Università degli Studi di Padova

Corso di Laurea Magistrale in *Informatica*

a.a. 2016/2017

Data Mining

Docente: Annamaria Guolo

Prova scritta del 30 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 $\hat{\beta}_0 = 0$ allora sicuramente

(a) $R^2 = 0$ (b) $R^2 = -1$ (c) $R^2 = 1$ (d) nessuna delle precedenti

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 200 conduce ad un valore della opportuna statistica test pari a -2.13. Allora il livello di significatività osservato (p-value) vale

(a) $2\min\{P(t_{198} < -2.13), P(t_{198} > -2.13)\}$
(b) $2\min\{P(t_{198} < 2.13), P(t_{198} > -2.13)\}$
(c) $P(t_{198} < -2.13)$
(d) $P(t_{198} > -2.13)$

3) I residui standardizzati

(a) hanno media nulla
(b) hanno varianza pari a quella dei residui non standardizzati
(c) sono sempre positivi
(d) sono sempre negativi

4) La devianza totale è pari a

(a) $\frac{devianza\ spiegata}{devianza\ residua}$
(b) $\frac{devianza\ residua}{devianza\ spiegata}$
(c) $devianza\ spiegata + devianza\ residua$
(d) $devianza\ spiegata - devianza\ residua$

5) Il coefficiente di correlazione ρ_{XY} è

(a) una probabilità
(b) una misura del legame lineare tra X e Y
(c) un intervallo di confidenza
(d) una misura del livello di significatività osservato

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 seguenti dati riferiti ad un'indagine svolta dal Canadian Survey of Labour and Income Dynamics nel 1994

- `wages`: guadagno orario (su scala logaritmica in base naturale)
- `education`: anni di istruzione
- `age`: età in anni
- `sex`: genere Maschio/Femmina
- `language`: lingua parlata English/French/Other

I dati si riferiscono a 300 osservazioni.

a) Viene stimato un modello di regressione lineare per spiegare il guadagno orario in funzione degli anni di istruzione e del genere. Di seguito l'output fornito da `R`

```
Call:
lm(formula = wages ~ sex + education + I(education^2), data = slid)

Residuals:
     Min       1Q   Median       3Q      Max
-1.75578 -0.31897  0.01735  0.31488  1.19888

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.627102   0.310179   8.470 1.17e-15 ***
sexMale         0.133321   0.055630   2.397  0.01717 *
education      -0.070156   0.045183  -1.553  0.12156
I(education^2)  0.004444   0.001648   2.696  0.00742 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4778 on 296 degrees of freedom
Multiple R-squared:  0.1374, Adjusted R-squared:  0.1286
F-statistic: 15.71 on 3 and 296 DF,  p-value: 1.655e-09
```

a.1) Di che natura sono le variabili esplicative considerate nel modello? Come si interpreta la stima del coefficiente associato a `sex`?

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 `t value` riportata nell'output? Come viene calcolata?

a.4) Costruire un intervallo di confidenza di livello 0.95 per il parametro associato a `education`, spiegando le eventuali assunzioni fatte.

a.5) Prevedere il guadagno orario (sulla scala originaria) per una donna con 15 anni di istruzione. Come cambia il risultato nel caso di un uomo?

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

```
Call:
lm(formula = wages ~ sex + language * education + I(education^2),
    data = slid)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63393 -0.31887  0.02166  0.30815  1.22297

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.468853   0.348133   7.092    1e-11 ***
sexMale                   0.131873   0.056209   2.346   0.0196 *
languageFrench            0.449934   0.532223   0.845   0.3986
languageOther             0.281458   0.331745   0.848   0.3969
education                -0.052333   0.048843  -1.071   0.2849
I(education^2)            0.003976   0.001726   2.303   0.0220 *
languageFrench:education -0.023459   0.041178  -0.570   0.5693
languageOther:education  -0.023564   0.026030  -0.905   0.3661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4792 on 292 degrees of freedom
Multiple R-squared:  0.1441, Adjusted R-squared:  0.1236
F-statistic: 7.022 on 7 and 292 DF,  p-value: 9.503e-08
```

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

b.2) Sulla base dell'output come si dovrebbe procedere nell'analisi?

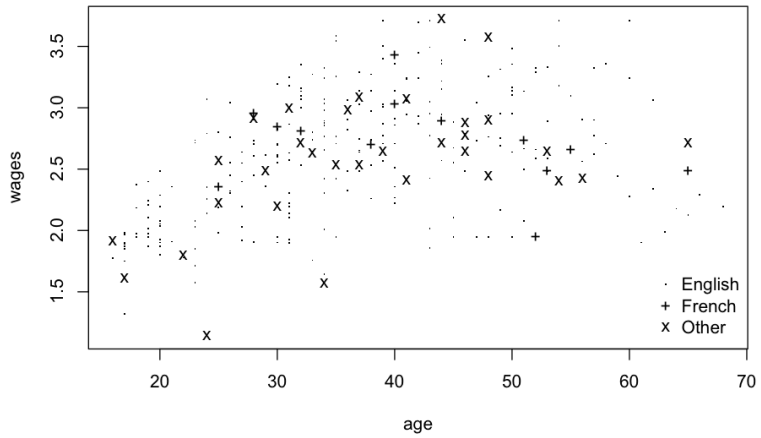
c) Si confrontano i due modelli fin qui stimati (`modello 1` senza considerare la variabile `language` e `modello 2` considerando la variabile `language`) per valutare se sia possibile scegliere il modello più semplice (`modello 1`) come modello per la previsione. Di seguito l'output fornito da `R`

```
Analysis of Variance Table

Model 1: wages ~ sex + education + I(education^2)
Model 2: wages ~ sex + language * education + I(education^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    296 67.586
2    292 67.060  4   0.52655 0.5732 0.6823
```

A cosa si riferisce l'analisi svolta e come si interpreta il risultato? È giustificabile la scelta del modello più semplice? Perchè?

d) Il seguente grafico riporta la distribuzione di `wages` in funzione della variabile `age`, distinguendo in base alla variabile `language`



Quali suggerimenti si possono cogliere dal grafico al fine di migliorare il modello fin qui stimato?

Informazioni utili

Quantili di una $N(0,1)$: $z_{0.01} = -2.33$ $z_{0.025} = -1.96$ $z_{0.05} = -1.64$ $z_{0.95} = 1.64$ $z_{0.975} = 1.96$ $z_{0.99} = 2.33$