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

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) $\frac{\text{devianza residua}}{\text{devianza totale}}$

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 $\alpha = 0.05$

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

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

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

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
```

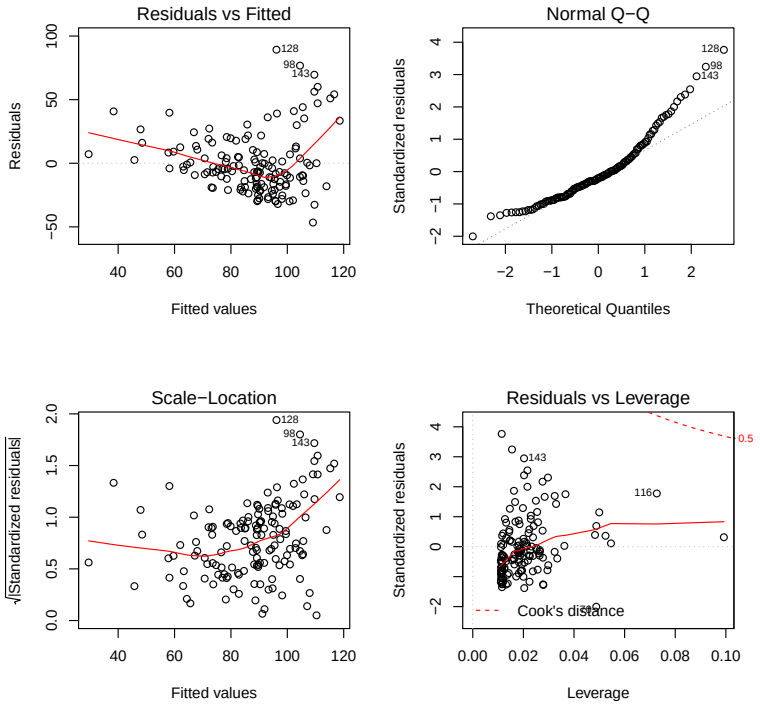
a.1) Scrivere l'espressione del modello stimato. Precisare come viene gestita la variabile qualitativa `cavalli` e quale livello (vale a dire quale potenza del motore) viene considerato di base.

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.

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?

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

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?

b) L'estensione del modello con l'inclusione della variabile `anni` risulta

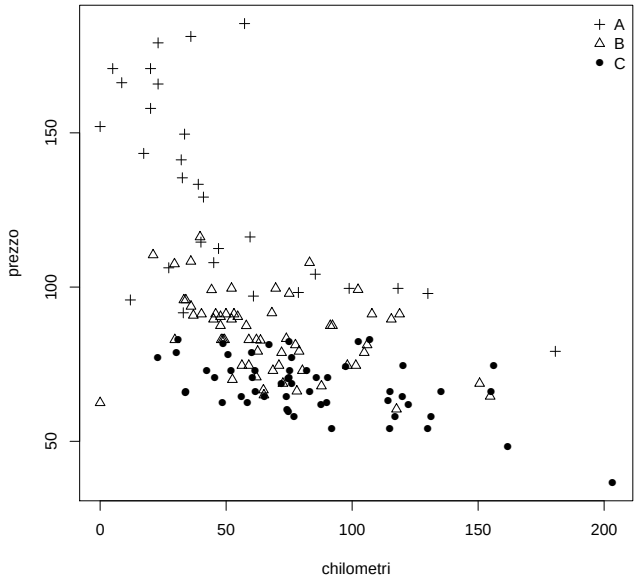
```
Call:
lm(formula = prezzo ~ chilometri + anni + cavalli, data = dati)

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
```

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

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.

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?

c) Il seguente grafico di dispersione tra `chilometri` e `prezzo` distingue le osservazioni riferite alle auto delle tre classi di età



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?

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

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

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