Università degli Studi di Padova

Corso di Laurea Magistrale in *Informatica*

a.a. 2016/2017

Data Mining

Docente: Annamaria Guolo

Prova scritta del 16 giugno 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) Se nel modello di regressione lineare semplice $Y = \beta_0 + \beta_1 X + \varepsilon$ stimato ai minimi quadrati si ottiene $\hat{\beta}_1 = 1$ allora

(a) $\rho_{XY} > 0$ (b) $\rho_{XY} < 0$ (c) $\rho_{XY} = 0$ (d) ρ_{XY} non determinabile

2) Sia dato il modello di regressione lineare semplice $Y = \beta_0 + \beta_1 X + \varepsilon$. Il test di verifica d'ipotesi $H_0 : \beta_1 \geq 0$ contro $H_1 : \beta_1 < 0$ basato su un campione di dimensione 30 conduce ad un valore della opportuna statistica test pari a 1.59. Allora il livello di significatività osservato (p-value) si calcola come

(a) $P(t_{28} < 1.59)$ (b) $P(t_{28} > 1.59)$ (c) $2\min\{P(t_{28} < 1.59), P(t_{28} > 1.59)\}$
(d) non calcolabile

3) I residui in un modello di regressione lineare che presenta un buon adattamento

(a) non hanno andamenti deterministici (b) hanno media pari alla media delle esplicative
(c) hanno varianza che cresce con il valore della risposta (d) hanno una distribuzione asimmetrica

4) Per stimare il modello di regressione lineare $Y = \beta_0 + \beta_1 X + \varepsilon$ sulla base di n osservazioni (y_i, x_i), il criterio dei minimi quadrati minimizza

(a) $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$ (b) $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$
(c) $\sum_{i=1}^{n}(y_i + \beta_0 + \beta_1 x_i)$ (d) $\sum_{i=1}^{n}(y_i + \beta_0 + \beta_1 x_i)^2$

5) L'errore di primo tipo è

(a) rifiutare H_0 quando H_0 è vera (b) rifiutare H_0 quando H_0 è falsa
(c) rifiutare H_1 quando H_0 è vera (d) rifiutare H_1 quando H_0 è falsa

Esercizio

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

Si considerino i dati riferiti a 111 rilevazioni giornaliere della qualità dell'aria a New York, nel periodo Maggio-Settembre 1973

- `Ozono`: logaritmo della concentrazione di ozono in parti per bilione
- `Radiazione` misurazione della radiazione solare in Langleys
- `Vento`: misurazione della velocità del vento in miglia orarie
- `Temperatura`: temperatura massima in gradi Farenheit
- `Mese`: mese di rilevazione (Maggio=5, Giugno=6, Luglio=7, Agosto=8, Settembre=9)

a) Viene stimato un modello di regressione lineare per spiegare la concentrazione di Ozono in funzione della radiazione solare e del mese di rilevazione. Di seguito l'output fornito da `R`

```
Call:
lm(formula = Ozono ~ Radiazione + Mese, data = dati)

Residuals:
     Min       1Q   Median       3Q      Max
-2.12182 -0.45821 -0.05062  0.48454  1.75260

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.0894291  0.1923459  10.863  < 2e-16 ***
Radiazione  0.0040489  0.0007287   5.556 2.10e-07 ***
Mese6       0.4013504  0.2666987   1.505   0.1354
Mese7       0.9181340  0.1947599   4.714 7.49e-06 ***
Mese8       1.0305130  0.1992029   5.173 1.11e-06 ***
Mese9       0.4483159  0.1885550   2.378   0.0192 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6823 on 105 degrees of freedom
Multiple R-squared:  0.4073, Adjusted R-squared:  0.379
F-statistic: 14.43 on 5 and 105 DF,  p-value: 9.538e-11
```

a.1) Di che natura sono le variabili esplicative considerate nel modello? Come viene gestita la variabile `Mese` da `R`?

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

a.3) Cosa rappresenta la quantità `Residual standard error` riportata nell'output? Come viene calcolata?

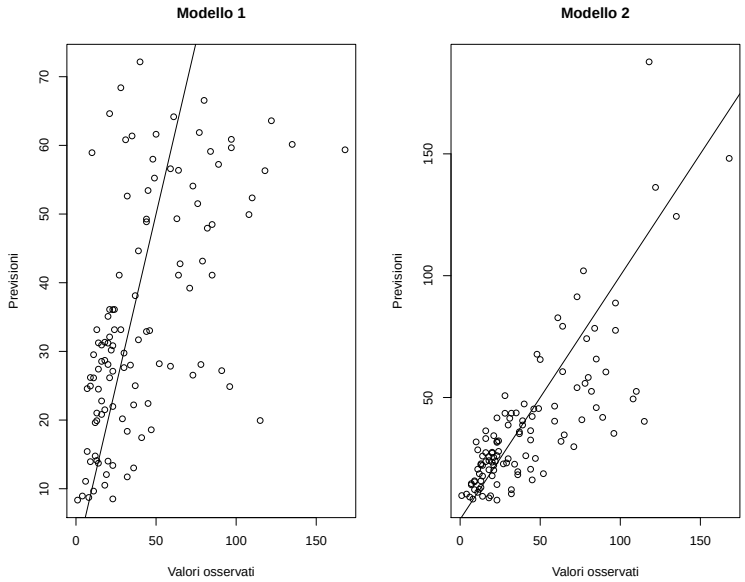
b) Si decide di estendere il modello con l'inclusione della variabile `Vento`. Il modello stimato è

```
Call:
lm(formula = Ozono ~ Radiazione + Mese + Vento + I(Vento^2),
    data = dati)

Residuals:
     Min       1Q   Median       3Q      Max
-2.27757 -0.31881 -0.02675  0.35808  1.11581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.3649500  0.4066855  10.733  < 2e-16 ***
Radiazione   0.0035803  0.0006104   5.865 5.47e-08 ***
Mese6        0.4382680  0.2224700   1.970 0.051522 .
Mese7        0.6168069  0.1687625   3.655 0.000407 ***
Mese8        0.7349607  0.1717170   4.280 4.20e-05 ***
Mese9        0.2966963  0.1588203   1.868 0.064587 .
Vento       -0.3142175  0.0665502  -4.722 7.41e-06 ***
I(Vento^2)   0.0099042  0.0029921   3.310 0.001286 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5679 on 103 degrees of freedom
Multiple R-squared:  0.5972, Adjusted R-squared:  0.5698
F-statistic: 21.82 on 7 and 103 DF,  p-value: < 2.2e-16
```

b.1) Sulla base dell'output è stato vantaggioso l'inserimento della variabile `Vento`?

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.

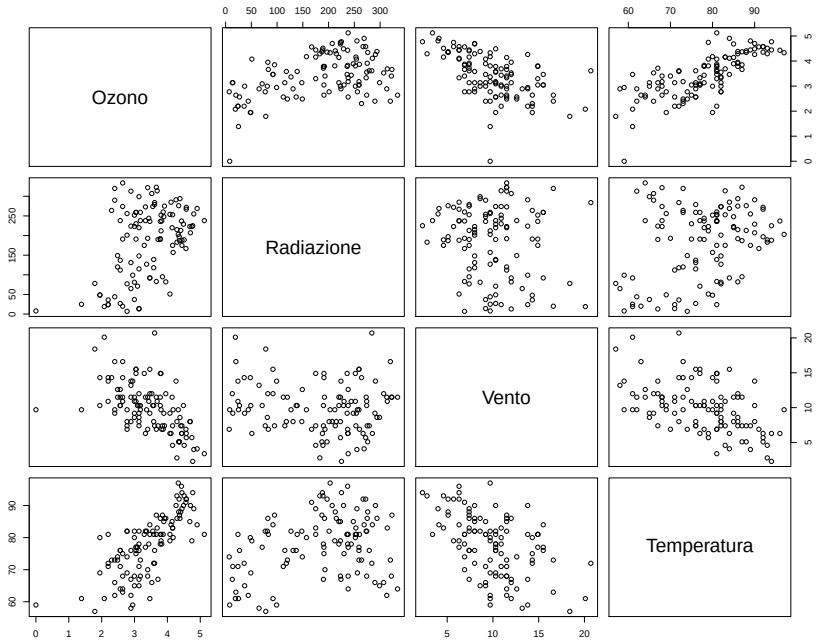
b.3) Il seguente grafico riporta la previsione di concentrazione di ozono (sulla scala originaria) sul training set usando i due modelli stimati, vale a dire il modello che non considera la variabile `Vento` (modello 1) ed il modello che considera la variabile `Vento` (modello 2). La retta tratteggiata è la bisettrice del primo-terzo quadrante. Come si interpreta il grafico? È utile per confrontare le performance dei due modelli? Cosa suggerisce?



b.4) Si valuti tramite verifica d'ipotesi al livello di significatività 0.01 se il parametro associato alla variabile `Vento` (che entra linearmente nel modello) si possa considerare pari a -0.2 oppure no, spiegando le eventuali assunzioni fatte.

b.5) Usando il secondo modello stimato, prevedere la concentrazione di ozono (sulla scala originale) nel mese di Luglio, nel caso di radiazione solare pari a 185 Langleys e velocità del vento pari a 10 miglia orarie. Come cambia la concentrazione di ozono se radiazione solare e velocità del vento rimangono costanti ma la rilevazione viene fatta in Settembre? Il risultato è ragionevole?

c) Il seguente grafico riporta i diagrammi di dispersione delle variabili continue del modello, prese a due a due



c.1) I grafici suggeriscono possibili estensioni del modello stimato `modello` che potrebbe migliorarne l'adattamento? Quali?

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

$$F_{0.025;2,103} = 0.0253 \;\; F_{0.025;103,2} = 0.261 \;\; F_{0.975;2,103} = 3.824 \;\; F_{0.975;103,2} = 39.488$$

$$F_{0.05;2,103} = 0.051 \;\; F_{0.05;103,2} = 0.324 \;\; F_{0.95;2,103} = 3.085 \;\; F_{0.95;103,2} = 19.486$$