Università degli Studi di Padova

Corso di Laurea Magistrale in *Informatica*

a.a. 2018/2019

Data Mining

Docenti: Antonio Canale, Manuela Cattelan, Davide Risso

Esempio di prova parziale[1]

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 stimato ai minimi quadrati $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ si ha $\hat{\beta}_1 = 0$, allora

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

2) Nell'ambito della verifica d'ipotesi, il livello di significatività osservato è

(a) compreso tra 0 e $+\infty$
(b) compreso tra -1 e 1
(c) l'errore di secondo tipo
(d) nessuna delle precedenti

3) In un modello di regressione lineare, la precisione delle stime ottenute con il criterio dei minimi quadrati si misura tramite

(a) lo standard error (b) l'indice di correlazione (c) la distorsione
(d) la somma dei residui

4) All'aumentare della numerosità campionaria n l'ampiezza degli intervalli di confidenza associati ai parametri del modello di regressione lineare

(a) diminuisce (b) aumenta (c) diminuisce fino ad un certo n e poi aumenta
(d) non varia

5) Gli errori ε nel modello di regressione lineare $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \ldots, n$, si assumono

(a) a media unitaria
(b) a varianza unitaria
(c) incorrelati con le esplicative
(d) incorrelati con Y

[1] Materiale predisposto da Annamaria Guolo

Esercizio.
Si considerino le informazioni su 397 docenti di un college statunitense nel periodo accademico 2008-2009 e relative agli anni di esperienza di insegnamento, al tipo di disciplina insegnata (A= teorica, B= applicata) ed allo stipendio per 9 mesi in dollari.

a) Si stima un modello di regressione lineare per spiegare lo stipendio in funzione degli anni di servizio e del tipo di disciplina insegnata. Di seguito l'output fornito da `R`

```
Call:
lm(formula = stipendio ~ anni.servizio + disciplina, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-77537 -19699  -5135  15631 106625

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   91335.8     3005.4  30.391  < 2e-16 ***
anni.servizio   862.8      109.2   7.904 2.73e-14 ***
disciplinaB   13184.0     2846.8   4.631 4.95e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27870 on 394 degrees of freedom
Multiple R-squared:  0.1579,
  Adjusted R-squared:  0.1536
F-statistic: 36.94 on 2 and 394 DF,  p-value: 1.983e-15
```

a.1) Scrivere l'espressione del modello stimato. Precisare come viene gestita la variabile qualitativa `disciplina` e quale livello viene considerato di base.

a.2) Commentare l'output del modello evidenziando i) significatività dei coefficienti associati alle stime, ii) possibilità di semplificazione del modello, iii) adattamento del modello tramite R^2.

a.3) Proporre un intervallo di confidenza di livello 0.95 per il parametro associato alla variabile `anni.servizio` spiegando le assunzioni fatte.

b) L'estensione del modello con l'inclusione dell'interazione tra `anni.servizio` e `disciplina` porta al seguente output

```
Call:
lm(formula = stipendio ~ anni.servizio * disciplina, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-86326 -19779  -4999  16091 102274

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                98038.0     3626.9   27.03  < 2e-16 ***
anni.servizio                526.8      150.1    3.51 0.000499 ***
disciplinaB                  857.4     4750.7    0.18 0.856873
anni.servizio:disciplinaB    695.2      215.9    3.22 0.001388 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27540 on 393 degrees of freedom
Multiple R-squared:  0.1795,
  Adjusted R-squared:  0.1733
F-statistic: 28.67 on 3 and 393 DF,  p-value: < 2.2e-16
```

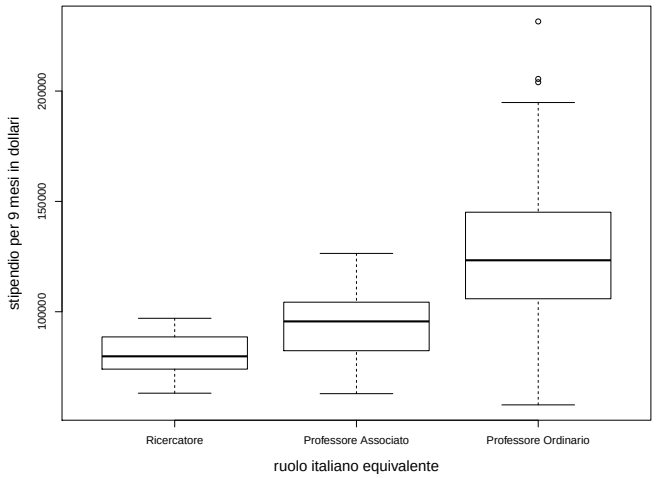
b.1) Ha senso mantenere l'interazione nel modello? Il modello è semplificabile? Perchè?

b.2) Confrontare i due modelli fin qui stimati in base a R^2 e commentare.

b.3) 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.4) Prevedere lo stipendio per un docente che insegna una disciplina teorica ed ha 20 anni di servizio. Per un docente con la stessa età di servizio, prevedere lo stipendio se insegnasse una disciplina applicata.

c) Il seguente grafico riporta la distribuzione dello stipendio distinta per ruolo ricoperto dal docente



c.1) Se si inserisse la variabile `ruolo` come esplicativa (senza interazioni) nel modello di regressione lineare che vede lo stipendio come risposta, quale sarebbe il livello base? Quante e quali variabili dummy sarebbero costruite?

c.2) Commentare il grafico. Cosa ci si potrebbe attendere in termini di significatività del parametro/dei parametri associato/associati alla variabile `ruolo` nel caso in cui la variabile `ruolo` venisse inserita nel modello?

Informazioni utili

Quantili di una Normale standard

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

Quantili di una F

$$F_{0.025;1,393} = 0.00098 \;\; F_{0.025;393,1} = 0.1975 \;\; F_{0.975;1,393} = 5.063 \;\; F_{0.975;393,1} = 1016.962$$

$$F_{0.05;1,393} = 0.0039 \;\; F_{0.025;393,1} = 0.2587 \;\; F_{0.95;1,393} = 3.865 \;\; F_{0.95;393,1} = 253.9898$$