{\frac{S - \mathbb{E}[S]}{\sqrt{\mathrm{Var}(S)}}}_{\substack{\approx N(0,1) \\ \text{per il TLC}}} > \frac{195 - \mathbb{E}[S]}{\sqrt{\mathrm{Var}(S)}}\right) \simeq 1 - \Phi\left(\frac{195 - \mathbb{E}[S]}{\sqrt{\mathrm{Var}(S)}}\right) = 1 - \Phi\left(\frac{195 - 50 \cdot \mathbb{E}[P_i]}{\sqrt{50 \cdot \mathrm{Var}(P_i)}}\right) \\ &= 1 - \Phi\left(\frac{195 - 50 \cdot 4}{\sqrt{50 \cdot (77/235)}}\right) = 1 - \Phi(-1.235) = \Phi(1.235) \simeq \frac{0.89065 + 0.89251}{2} \\ &= 89.158\%\,.\end{aligned}$$

(f) Per una ben nota proprietà della media, abbiamo

$$\begin{aligned}\mathbb{E}[R] &= \mathbb{E}\left[\frac{48400}{4 + X}\right] = 48400 \cdot \mathbb{E}\left[\frac{1}{4 + X}\right] = 48400 \cdot \int_{-\infty}^{+\infty} \frac{1}{4 + x} f(x)\mathrm{d}x \\ &= 48400 \cdot \int_{-1}^{1} \frac{1}{4 + x} \cdot \frac{3}{94}(16 - x^2)\mathrm{d}x = \frac{72600}{47} \int_{-1}^{1} (4 - x)\mathrm{d}x = \frac{72600}{47}\left[4x - \frac{x^2}{2}\right]_{x=-1}^{x=1} = \frac{72600}{47} \cdot 8 \\ &\simeq 12\,357.45\,\Omega\,.\end{aligned}$$

In alternativa, il valore atteso si poteva calcolare in modo approssimato utilizzando il metodo delta. Infatti,

$$\mathbb{E}[R] = \mathbb{E}\left[\frac{48400}{4 + X}\right] \underset{\delta}{\simeq} \frac{48400}{4 + \mathbb{E}[X]} = \frac{48400}{4 + 0} = 12\,100\,\Omega\,.$$

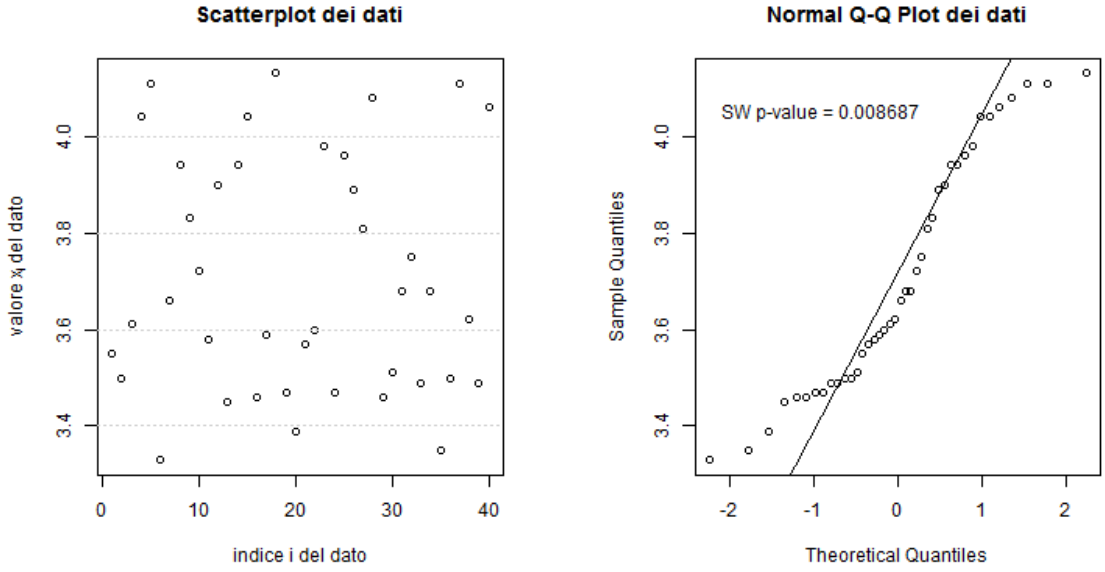
Problema 2. Lo stabilimento siderurgico XYZ produce e vende rotoli d'acciaio di qualità semidolce, cioè contenenti un tenore di carbonio (= percentuale di carbonio nella lega d'acciaio) che non deve essere superiore al 4.00 ‰.

Per essere ragionevolmente sicuri di rispettare questo standard, i tecnici hanno programmato l'impianto per generare rotoli con un tenore di carbonio atteso μ minore del 3.60 ‰. Inoltre, per controllare periodicamente che μ non superi questa soglia, ogni mese essi analizzano 40 rotoli a caso tra gli ultimi prodotti e ne verificano la compatibilità con l'ipotesi che μ sia minore del 3.60 ‰. Naturalmente, per i tecnici l'errore più grave consiste nel concludere che l'impianto debba essere revisionato quando in realtà non ce n'è alcun bisogno.

Nel controllo periodico di questo mese, le analisi dei 40 rotoli hanno dato i risultati nella tabella seguente. A fianco, sono riportate anche la media e la deviazione standard campionaria dei dati in tabella (tutti i valori sono in ‰):

3.55	3.50	3.61	4.04	4.11	3.33	3.66	3.94	3.83	3.72	
3.58	3.90	3.45	3.94	4.04	3.46	3.59	4.13	3.47	3.39	$\bar{x} = 3.7075$
3.57	3.60	3.98	3.47	3.96	3.89	3.81	4.08	3.46	3.51	$s = 0.2443909$
3.68	3.75	3.49	3.68	3.35	3.50	4.11	3.62	3.49	4.06	

Sotto, infine, si riporta lo scatterplot e il normal Q-Q plot col risultato del test di Shapiro-Wilks per i dati precedenti:



(a) Si può affermare in base ai dati che questo mese il tenore di carbonio in un rotolo qualsiasi è una variabile aleatoria gaussiana?

(b) Impostate un opportuno test per stabilire se i dati sono compatibili con l'ipotesi che questo mese il tenore di carbonio atteso in un rotolo qualsiasi sia minore del 3.60 ‰, o se al contrario c'è evidenza che l'impianto abbia bisogno di essere revisionato. Scrivete le ipotesi nulla e alternativa del test e la regola di rifiuto al livello α.

(c) A quale tipo di campione si applica il test del punto precedente? Si può applicare ai dati in tabella?

(d) Calcolate il p-value del test del punto (b) e traetene una conclusione. Si tratta di una conclusione debole o forte?

Prima di essere consegnato ai clienti, ogni rotolo viene ulteriormente analizzato da un dispositivo elettronico, che ne controlla l'effettivo tenore di carbonio X. Se X non supera il 4.00 ‰, il rotolo viene consegnato; altrimenti, esso viene automaticamente scartato e mandato in fonderia per essere rifuso.

(e) Fornite una stima puntuale e una stima intervallare al livello di confidenza del 90% per la probabilità che un rotolo qualsiasi tra quelli prodotti questo mese venga scartato.

Risultati.

(a) Vediamo che nel normal Q-Q plot i dati non si allineano lungo la Q-Q line, e inoltre il p-value del test di Shapiro-Wilk è molto basso (solo lo 0.8687%). Dunque non possiamo accettare l'ipotesi nulla che X sia una v.a. gaussiana (= forte evidenza contro l'ipotesi nulla di normalità).

(b) Dobbiamo fare un test per le ipotesi statistiche

$$H_0 : \mu \leq 3.60\,‰ =: \mu_0 \qquad \text{contro} \qquad H_1 : \mu > \mu_0\,,$$

in quanto siamo disposti a fermare e revisionare l'impianto solo se c'è forte evidenza che esso non funzioni (e dunque mettiamo $\mu > 3.60\,‰$ nell'ipotesi alternativa). Benché il campione non sia gaussiano, esso è comunque numeroso ($n = 40 > 30$). Dunque possiamo fare un T-test approssimato e usare la seguente regola di rifiuto al livello α:

$$\text{“ rifiuto } H_0 \text{ se } T_0 := \frac{\bar{X} - \mu_0}{S}\sqrt{n} > z_{1-\alpha}\text{ ”}. \qquad (*)$$

(c) Come già detto nel punto precedente, il test si applica a un campione numeroso.

(d) Il p-value del test con la regola di rifiuto $(*)$ è il valore di α che soddisfa l'equazione

$$t_0 = z_{1-\alpha} \quad \Leftrightarrow \quad \Phi(t_0) = \Phi(z_{1-\alpha}) = 1 - \alpha \Leftrightarrow \quad \alpha = 1 - \Phi(t_0)\,.$$

Dobbiamo dunque calcolare la realizzazione della statistica test t_0 coi dati misurati:

$$t_0 = \frac{3.7075 - 3.60}{0.2443909}\sqrt{40} = 2.7820\,,$$

ottenendo

$$p\text{-value} = 1 - \Phi(2.7820) = 1 - 0.99728 = 0.00272 = 0.272\%\,.$$

Con un p-value così basso non è possibile accettare H_0, e bisogna dunque accettare l'ipotesi alternativa che $\mu > 3.60\,‰$, cioè che l'impianto vada fermato e revisionato. Avendo accettato H_1, la conclusione del test è *forte*.

(e) Si richiede di calcolare una stima puntuale e una intervallare per la probabilità $p = \mathbb{P}\,(X > 4.00\,‰)$. Si tratta dunque di stimare il parametro p di un campione bernoulliano numeroso $Y_1, \ldots, Y_{40}$, dove

$$Y_i = \begin{cases} 1 & \text{se } X_i > 4.00\,‰ \\ 0 & \text{se } X_i \leq 4.00\,‰ \end{cases} \sim B(1, p)\,.$$

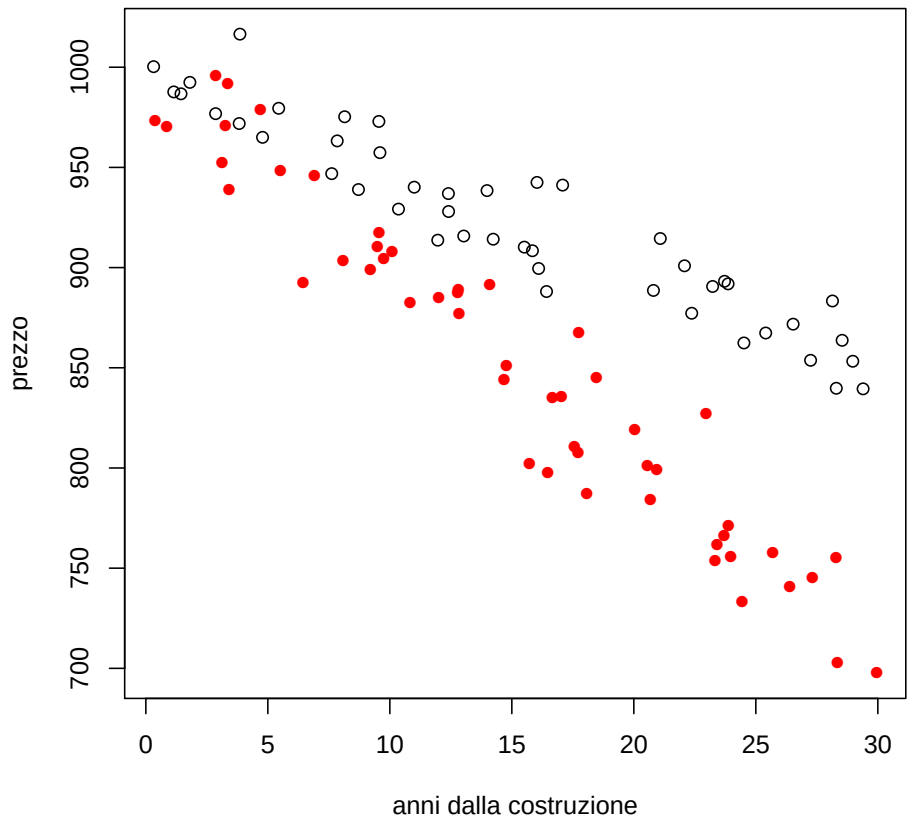
Una stima puntuale di p è data dalla frequenza campionaria

$$\bar{y} = \frac{\text{numero rotoli scartati}}{\text{numero rotoli analizzati}} = \frac{7}{40} = 0.17500 = 17.500\%\,.$$

Invece, una stima intervallare al livello del 90% è l'$IC_p(\gamma = 90\%)$ bilatero seguente:

$$p \in \left(\bar{y} \pm z_{\frac{1+\gamma}{2}}\sqrt{\frac{\bar{y}(1-\bar{y})}{n}}\right) = \left(\frac{7}{40} \pm z_{\frac{1+0.90}{2}}\sqrt{\frac{\frac{7}{40}\left(1-\frac{7}{40}\right)}{40}}\right) = (0.17500 \pm 1.645 \cdot 0.06008)$$
$$= (0.17500 \pm 0.09883) = (7.617\%\,,\, 27.383\%)\,.$$

Problema 3. Simone sta cercando un trilocale di circa 90 m^2 a Milano. La sua agenzia immobiliare di fiducia gli ha fornito i dati riguardanti 96 trilocali locati in due diverse zone del centro di Milano. In particolare, Simone è interessato a studiare come varia il `prezzo` dell'appartamento (in migliaia di euro), in dipendenza del numero di `anni` da cui è stato costruito e della `zona` in cui si trova. In figura sono rappresentati i dati raccolti. *Per tutti i modelli considerati si suppongano verificate le ipotesi del modello lineare gaussiano.*



(a) Impostare un modello di regressione semplice tra `prezzo` e `anni`.

(b) Sulla base dell'output riportato in Figura 1, il modello risulta un buon modello in termini di variabilità spiegata? Perché?

(c) Impostare un secondo modello di regressione lineare gaussiano con risposta il `prezzo` e regressori `anni`, `zona` e l'interazione tra `anni` e `zona`. Impostare infine un terzo modello con regressori solo `anni` e interazione tra `anni` e `zona`.

(d) Sulla base dell'output riportato in Figura 1, i regressori nei due modelli del punto (c) sono tutti significativi? Perché?

(e) Sulla base delle risposte precedenti, quale di tutti i tre modelli risulta il migliore?

(f) Per il modello scelto, dare una previsione puntuale del prezzo per un appartamento costruito da 5 anni che si trovi nella zona B.

```
Call:
lm(formula = prezzo ~ anni)

Residuals:
     Min       1Q   Median       3Q      Max
-113.158  -49.684    1.634   45.577  122.322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 992.4221    12.8721  77.098  < 2e-16 ***
anni         -6.2226     0.7443  -8.361 5.56e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.62 on 94 degrees of freedom
Multiple R-squared:  0.4265,    Adjusted R-squared:  0.4204
F-statistic:  69.9 on 1 and 94 DF,  p-value: 5.556e-13
```

```
Call:
lm(formula = prezzo ~ anni + zona + anni:zona)

Residuals:
    Min      1Q  Median      3Q     Max
-42.283 -12.756  -1.264  10.632  51.872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 997.3432     5.4930 181.567  < 2e-16 ***
anni         -2.9456     0.3148  -9.356 5.17e-15 ***
zonaB        -1.8721     7.7844  -0.240     0.81
anni:zonaB   -6.6447     0.4499 -14.768  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.62 on 92 degrees of freedom
Multiple R-squared:  0.9487,    Adjusted R-squared:  0.9471
F-statistic: 567.6 on 3 and 92 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = prezzo ~ anni + anni:zona)

Residuals:
    Min      1Q  Median      3Q     Max
-42.464 -12.578  -1.299  10.942  52.040

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 996.4111     3.8724  257.31   <2e-16 ***
anni         -2.8995     0.2485  -11.67   <2e-16 ***
anni:zonaB   -6.7390     0.2190  -30.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.53 on 93 degrees of freedom
Multiple R-squared:  0.9487,    Adjusted R-squared:  0.9476
F-statistic:   860 on 2 and 93 DF,  p-value: < 2.2e-16
```

Figura 1: Summary dei tre modelli del Problema 3.

Risultati.

(a) Il modello di regressione semplice è

$$\texttt{prezzo}_i = \beta_0 + \beta_1 \texttt{anni}_i + \varepsilon_i \qquad \varepsilon_1, \ldots, \varepsilon_{96} \text{ i.i.d. } \sim N(0, \sigma^2)\,.$$

(b) La variabilità spiegata dal modello di regressione semplice è il valore del suo R^2, e cioè $r^2 = 0.4265 = 42.65\%$. Questo valore è molto al di sotto dell'81%, e dunque il modello NON spiega bene la variabilità dei dati di output.

(c) Le relazioni ipotizzate dai due modelli sono rispettivamente:

$$\begin{aligned} &\texttt{prezzo}_i = \beta_0 + \beta_1 \texttt{anni}_i + \beta_2 \texttt{zona}_i + \beta_3 \texttt{anni}_i \cdot \texttt{zona}_i + \varepsilon_i \\ &\texttt{prezzo}_i = \beta_0 + \beta_1 \texttt{anni}_i + \beta_2 \texttt{anni}_i \cdot \texttt{zona}_i + \varepsilon_i \end{aligned} \qquad \varepsilon_1, \ldots, \varepsilon_{96} \text{ i.i.d. } \sim N(0, \sigma^2)\,.$$

(d) No, nel primo modello di regressione multipla il regressore `zona` non è significativo, perché il p-value del test per le ipotesi

$$H_0 : \beta_2 = 0 \qquad \text{contro} \qquad H_1 : \beta_2 \neq 0$$

è troppo elevato (p-value $= 81\% \gg 5\%$)

(e) Il migliore dei tre modelli è il terzo. Infatti, il modello di regressione semplice ha un R^2_{adj} pari solo al 42.04%, e dunque spiega la variabilità degli output molto meno dei modelli di regressione multipla, che hanno invece R^2_{adj} simili e pari rispettivamente al 94.71% e al 94.76%. Dei due modelli di regressione multipla, poi, solo l'ultimo ha tutti i predittori significativi, mentre come abbiamo visto nel punto precedente nell'altro il regressore `zona` non è significativo.

(f) Usando il terzo modello, la previsione puntuale richiesta in corrispondenza del nuovo input $\texttt{anni}^* = 5$ e $\texttt{zona}^* = 1$ è

$$\widehat{\texttt{prezzo}}^* = \hat\beta_0 + \hat\beta_1 \texttt{anni}^* + \hat\beta_2 \texttt{anni}^* \cdot \texttt{zona}^* = 996.4111 - 2.8995 \cdot 5 - 6.7390 \cdot 5 \cdot 1 = 948.2186$$