Università degli Studi di Padova

Corso di Laurea Magistrale in *Informatica*

a.a. 2016/2017

Data Mining

Docente: Annamaria Guolo

Prova parziale del 20 aprile 2017: SOLUZIONE

ISTRUZIONI: La durata della prova è di 1 ora. La prova va svolta su questi fogli. Eventuali fogli di brutta copia possono essere richiesti, ma non verranno corretti. Non scrivere in matita. In caso di errore, barrare la parte errata, non utilizzare un correttore (bianchetto).

Nome:_____ Cognome:_____ Matricola:_____

Domande a risposta multipla
Solo una delle risposte è corretta. Segnare con una crocetta la risposta corretta. Le risposte sbagliate o non date valgono zero punti.

1) Nel modello di regressione lineare stimato ai minimi quadrati l'indice R^2 è pari a

(a) $\frac{\text{devianza spiegata}}{\text{devianza totale}}$ (b) $\frac{\text{devianza totale}}{\text{devianza residua}}$ (c) $\frac{\text{devianza residua}}{\text{devianza spiegata}}$ (d) $\frac{\text{devianza residua}}{\text{devianza totale}}$

La risposta corretta è la (a).

2) La verifica d'ipotesi per il confronto tra due modelli lineari annidati condotta tramite la statistica F rifiuta l'ipotesi nulla di passaggio dal modello più grande al modello più piccolo al livello di significatività α

(a) per valori alti e bassi di F
(b) per valori bassi di F
(c) per valori alti di F
(d) per valori di F minori di $\alpha = 0.05$

La risposta corretta è la (c).

3) In un modello di regressione lineare, il problema della multicollinearità deriva da

(a) bassa correlazione tra gli errori ε e la risposta
(b) bassa correlazione tra tutte le esplicative
(c) alta correlazione tra almeno due esplicative
(d) bassa correlazione tra almeno un errore ε e le esplicative

La risposta corretta è la (c).

4) Nel modello di regressione lineare $Y = \beta_0 + \beta_1 X + \varepsilon$, il livello di significatività osservato (p-value) pari a 0.83 per il test $H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$ suggerisce

(a) di eliminare X dal modello
(b) di mantenere X nel modello
(c) di eliminare β_0 dal modello
(d) che vi sono osservazioni anomale

La risposta corretta è la (b).

5) Il *residual standard error* (RSE) per un modello $Y = \beta_0 + \beta_1 X + \varepsilon$ con errori che si assumono $N(0, \sigma^2)$ e che viene stimato ai minimi quadrati è

(a) la stima di β_1 (b) la stima della media degli errori (c) la stima di σ
(d) il p-value associato al test di bontà di adattamento del modello

La risposta corretta è la (c).

Esercizio

Rispondere su questi fogli in modo conciso e chiaro. Per i calcoli, riportare tutti i passaggi, non solo il risultato finale.

Si considerino le informazioni su 145 auto usate e relative a

- `prezzo` (in centinaia di euro)
- `chilometri` percorsi (in migliaia)
- `cavalli` potenza: la variabile assume valore TRUE se i cavalli sono maggiori di 100 e FALSE se i cavalli sono minori o uguali a 100
- `anni`: la variabile assume valore A se l'età è minore o uguale a 3 anni, B se l'età è tra 4 e 5 anni (estremi inclusi), C se l'età è maggiore o uguale a 6 anni

a) Viene stimato un modello di regressione lineare per spiegare il prezzo dell'auto usata in funzione dei chilometri percorsi e della potenza del motore. Di seguito l'output fornito da `R`

```
Call:
lm(formula = prezzo ~ chilometri + cavalli, data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-46.599 -16.590  -4.753   9.116  89.268

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.09941    5.26824  20.709  < 2e-16 ***
chilometri   -0.39153    0.05393  -7.259 2.34e-11 ***
cavalliTRUE   9.49269    4.13367   2.296   0.0231 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.85 on 142 degrees of freedom
Multiple R-squared:  0.3134,
  Adjusted R-squared:  0.3038
F-statistic: 32.41 on 2 and 142 DF,  p-value: 2.536e-12
```

a.1) Scrivere l'espressione del modello stimato. Precisare come viene gestita la variabile qualitativa `cavalli` e quale livello (vale a dire quale potenza del motore) viene considerato di base.
Il modello stimato è

$$\widehat{\texttt{prezzo}} = 109.09941 - 0.39153 \times \texttt{chilometri} + 9.49269 \times I_{\texttt{cavalli=TRUE}}$$

dove $I_{\texttt{cavalli=TRUE}}$ è la variabile indicatrice che assume valore 1 se `cavalli=TRUE` e valore 0 se `cavalli=FALSE`. Questo perchè la variabile `cavalli` è una variabile qualitativa con due livelli, TRUE e FALSE, dei quali il livello FALSE è considerato come livello base da `R`, dato l'ordinamento alfabetico dei livelli. La variabile qualitativa con due livelli viene tradotta in una variabile numerica (quantitativa) che assume valori 0 e 1 come specificato sopra. Di conseguenza, il coefficiente 9.49269 indica l'aumento di prezzo per un'auto con oltre 100 cavalli rispetto ad un'auto con meno (o con esattamente) di 100 cavalli.

a.2) Commentare l'output evidenziando la significatività dei coefficienti e la possibilità di semplificazione del modello, interpretando i segni e i valori dei coefficienti stimati (vale a dire l'associazione delle esplicative con la risposta), valutando l'adattamento del modello tramite R^2.

L'output del modello indica che entrambi i coefficienti associati alle variabili esplicative sono significativamente diversi da 0, dato che il p-value (livello di significatività osservato) associato al test di verifica dipotesi per luguaglianza a zero dei coefficienti è basso. Di conseguenza il modello non è semplificabile. Se invece del classico livello di riferimento 0.05 con cui confrontare il p-value si considerasse il livello 0.01, allora il coefficiente associato a `cavalliTRUE`

sarebbe considerato pari a 0 e la corrispondente variabile si potrebbe eliminare dal modello.
I risultati indicano che al crescere dei chilometri percorsi il costo dell'auto diminuisce: in particolare, un aumento di 1000 chilometri percorsi diminuisce il prezzo di 0.39153 centinaia di euro (39.153 euro). Inoltre, se la potenza del motore supera i 100 cavalli allora il prezzo dell'auto aumenta di 9.49269 centinaia di euro rispetto a quello di un'auto con meno (o al più) di 100 cavalli.
Il valore basso dell'indice R^2 suggerisce che il modello non presenta un buon adattamento alle osservazioni: potrebbe essere migliorabile tramite l'inserimento di nuove variabili, l'inserimento di eventuali interazioni tra le variabili o di termini polinomiali per le variabili esplicative quantitative.

a.3) A quale verifica d'ipotesi si riferiscono i valori della statistica F e del suo p-value riportati nell'ultima riga dell'output? Come si interpretano i risultati?

Il valore si riferisce alla statistica che valuta l'uguaglianza a zero dei coefficienti associati alle esplicative, vale a dire $H_0 : \beta_1 = \beta_2 = 0$ (dove β_1 è il coefficiente associato a `chilometri` e β_2 il coefficiente associato a `cavalliTRUE`) contro l'alternativa H_1 che almeno uno dei due coefficienti sia non nullo. La statistica si scrive come

$$F = \frac{R^2}{1 - R^2} \frac{142}{2}$$

ed ha distribuzione $F_{2,142}$ sotto l'ipotesi nulla. Il valore del p-value molto basso indica che non è giustificabile il passaggio dal modello con due variabili esplicative al modello senza variabili esplicative.

a.4) Proporre un intervallo di confidenza di livello 0.90 per il parametro associato alla variabile `chilometri`, spiegando le eventuali assunzioni fatte.

Indicando con β_1 il coefficiente associato a `chilometri`, con $\hat{\beta}_1$ la sua stima e con $SE(\hat{\beta}_1)$ lo standard error, si ha che

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-3}$$

e quindi l'intervallo richiesto è

$$\hat{\beta}_1 - t_{0.95,n-3} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{0.95,n-3} SE(\hat{\beta}_1)$$

Poichè $n - 3 = 142$ è un valore molto elevato, la distribuzione t di Student si può approssimare con la distribuzione $N(0, 1)$ (normale standard): l'intervallo allora si può approssimare con il seguente intervallo che utilizza i quantili di $N(0, 1)$

$$\hat{\beta}_1 - z_{0.95} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + z_{0.95,n-3} SE(\hat{\beta}_1)$$

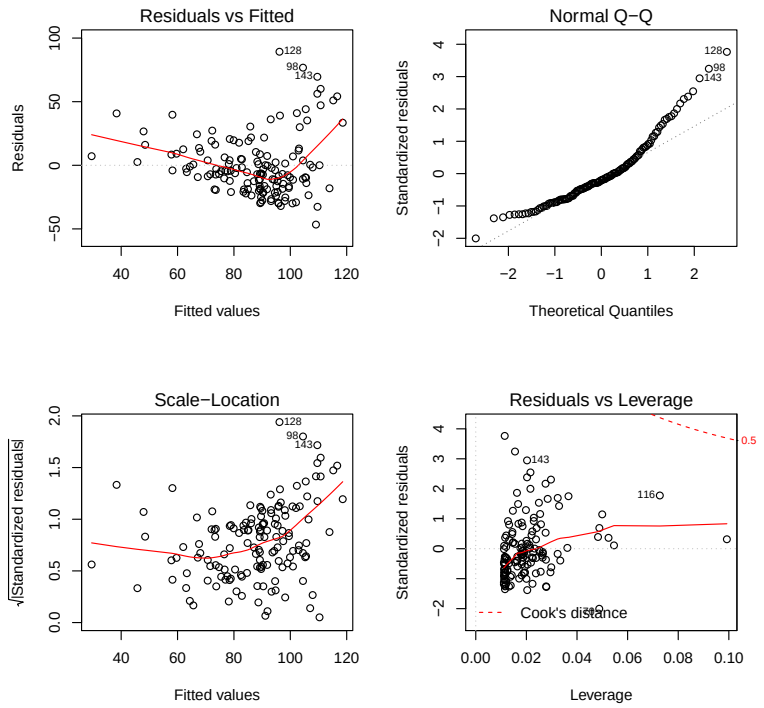
vale a dire

$$-0.39153 - 1.64 \times 0.05393 \leq \beta_1 \leq -0.39153 + 1.64 \times 0.05393$$

da cui si ricava

$$-0.4799752 \leq \beta_1 \leq -0.39153$$

a.5) Il seguente grafico riporta l'output di default di `R` riferito all'analisi dei residui del modello
Il modello presenta un buon adattamento? Presenta dei problemi? Se si, quali sono delle possibili soluzioni?



Il grafico dei residui indica che il modello non presenta un buon adattamento. Infatti, il primo grafico (grafico di dispersione dei residui) mostra un andamento deterministico che non dovrebbe in realtà esserci. Inoltre, la media dei residui non sembra pari a 0 e la varianza dei residui non appare costante, come dovrebbe essere sulla base delle ipotesi che il modello di regressione pone sugli errori ε. Inoltre, nemmeno la normalità dei residui è soddisfatta, come evidenziato nel secondo grafico: i quantili empirici, infatti, si discostano da quelli teorici di una normal standard. Possibili soluzioni prevedono l'inserimento di nuove variabili esplicative, eventualmente con interazioni e/o l'inserimento di termini polinomiali associati alle variabili quantitative.

Per completare l'analisi dei residui, non sembrano essere presenti valori anomali: sebbene `R` evidenzi delle osservazioni, queste non rappresentano osservazioni anomale dato che la distanza di Cook non è elevata.

b) L'estensione del modello con l'inclusione della variabile `anni` risulta (si veda pagina successiva)

b.1) Come si interpretano i valori dei coefficienti associati ai livelli della variabile `anni`? Sono valori ragionevoli?

I coefficienti indicano che il prezzo dell'auto diminuisce al crescere dell'età del veicolo, il che è in linea con quanto ci si attende nella realtà. In particolare, per un'auto di 4-5 anni il prezzo diminuisce di 40.4693 centinaia di euro rispetto ad un'auto di 3 anni al massimo, mentre per un'auto di 6 o più anni il prezzo diminuisce di 53.3891 centinaia di euro rispetto ad un'auto di 3 anni al massimo.

b.2) Confrontare i due modelli fin qui stimati calcolando la statistica F, spiegando la verifica d'ipotesi condotta e commentando il risultato. Considerare il livello di significatività 0.05.

Consideriamo la verifica d'ipotesi $H_0 : \beta_2 = \beta_3 = 0$ contro H_1 che considera almeno uno dei due coefficienti non nullo: in questo modo, sotto H_1 consideriamo che la variable `anni` sia inserita nel modello (basta che almeno uno dei coefficienti associati ai livelli sia significativo).

```
Call:
lm(formula = prezzo ~ chilometri + anni + cavalli, data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-44.173  -8.300  -0.102   6.807  55.455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 136.4331     4.0161  33.972  < 2e-16 ***
chilometri   -0.2215     0.0367  -6.036 1.34e-08 ***
anniB       -40.4693     3.5125 -11.522  < 2e-16 ***
anniC       -53.3891     3.7538 -14.223  < 2e-16 ***
cavalliTRUE   6.2281     2.6545   2.346   0.0204 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.22 on 140 degrees of freedom
Multiple R-squared:  0.7242,
  Adjusted R-squared:  0.7163
F-statistic: 91.91 on 4 and 140 DF,  p-value: < 2.2e-16
```

Il fatto che i modelli siano annidati permette di fare riferimento alla statistica

$$F = \frac{RSS_0 - RSS}{RSS}\frac{n-p-1}{q},$$

dove $q = 2$, $n - p - 1 = 145 - 4 - 1 = 140$, RSS_0 è la somma dei quadrati dei residui nel modello con meno variabili e RSS è la somma dei quadrati dei residui nel modello con più variabili. Considerando che $RSS_0 = RSE^2(n-p-1) = 23.85^2 \times 142 = 80772.8$ e che $RSS = 15.22^2 \times 140 = 32430.78$ il valore osservato sui dati di F è

$$\frac{80772.8 - 32430.78}{32430.78}\frac{140}{2} = 104.3435$$

Confrontiamo questo valore con il quantile 0.95 di una $F_{2,140}$, che vale 3.061. Poichè il valore osservato è maggiore, si rifiuta l'ipotesi nulla al livello 0.05: la variabile `anni` va mantenuta nel modello, vale a dire che non è giustificabile il passaggio al modello più semplice.

b.3) Usando il secondo modello stimato, prevedere il costo di un'auto di 2 anni, con 120 cavalli potenza e con 90000 chilometri percorsi (attenzione alla scala di misura delle variabili). Come cambia il costo per un'auto con le stesse caratteristiche ma con 7 anni di età? Il risultato è ragionevole?

Auto di 2 anni con le caratteristiche indicate:

$$\widehat{\texttt{prezzo}} = 136.4331 - 0.2215 \cdot 90 + 6.2281 = 122.7262 \text{ centinaia di euro}$$

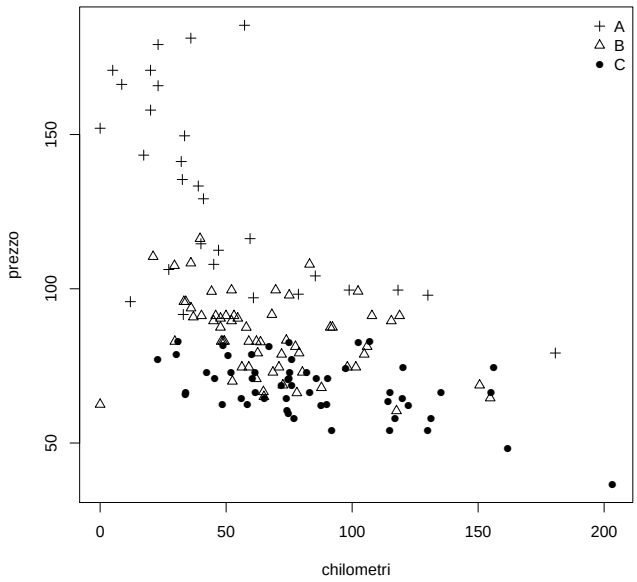
Auto di 7 anni con le caratteristiche indicate:

$$\widehat{\texttt{prezzo}} = 136.4331 - 0.2215 \cdot 90 - 53.3891 + 6.2281 = 69.3371 \text{ centinaia di euro}$$

È ragionevole che il prezzo dell'auto diminuisca al crescere dell'età del veicolo.

c) Il seguente grafico di dispersione tra `chilometri` e `prezzo` distingue le osservazioni riferite alle auto delle tre classi di età (si veda la pagina successiva)

c.1) Sulla base del grafico, si può ipotizzare che un modello lineare che spieghi il prezzo dell'auto in funzione dei chilometri, degli anni e della loro interazione abbia senso? Se si, perchè?



Si, ha senso un modello che includa le due variabili ed anche la loro interazione. Osservando il grafico si nota che che i punti associati ai tre livelli della variabile `anni` seguono dispersioni diverse: livelli più alti di prezzo si hanno tendenzialmente per il gruppo A, livelli medi per il gruppo B e livelli bassi per il gruppo C. In tutte e tre i casi al crescere dei chilometri diminuisce il prezzo del veicolo, con una precisazione: supponendo di stimare 3 rette di regressione che leghino il prezzo ai chilometri, una per ciascun livello di `anni`, si nota che le rette si intersecherebbero, vale a dire che la loro inclinazione, seppur decrescente data la relazione con `chilometri`, sarebbe diversa per ciascun gruppo. Ci si attende quindi che il termine di interazione sia significativo.

Informazioni utili

Quantili di una $N(0,1)$

$$z_{0.01} = -2.33 \;\; z_{0.025} = -1.96 \;\; z_{0.05} = -1.64 \;\; z_{0.95} = 1.64 \;\; z_{0.975} = 1.96 \;\; z_{0.99} = 2.33$$

Quantili di una F

$$F_{0.025;2,140} = 0.0253 \;\; F_{0.025;140,2} = 0.264 \;\; F_{0.975;2,140} = 3.788 \;\; F_{0.975;140,2} = 39.491$$

$$F_{0.05;2,140} = 0.051 \;\; F_{0.05;140,2} = 0.327 \;\; F_{0.95;2,140} = 3.061 \;\; F_{0.95;140,2} = 19.489$$