Università degli Studi di Padova

Corso di Laurea Magistrale in *Informatica*

a.a. 2016/2017

Data Mining

Docente: Annamaria Guolo

Esempio di prova parziale: 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) 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

La risposta corretta è la risposta (a).

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

La risposta corretta è la risposta (d).

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

La risposta corretta è la risposta (a).

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

La risposta corretta è la risposta (a).

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

La risposta corretta è la risposta (c).

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.
Il modello stimato è

$$\widehat{\texttt{Stipendio}} = 91335.8 + 862.8 \times \texttt{anni.servizio} + 13184 \times I_{disciplina=B}$$

dove $I_{disciplina=B}$ è la variabile indicatrice che assume valore 1 se `disciplina=B` e valore 0 se `disciplina=B`. Questo perchè la variabile `disciplina` è una variabile qualitativa con due livelli, A e B, dei quali il livello A 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 13184 indica l'aumento di stipendio per un soggetto che insegna la disciplina B rispetto ad un soggetto che insegna la disciplina A.

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.
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 d'ipotesi per l'uguaglianza a zero dei coefficienti è molto basso (praticamente nullo). Di conseguenza il modello non è semplificabile. 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) Proporre un intervallo di confidenza di livello 0.95 per il parametro associato alla variabile `anni.servizio` spiegando le assunzioni fatte.
Indicando con β_1 il coefficiente associato a `anni.servizio`, 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

$$\hat{\beta}_1 - t_{0.975,n-3} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{0.975,n-3} SE(\hat{\beta}_1)$$

Poichè $n - 3 = 394$ 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 $N(0, 1)$

$$\hat{\beta}_1 - z_{0.975} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + z_{0.975,n-3} SE(\hat{\beta}_1)$$

vale a dire

$$862.8 - 1.96 \times 109.2 \leq \beta_1 \leq 862.8 + 1.96 \times 109.2$$

da cui si ricava

$$648.768 \leq \beta_1 \leq 1076.832$$

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?

Ha senso mantenere il termine di interazione nel modello in quanto il livello di significatività osservato associato alla variabili di interazione è molto piccolo e quindi il test per la significatività del coefficiente corrispondente viene rifiutato. Sebbene la variabile `disciplina` non sia significativa, il modello non è semplificabile in base al principio di gerarchia, essendo la variabile presente in interazione ed essendo l'interazione significativa.

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

Il passaggio al secondo modello comporta un aumento di R^2 da 0.1579 a 0.1795. Considerato che il nuovo termine inserito nel modello è significativo, l'aumento di R^2 è effettivamente indicazione di un miglioramento dell'adattamento.

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.

Consideriamo la verifica d'ipotesi $H_0 : \beta_3 = 0$ contro $H_1 : \beta_3 \neq 0$, dove β_3 indica il coefficiente associato all'interazione tra `anni.servizio` e `disciplina`. 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 $q = 1$, $n - p - 1 = 397 - 3 - 1 = 393$, 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 $RSS_0 = RSE^2(n - p - 1) = 27870^2 \times 394 = 306034338600$ e che $RSS = 27540^2 \times 393 = 298071478800$ il valore osservato sui dati di F è

$$\frac{306034338600 - 298071478800}{298071478800} \frac{393}{1} = 10.49884$$

Confrontiamo questo valore con il quantile 0.95 di una $F_{1,393}$, che vale 3.865. Poichè il valore osservato è maggiore, si rifiuta l'ipotesi nulla al livello 0.05: la variabile va mantenuta nel modello, non è giustificabile il passaggio al modello più semplice senza interazione. In questo caso, poichè i modelli differiscono per una esplicativa, il valore osservato di F coincide con il valore osservato della statistica t associata alla significatività del coefficiente del termine di interazione, vale a dire al quadrato di 3.22 (a meno di arrotondamenti).

Poichè la richiesta indica di considerare il livello di significatività 0.05, si prevede che la verifica d'ipotesi venga condotta nel modo sopra descritto piuttosto che tramite il calcolo del livello di significatività osservato. Se lo si volesse calcolare, comunque, il risultato sarebbe

$$P(F_{1,393} \geq 10.49884) = 1 - P(F_{1,393} < 10.49884) = 0.0013,$$

valore piccolo che conferma il rifiuto dell'ipotesi nulla.

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.

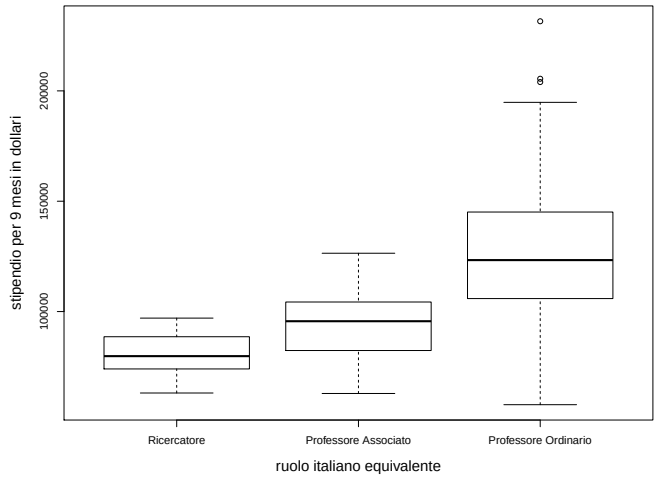
Docente disciplina B:

$$\widehat{\texttt{Stipendio}} = 98038 + 526.8 \times 20 + 857.4 + 695.2 \times 20 = 123335.4$$

Docente disciplina A:

$$\widehat{\texttt{Stipendio}} = 98038 + 526.8 \times 20 = 108574$$

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?

Il livello base sarebbe `Professore Associato`, dato che `R` segue di default l'ordine alfabetico. Poichè la variabile presenta 3 livelli, si costruiscono due variabili dummy, o variabili indicatrici: la prima assume valore 0 se il soggetto non è Ricercatore e 1 se lo è, mentre la seconda assume valore 0 se il soggetto non è Professore Ordinario e 1 se lo è.

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?

Il grafico indica che la distribuzione della variabile `stipendio` varia a seconda dei tre livelli di `ruolo`. In particolare, la mediana (e tendenzialmente anche la media) dello stipendio per un ricercatore è minore di quella di un professore associato, a sua volta minore di quella di un professore ordinario. La variabilità dello stipendio di un professore ordinario è la maggiore.
Ci si attende che, una vola inserita la variabile nel modello, vi sia una significatività dei due coefficienti associati ai livelli corrispondenti a ricercatore e professore ordinario. Inoltre, sulla base delle mediane e medie dei boxplot, ci si attende che il coefficiente associato a ricercatore sia di segno negativo, indicando quindi una diminuzione di stipendio medio rispetto al professore associato (il livello base), mentre ci si attende che il coefficiente associato a professore ordinario sia di segno positivo, indicando quindi un aumento medio di stipendio rispetto al professore associato.
Infine, il boxplot riferito allo stipendio per professore ordinario indica la presenza di tre valori estremi, vale a dire di tre soggetti con stipendio più elevato rispetto agli altri. È possibile che a questi soggetti siano associati tre residui del modello classificabili come valori anomali: la verifica andrebbe fatta sulla base dell'analisi dei residui, in particolare dei valori leva e della distanza di Cook.

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

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

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