Università degli Studi di Padova

Corso di Laurea Magistrale in *Informatica*

a.a. 2016/2017

Data Mining

Docente: Annamaria Guolo

Prova scritta del 21 settembre 2017

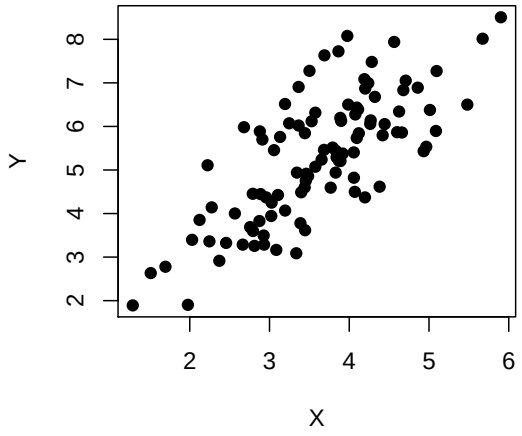
ISTRUZIONI: La durata della prova è di 1 ora. La prova va svolta su questi fogli. Eventuali fogli di brutta copia possono essere richiesti, ma non verranno corretti. Non scrivere in matita. In caso di errore, barrare la parte errata, non utilizzare un correttore (bianchetto).

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) Il seguente grafico riporta la distribuzione delle coppie di osservazioni dalle variabili Y e X. Un valore ragionevole per il coefficiente di correlazione lineare tra X e Y è



(a) 0 (b) -1 (c) 1 (d) 0.7

2) Il livello di significatività osservato (detto anche p-value) è

(a) una probabilità
(b) l'errore di secondo tipo
(c) una variabile casuale
(d) una misura del legame lineare tra X e Y

3) All'aumentare della numerosità campionaria n la varianza (o, se si preferisce, la sua radice quadrata, vale a dire lo standard error) associata agli stimatori dei parametri di un modello di regressione lineare

(a) diminuisce (b) aumenta (c) resta invariata
(d) diminuisce fino ad un certo n e poi aumenta

4) Il residuo in un modello in cui y_i indica l'osservazione i-esima della variabile risposta e $\hat{y}_i$ la previsione basata sul modello è pari a

(a) $(y_i - \hat{y}_i)^2$ (b) $y_i - \hat{y}_i$ (c) $|y_i - \hat{y}_i|$ (d) $\sqrt{y_i - \hat{y}_i}$

5) Al crescere della devianza residua di un modello

(a) R^2 diminuisce (b) R^2 aumenta (c) la devianza totale diminuisce
(d) la devianza spiegata resta invariata

6) Sia dato il modello di regressione lineare semplice $Y = \beta_0 + \beta_1 X + \varepsilon$. Sia 1.325 la stima ai minimi quadrati di β_1 e 0.23 il suo standard error. La statistica test per la verifica d'ipotesi per $H_0 : \beta_1 \neq 1$ contro l'alternativa bilaterale assume valore

(a) 1.41 (b) 5.76 (c) 0.708 (d) 4.35

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 alle caratteristiche ed alle spese sostenute da 1002 possessori di carta di credito.

- `expenditure`: spesa media mensile in dollari (su scala logaritmica in base naturale)
- `age`: età in anni
- `selfemp`: lavoratore autonomo (yes/no)
- `income`: guadagno annuale (in 10,000 dollari)

a) Viene stimato un modello di regressione lineare per spiegare la spesa sostenuta tramite carta di credito in funzione di `age`, `income`, `selfemp`. Di seguito l'output fornito da `R`

```
Call:
lm(formula = expenditure ~ age + income + selfemp, data = dati)

Residuals:
      Min        1Q    Median        3Q       Max
-6.092960 -0.568312  0.141575  0.795429  2.708812

Coefficients:
              Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  4.70384754  0.13274635 35.43485 < 2.22e-16 ***
age         -0.01337835  0.00397093 -3.36907 0.00078307 ***
income       0.19286403  0.02343480  8.22981 5.8177e-16 ***
selfempyes  -0.30013695  0.15898459 -1.88784 0.05933777 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.18068 on 998 degrees of freedom
Multiple R-squared:  0.0650112,
  Adjusted R-squared:  0.0622007
F-statistic: 23.1308 on 3 and 998 DF,  p-value: 1.76535e-14
```

a.1) Scrivere l'espressione del modello (generico) per il quale si è deciso di procedere alla stima e l'espressione del modello stimato.

a.2) Di che natura sono le variabili esplicative considerate nel modello? Come si interpreta la stima del coefficiente associato a `selfemp`?

a.3) 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.4) Cosa rappresenta la quantità `F statistic` riportata nell'output? Come viene calcolata?

a.5) Costruire un intervallo di confidenza di livello 0.9 per il parametro associato a `income`, spiegando le eventuali assunzioni fatte.

a.6) Prevedere la spesa media (sulla scala originaria) un lavoratore autonomo di 35 anni con un guadagno annuale di 40,000 dollari. Come cambia il risultato nel caso di un lavoratore non autonomo? Commentare.

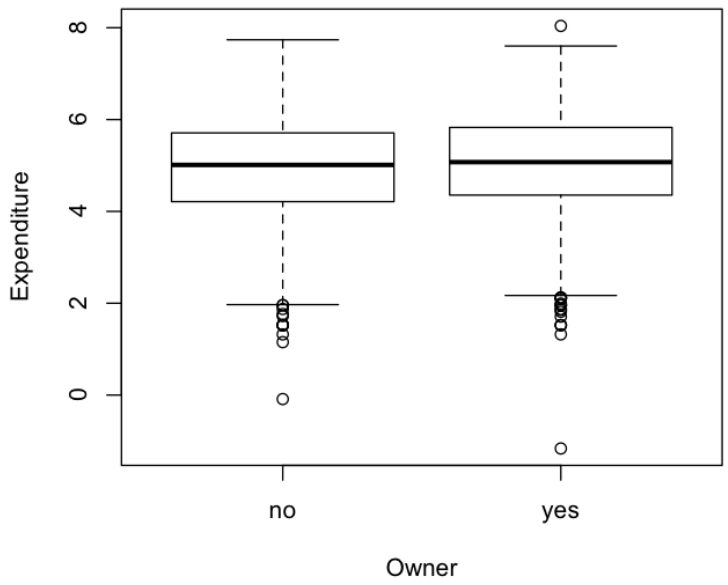
b) Si decide di estendere il modello con l'inclusione dell'interazione tra la variabile `income` e la variabile `selfemp`. Il modello stimato è

```
Call:
lm(formula = expenditure ~ age + income * selfemp, data = dati)

Residuals:
      Min        1Q    Median        3Q       Max
-6.093290 -0.568690  0.149711  0.807821  2.706777

Coefficients:
                    Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)        4.71050384  0.13337194 35.31855 < 2.22e-16 ***
age               -0.01316063  0.00399302 -3.29591  0.0010156 **
income             0.18881187  0.02462959  7.66606 4.1991e-14 ***
selfempyes        -0.48243013  0.37512796 -1.28604  0.1987272
income:selfempyes  0.04256613  0.07933178  0.53656  0.5916924
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.1811 on 997 degrees of freedom
Multiple R-squared:  0.0652811,
  Adjusted R-squared:  0.061531
F-statistic: 17.4077 on 4 and 997 DF,  p-value: 8.12974e-14
```

b.1) Sulla base dell'output è stato vantaggioso l'inserimento dell'interazione? Perchè?

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.

c) Si vuole valutare l'inserimento della variabile `owner` che considera la proprietà o meno dell'abitazione da parte del possessore della carta di credito. A tal fine si disegna il seguente grafico.



A cosa si riferisce il grafico e come si interpreta? Ci si aspetta una relazione tra `expenditure` e `owner` sulla base del grafico?

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

$$F_{0.025;1,997} = 0.00098 \;\; F_{0.025;997,1} = 0.1984 \;\; F_{0.975;1,997} = 5.039 \;\; F_{0.975;997,1} = 1017.747$$

$$F_{0.05;1,997} = 0.0039 \;\; F_{0.05;997,1} = 0.2597 \;\; F_{0.95;1,997} = 3.8508 \;\; F_{0.95;997,1} = 254.1864$$