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

V appello di Statistica per Ingegneria Fisica
19 Febbraio 2020

Cognome, Nome e Numero di matricola:

Problema 1. Daniela è addetta alla vendita di cucine in un piccolo negozio d'arredamento. Svolgendo questo lavoro ormai da molti anni, ha imparato che il numero di cucine vendute in una settimana qualsiasi si può modellizzare con una variabile aleatoria avente densità di Poisson. Sa inoltre che le vendite realizzate in settimane diverse si possono assumere indipendenti e identicamente distribuite. Infine, l'esperienza le ha insegnato che in un anno (= 52 settimane) mediamente riesce a vendere 15.6 cucine.

(a) Sia X il numero di cucine vendute da Daniela la prossima settimana. Calcolare il valore atteso e la varianza di X.

(b) Calcolare la probabilità che la prossima settimana Daniela non riesca a vendere nemmeno una cucina.

(c) Calcolare la probabilità che Daniela non riesca a vendere nemmeno una cucina per 2 o più delle prossime 4 settimane.

(d) Sia Y il numero di cucine vendute da Daniela in un anno. Qual è la densità *esatta* della variabile aleatoria Y?

(e) Il capo ha promesso a Daniela che le aumenterà lo stipendio se l'anno prossimo riuscirà a vendere almeno 20 cucine in tutto. Calcolate la probabilità (eventualmente *approssimata*) che Daniela ottenga l'aumento.

Risultati.

(a) Se X_i è il numero di cucine vendute nell'i-esima settimana, con $i = 1, 2, \ldots, 52$, allora il numero di cucine vendute in un anno è la v.a.

$$Y = X_1 + X_2 + \ldots + X_{52}\,. \qquad (*)$$

Sappiamo dal testo che le v.a. $X_1, X_2, \ldots, X_{52}$ sono i.i.d., e che inoltre $\mathbb{E}[Y] = 15.6$. Di conseguenza,

$$\begin{aligned} 15.6 = \mathbb{E}[Y] &= \mathbb{E}[X_1 + X_2 + \ldots + X_{52}] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots + \mathbb{E}[X_{52}] = 52\,\mathbb{E}[X_i] \\ \Rightarrow \quad \mathbb{E}[X_i] &= \frac{15.6}{52} = 0.3\,. \end{aligned}$$

Poiché il valore atteso e la varianza della densità di Poisson sono uguali tra loro, abbiamo anche $\mathrm{Var}(X_i) = 0.3$. In altre parole, $\mathbb{E}[X] = \mathrm{Var}(X) = 0.3$.

(b) Dobbiamo calcolare la probabilità

$$\mathbb{P}(X = 0) = p_X(0) = \mathrm{e}^{-0.3}\,\frac{0.3^0}{0!} = \mathrm{e}^{-0.3} = 74.082\%\,,$$

dove $p_X(k) = \mathrm{e}^{-0.3}\,0.3^k/k!$ è la densità di X.

(c) Sia Z la v.a. che conta il numero di settimane tra le prossime 4 in cui Daniela non riesce a vendere nemmeno una cucina. Allora $Z \sim B(4, q)$, dove $q = \mathrm{e}^{-0.3}$ è la probabilità trovata al punto precedente. La probabilità richiesta è

$$\begin{aligned} \mathbb{P}(Z \geq 2) &= \sum_{k=2}^{4} p_Z(k) = 1 - \sum_{k=0}^{1} p_Z(k) = 1 - \sum_{k=0}^{1} \binom{4}{k} q^k (1-q)^{4-k} = 1 - \left[(1-q)^4 + 4q(1-q)^3\right] \\ &= 1 - \left[(1-\mathrm{e}^{-0.3})^4 + 4\mathrm{e}^{-0.3}(1-\mathrm{e}^{-0.3})^3\right] = 94.390\%\,. \end{aligned}$$

(d) Come visto nell'equazione $(*)$, il numero di cucine Y vendute in un anno è la somma delle 52 v.a. i.i.d. $X_1, X_2, \ldots, X_{52}$, in cui ciascuna X_i ha densità $\mathcal{P}(0.3)$. Per la ben nota proprietà di riproducibilità della densità di Poisson (= la somma di v.a. poissoniane indipendenti ha ancora densità di Poisson), la densità di Y è anch'essa poissoniana. Sappiamo dal testo che $\mathbb{E}[Y] = 15.6$. Di conseguenza,

$$Y \sim \mathcal{P}(15.6)\,.$$

(e) Poiché $15.6 \gg 5$, possiamo usare il TLC per approssimare la densità $\mathcal{P}(15.6)$ con la densità normale, cioè

$$\mathcal{P}(15.6) \approx N(15.6\,,\, 15.6)$$

(si ricordi che, per la densità poissoniana, valore atteso e varianza coincidono). Con la correzione di continuità, la probabilità richiesta è dunque

$$\begin{aligned}\mathbb{P}(Y \geq 20) = \mathbb{P}(Y \geq 19.5) &= \mathbb{P}\left(\underbrace{\frac{Y - \mathbb{E}[Y]}{\sqrt{\mathrm{Var}(Y)}}}_{\approx N(0,1)} \geq \frac{19.5 - 15.6}{\sqrt{15.6}}\right) \simeq 1 - \Phi\left(\frac{19.5 - 15.6}{\sqrt{15.6}}\right) \\ &= 1 - \Phi(0.9874) \simeq 1 - 0.83891 = 16.109\%\,.\end{aligned}$$

Senza correzione di continuità, invece,

$$\mathbb{P}(Y \geq 20) = \simeq 1 - \Phi\left(\frac{20 - 15.6}{\sqrt{15.6}}\right) = 1 - \Phi(1.1140) \simeq 1 - \frac{0.86650 + 0.86864}{2} = 13.243\%\,.$$

Problema 2. La ACME è una grande multinazionale che produce apparecchiature per le radiocomunicazioni. Recentemente, i suoi tecnici hanno ideato un nuovo prototipo di antenna per le trasmissioni a lungo raggio, che dovrebbe sensibilmente migliorare le prestazioni di tutti i modelli d'antenna precedenti.

Per testare la loro invenzione, i tecnici hanno installato due antenne del nuovo tipo a 50 Km di distanza l'una dall'altra, e hanno poi inviato dalla prima antenna un segnale formato da 128 bit, tutti zero:

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Questo è il corrispondente segnale di 128 bit che hanno ricevuto dall'altra antenna:

0 0 0 0 0 0 0 0 0 0 **1** 0 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 **1** 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0

Se la seconda antenna ha ricevuto 1 anziché 0, significa che quel bit si è corrotto durante la trasmissione.

(a) Determinate un intervallo di confidenza bilatero al livello del 95% per la probabilità che un bit inviato a 50 Km di distanza col nuovo prototipo d'antenna si corrompa durante la trasmissione.

(b) I tecnici vorrebbero che l'ampiezza dell'intervallo precedente non superasse il 5%. Possono dunque ritenersi soddisfatti di quanto ottenuto col loro segnale di 128 bit? In caso contrario, qual è il numero minimo di bit sicuramente sufficienti a ottenere ciò che vogliono?

(c) Con i vecchi modelli di antenna della ACME, più del 10% dei bit inviati a 50 Km di distanza venivano corrotti durante la trasmissione. Impostate un opportuno test al livello α per stabilire dai dati se il nuovo prototipo abbia prestazioni significativamente migliori dei vecchi modelli.

(d) Qual è il p-value del test precedente? Con un valore simile, secondo voi è ragionevole che la ACME investa per avviare la produzione del nuovo prototipo su larga scala?

(e) Quale sarebbe la potenza del test del punto (c), fatto al livello di significatività $\alpha = 5\%$, se un bit trasmesso a 50 Km di distanza con la nuova antenna avesse una probabilità di corrompersi pari solamente al 6%?

Risultati.

(a) Si tratta di un intervallo di confidenza per il parametro p di un campione bernoulliano numeroso $X_1, X_2, \ldots, X_{128}$, in cui

$$X_i = \begin{cases} 0 & \text{se l'}i\text{-esimo bit è trasmesso senza errori} \\ 1 & \text{se l'}i\text{-esimo bit si corrompe durante la trasmissione} \end{cases}$$

e dunque $X_i = B(1, p)$, dove p è la probabilità che un bit qualsiasi si corrompa durante la trasmissione. Un $IC_p(95\%)$ bilatero è

$$p \in \left(\bar{x} \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right) ,$$

in cui $\gamma = 0.95$, $n = 128$ e $\bar{x} = 9/128$ è la frequenza empirica di bit corrotti nel segnale trasmesso dai tecnici. Abbiamo $z_{\frac{1+0.95}{2}} = z_{0.975} = 1.96$, e dunque l'intervallo precedente diventa

$$p \in \left(\frac{9}{128} \pm 1.96 \sqrt{\frac{\frac{9}{128}\left(1 - \frac{9}{128}\right)}{128}} \right) = (0.07031 \pm 0.04429) = (2.602\% \,,\, 11.461\%) \,,$$

(b) L'ampiezza dell'intervallo precedente è $2 \cdot 0.04429 = 8.859\% > 5\%$, e dunque è un valore insoddisfacente per i tecnici della ACME. Per esser sicuri di ridurre tale ampiezza al valore massimo del 5%, i tecnici devono trovare n per cui

$$\text{ampiezza} = 2 \cdot 1.96 \sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}} \leq 0.05$$

qualunque sia l'effettivo valore di $\bar{x}$ che poi misureranno. Sappiamo che, poiché $\bar{x} \in [0,1]$, allora $\bar{x}(1-\bar{x}) \in [0, 1/4]$. Di conseguenza,

$$\text{ampiezza} = 2 \cdot 1.96 \sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}} \leq 2 \cdot 1.96 \sqrt{\frac{\frac{1}{4}}{n}} = \frac{1.96}{\sqrt{n}}\,.$$

Per essere sicuri che l'ampiezza non superi il 5% è dunque sufficiente che

$$\frac{1.96}{\sqrt{n}} \leq 0.05 \quad \Leftrightarrow \quad n \geq \left(\frac{1.96}{0.05}\right)^2 = 1536.64\,.$$

Occorrono dunque almeno 1537 bit.

(c) Sappiamo dal testo del problema che coi vecchi modelli della ACME la probabilità p di errore per un singolo bit era maggiore del 10%. Poiché vogliamo verificare se il nuovo prototipo abbia prestazioni *significativamente migliori* dei vecchi modelli (cioè cerchiamo *evidenza forte* del fatto che la nuovo antenna sia effettivamente migliore di quelle vecchie), scegliamo quest'ultima affermazione come ipotesi alternativa del nostro test:

$$H_0 : p \geq 10\% =: p_0 \qquad \text{vs.} \qquad H_0 : p < p_0\,.$$

Possiamo dunque fare un test per la frequenza incognita di un campione bernoulliano numeroso, che in questo caso al livello α consiste nella regola

$$\text{“ rifiuto } H_0 \text{ se } Z_0 := \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} < -z_{1-\alpha}\text{ ”}\,.$$

(d) Il p-value del test precedente si ottiene dalla regola di rifiuto, uguagliando il quantile con la realizzazione della statistica test:

$$z_0 = -z_{1-\alpha} \quad \Leftrightarrow \quad \Phi(-z_0) = \Phi(z_{1-\alpha}) = 1-\alpha \quad \Leftrightarrow \quad \alpha = 1 - \Phi(-z_0)\,,$$

dove

$$z_0 = \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{\frac{9}{128} - 0.10}{\sqrt{0.10 \cdot (1-0.10)}}\sqrt{128} = -1.1196\,.$$

Il p-value è pertanto

$$\alpha = 1 - \Phi(-(-1.1196)) = 1 - \Phi(1.1196) = 1 - 0.86864 = 13.136\%\,.$$

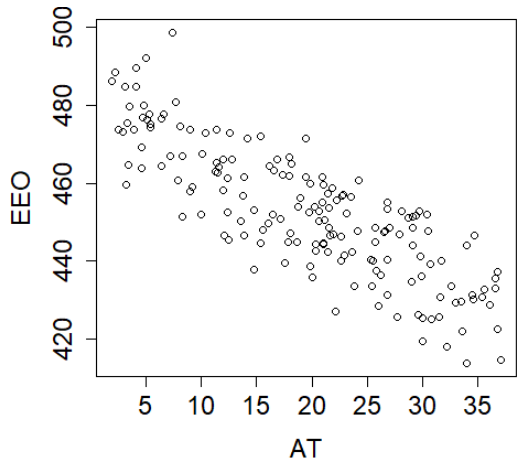
Con un p-value così alto, non c'è nessun motivo per dubitare di H_0. In conclusione, i dati non mostrano alcuna evidenza che il nuovo prototipo sia migliore delle vecchie antenne.

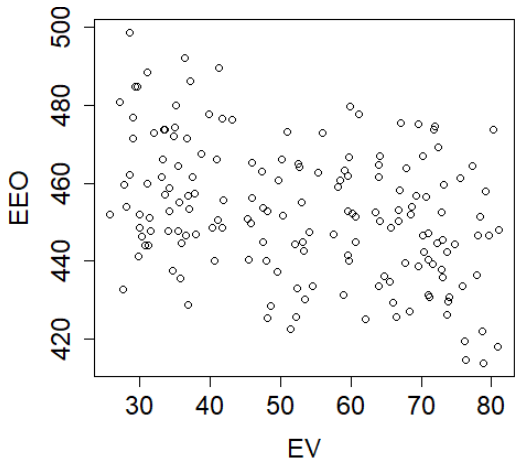
(e) La potenza del test al livello $\alpha = 5\%$ calcolata quando $p = 6\%$ è

$$\begin{aligned}
\pi(0.06) &= \mathbb{P}_{p=0.06}\,(\text{“rifiuto } H_0\text{”}) = \mathbb{P}_{p=0.06}\left(\frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} < -z_{1-0.05}\right) \\
&= \mathbb{P}_{p=0.06}\left(\bar{X} < -z_{0.95}\sqrt{\frac{p_0(1-p_0)}{n}} + p_0\right) \\
&= \mathbb{P}_{p=0.06}\left(\frac{\bar{X} - p}{\sqrt{p(1-p)}}\sqrt{n} < -z_{0.95}\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} + \frac{p_0 - p}{\sqrt{p(1-p)}}\sqrt{n}\right) \\
&= \mathbb{P}_{p=0.06}\left(\underbrace{\frac{\bar{X} - p}{\sqrt{p(1-p)}}\sqrt{n}}_{\approx N(0,1)} < -1.645\sqrt{\frac{0.10 \cdot (1-0.10)}{0.06 \cdot (1-0.06)}} + \frac{0.10 - 0.06}{\sqrt{0.06 \cdot (1-0.06)}}\sqrt{128}\right) \\
&= \Phi\left(-1.645\sqrt{\frac{0.10 \cdot (1-0.10)}{0.06 \cdot (1-0.06)}} + \frac{0.10 - 0.06}{\sqrt{0.06 \cdot (1-0.06)}}\sqrt{128}\right) = \Phi(-0.1724) \\
&= 1 - \Phi(0.1724) = 1 - 0.56749 = 43.251\%\,.
\end{aligned}$$

dove abbiamo usato il fatto che, poiché $n = 128$ è grande, per il TLC si ha $\bar{X} \approx N\left(p\,,\, \frac{p(1-p)}{n}\right)$.

Problema 3. Si vuole monitorare la produzione di energia di una centrale termoelettrica a ciclo combinato. A tal fine, sono disponibili 168 misurazioni medie orarie di *Ambient Temperature* (`AT`) misurata in gradi centigradi, di *Exhaust Vacuum* (`EV`) misurata in centimetri di mercurio e delle corrispondenti produzioni nette orarie di energia elettrica (*Electrical Energy Output*, `EEO`) misurate in megawatt. Le figure sottostanti rappresentano i dati raccolti:





Nelle due pagine seguenti sono riportati gli output e i grafici di diagnostica dei residui relativi a tre modelli di regressione lineare: il primo modello considera il solo regressore `AT`, il secondo modello considera il solo regressore `EV` e il terzo modello considera entrambi i regressori.

(a) Quali sono le ipotesi alla base dei modelli considerati? Per quali modelli sono verificate?

(b) I tre modelli sono globalmente significativi? Giustificare la risposta.

(c) Indicare, per ciascun modello, la percentuale di variabilità spiegata.

(d) Quale dei tre modelli è preferibile? Giustificare la risposta.

(e) Scrivere il modello ipotizzato e quello stimato per il caso scelto al punto (d).

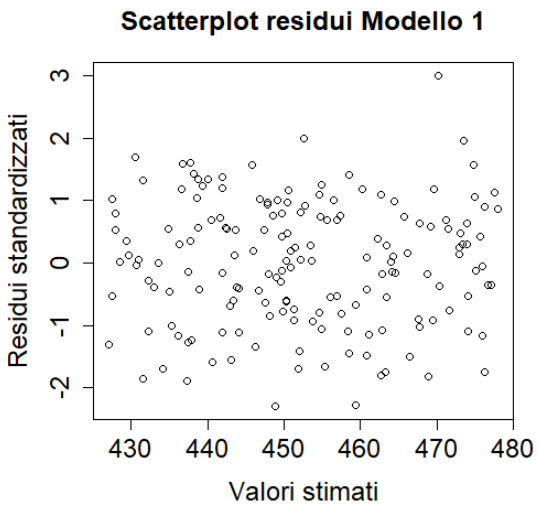
(f) Fornire una stima puntuale e un intervallo di confidenza bilatero (livello di confidenza 95%) per la produzione netta oraria di energia elettrica attesa per valori nulli dei regressori.

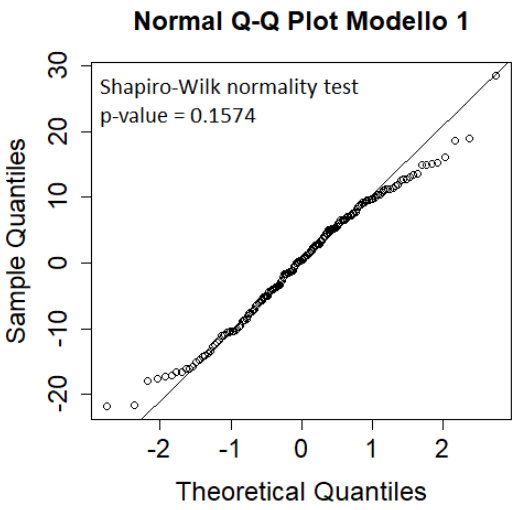
(g) Con un opportuno test (livello di significitatività 5%), verificare se c'è evidenza statistica che, per valori nulli dei regressori, la produzione netta oraria di energia elettrica attesa sia diversa da 500 megawatt.

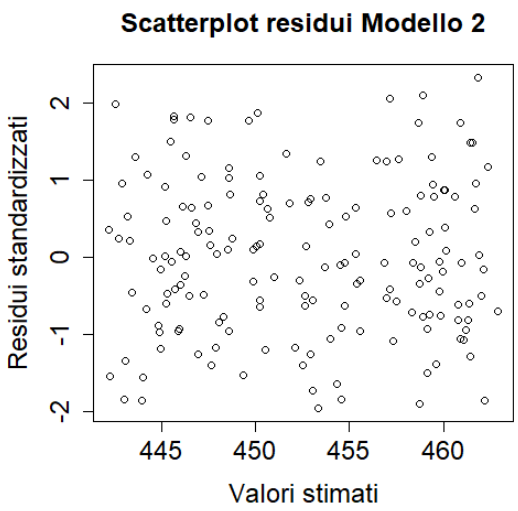
```
Call:
lm(formula = EEO ~ AT)

Residuals:
     Min       1Q   Median       3Q      Max
-21.7543  -7.0888   0.4298   7.0607  28.5160

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 480.81383    1.64357  292.54   <2e-16 ***
AT           -1.44841    0.07563  -19.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.493 on 166 degrees of freedom
Multiple R-squared:  0.6884,    Adjusted R-squared:  0.6866
F-statistic: 366.8 on 1 and 166 DF,  p-value: < 2.2e-16
```

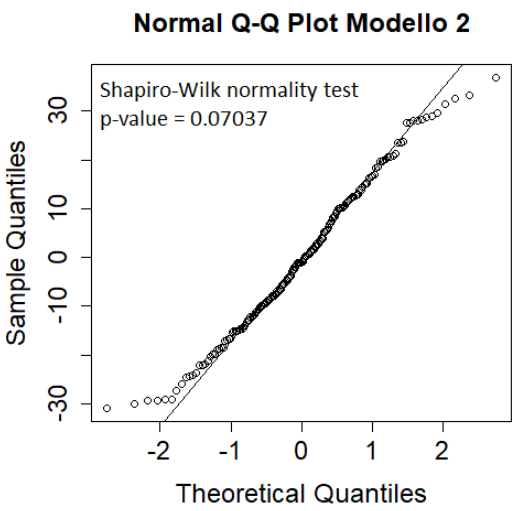
```
Call:
lm(formula = EEO ~ EV)

Residuals:
    Min      1Q  Median      3Q     Max
-30.855 -11.722  -0.996  11.706  36.829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 472.52158    4.12438 114.568  < 2e-16 ***
EV           -0.37371    0.07402  -5.048 1.16e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.83 on 166 degrees of freedom
Multiple R-squared:  0.1331,    Adjusted R-squared:  0.1279
F-statistic: 25.49 on 1 and 166 DF,  p-value: 1.162e-06
```

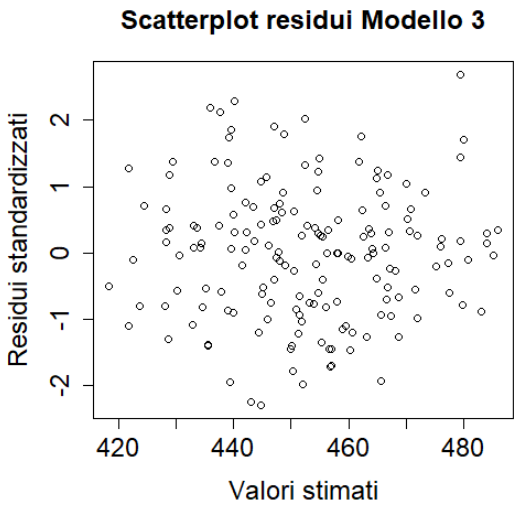
```
Call:
lm(formula = EEO ~ AT + EV)

Residuals:
     Min       1Q   Median       3Q      Max
-16.3772  -5.4103   0.4179   4.4175  19.2119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 500.92673    2.17320  230.50   <2e-16 ***
AT           -1.45092    0.05703  -25.44   <2e-16 ***
EV           -0.37703    0.03346  -11.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.158 on 165 degrees of freedom
Multiple R-squared:  0.8239,    Adjusted R-squared:  0.8218
F-statistic:   386 on 2 and 165 DF,  p-value: < 2.2e-16
```

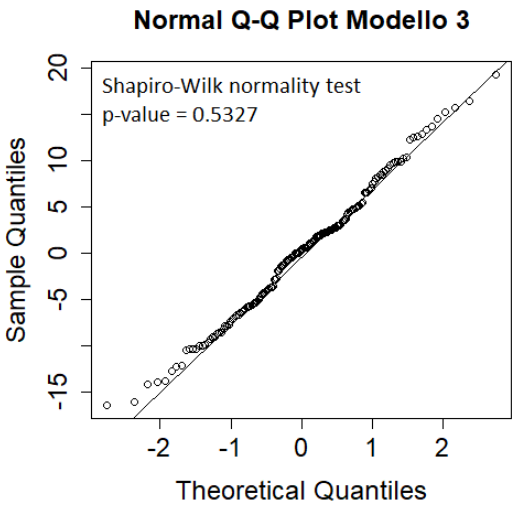

Scatterplot residui Modello 1
Residui standardizzati
Valori stimati


Normal Q-Q Plot Modello 1
Shapiro-Wilk normality test
p-value = 0.1574
Sample Quantiles
Theoretical Quantiles


Scatterplot residui Modello 2
Residui standardizzati
Valori stimati


Normal Q-Q Plot Modello 2
Shapiro-Wilk normality test
p-value = 0.07037
Sample Quantiles
Theoretical Quantiles


Scatterplot residui Modello 3
Residui standardizzati
Valori stimati


Normal Q-Q Plot Modello 3
Shapiro-Wilk normality test
p-value = 0.5327
Sample Quantiles
Theoretical Quantiles

Risultati.

(a) Le ipotesi alla base di un modello di regressione sono che i residui siano normali e omoschedastici. L'ipotesi di omoschedasticità è soddisfatta per tutti i modelli, come si può vedere dagli scatterplot dei residui che non presentano particolari pattern. Anche l'ipotesi di normalità è verificata per tutti i modelli, dato che i quantili teorici ed empirici nel normal Q-Q plot seguono l'andamento lineare e il p-value dello Shapiro-test è maggiore di 0.05.

(b) Tutti i modelli sono globalmente significativi. Infatti, gli F-test sulla significatività globale della regressione (che per i primi due modelli di regressione lineare semplice coincidono con i test sulla significatività del regressore) hanno p-value molto bassi: nel primo modello il p-value è inferiore a $2.2 \cdot 10^{-16}$ (`p-value: < 2.2e-16`), nel secondo modello il p-value è $1.162 \cdot 10^{-6}$ (`p-value: 1.162e-06`), nel terzo modello il p-value è inferiore a $2.2 \cdot 10^{-16}$ (`p-value: < 2.2e-16`).

(c) La percentuale di variabilità spiegata dal primo modello è 68.84% ($R^2 = 0.6884$). La percentuale di variabilità spiegata dal secondo modello è 13.31% ($R^2 = 0.1331$). La percentuale di variabilità spiegata dal terzo modello è 82.39%($R^2 = 0.8239$).

(d) Poiché per tutti i modelli sono soddisfatte le ipotesi di normalità e di omoschedasticità e tutti risultano globalmente significativi, sceglieremo il modello che risulta migliore in termini di variabilità spiegata. Per determinare il modello migliore in termini di variabilità spiegata, dobbiamo confrontare l'R^2_{adj} dei tre modelli, in quanto uno di essi (il terzo) coinvolge più di un predittore (due predittori). Abbiamo:

$$R^2_{adj,1} = 0.6866 \qquad R^2_{adj,2} = 0.1279 \qquad R^2_{adj,3} = 0.8218$$

Quindi, il migliore in termini di variabilità spiegata risulta essere il modello 3.

(e) Modello ipotizzato:

$$EEO = \beta_0 + \beta_1 AT + \beta_2 EV + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

Modello stimato:

$$\hat{EEO} = \hat{\beta}_0 + \hat{\beta}_1 AT + \hat{\beta}_2 EV = 500.92673 - 1.45092 AT - 0.37703 EV.$$

(f) La produzione netta oraria di energia elettrica per valori nulli dei regressori coincide con il parametro β_0. Una stima puntuale di β_0 è

$$\hat{\beta}_0 = 500.92673.$$

Un intervallo di confidenza bilatero per β_0 è

$$[\hat{\beta}_0 - t_{\frac{0.05}{2},165}\,\mathrm{se}(\hat{\beta}_0)\,;\, \hat{\beta}_0 + t_{\frac{0.05}{2},165}\,\mathrm{se}(\hat{\beta}_0)].$$

Dalla tavola della distribuzione t di Student si ricava $t_{\frac{0.05}{2},165} \simeq t_{\frac{0.05}{2},\infty} = 1.960$. Si ottiene quindi l'intervallo di confidenza:

$$[500.92673 - 1.960 \cdot 2.17320\,;\, 500.92673 + 1.960 \cdot 2.17320] = [496.6673\,;\, 505.1862].$$

(g) Bisogna fare un test per le ipotesi

$$H_0 : \beta_0 = 500 \qquad \text{contro} \qquad H_1 : \beta_0 \neq 500$$

Si tratta di un T-test sul coefficiente di regressione β_0. La statistica test è

$$t_0 = \frac{\hat{\beta}_0 - 500}{se(\hat{\beta}_0)} = \frac{500.92673 - 500}{2.17320} = 0.4264357$$

Rifiuto H_0 se $|t_0| \geq t_{\frac{0.05}{2},165}$ cioè se $0.4264357 \geq 1.960$. Dunque non possiamo rifiutare H_0. Ne concludiamo che al 5% di significatività non c'è evidenza statistica che la produzione netta oraria di energia elettrica attesa per valori nulli dei regressori sia diversa da 500 megawatt. Alternativamente, si poteva arrivare alla stessa conclusione osservando che il valore 500 è compreso nell'intervallo di confidenza bilatero per il parametro β_0 (con livello di confidenza pari a 1- livello significatività del test richiesto) calcolato al punto precedente.