# ESERCIZIO 1 -----------------------

#Introduzione

x = c(150:189) #crea un vettore x contenente i numeri 150,151,...,188,189

alpha = -90; beta = 0.8 #definizione di alpha e beta

ym = alpha + (beta\*x) #vettore contenente i valori alpha+(beta\*x)

set.seed(236704) #innesco generatore di numeri con numero di matricola

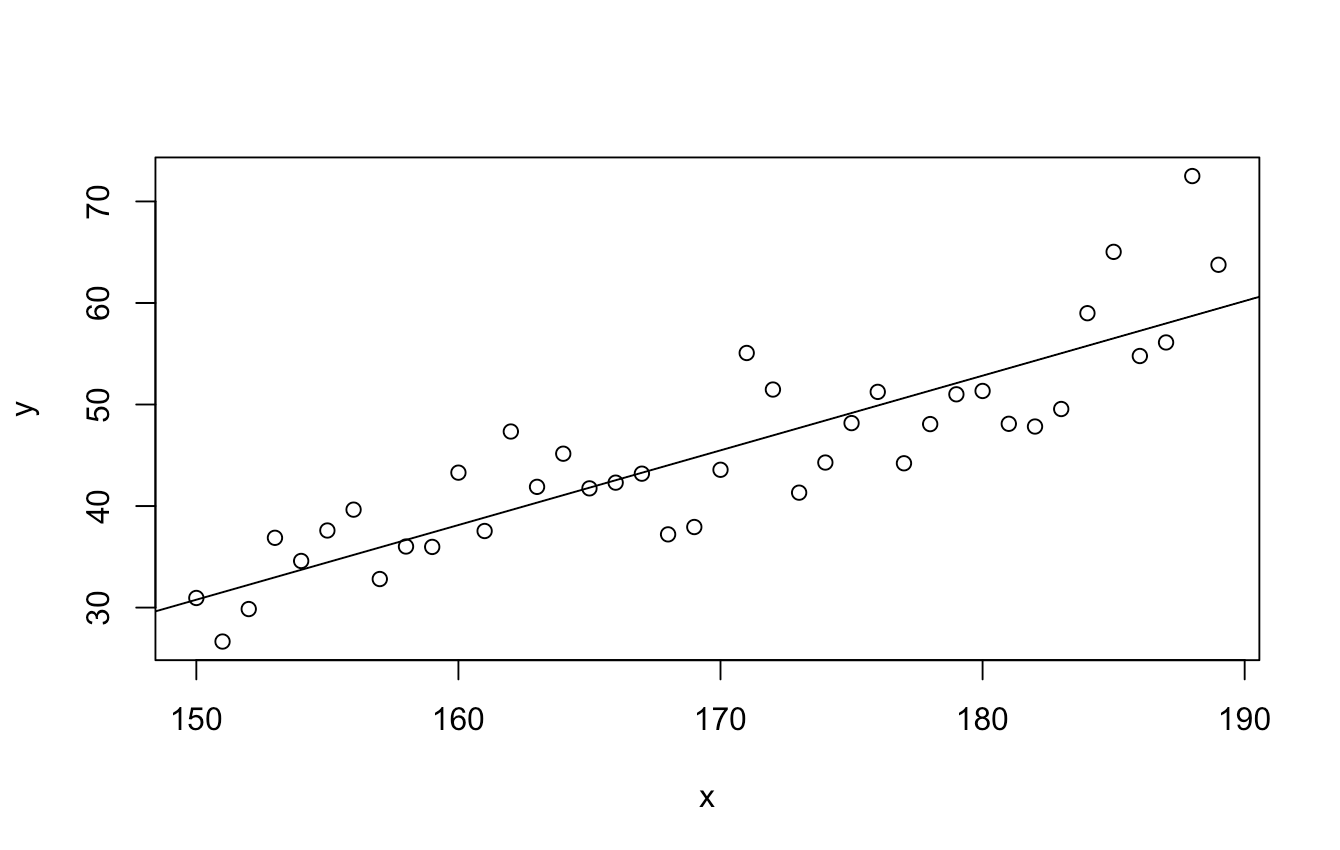
e = rnorm(40, mean = 0, sd = 5) #campionamento di 40 valori da N(0,5^2)

y = ym+e #vettore dei valori osservati

cor(x,y) #correlazione tra x e y con R

plot(x,y) #scatterplot

SOL:



> cor(x,y) #correlazione tra x e y con R

[1] 0.8720858

# Modello di regressione lineare y = alpha + beta(x)

modello = lm(y ~ x) #modello di regressione lineare

summary(modello) #sommario del modello di regressione lineare per trovare alpha

#(-106.16199) e beta (0.89385)

plot(x,y); abline(modello) #grafico con anche la retta di regressione

Ymedia = mean(y) #valore medio di Y

Yprevista = modello$fitted.values #valore previsto di Y per ogni X

TSS = sum((y-Ymedia)^2) #Total sum of squares

SSR = sum((Yprevista-Ymedia)^2) #Somma dei quadrati della regressione

SSE = sum((y-Yprevista)^2)#Sum of square error

s = sqrt(SSE/(length(y)-2)) #Residual standard error

s2 = s^2 #varianza condizionata

R2 = (TSS-SSE)/(TSS) #R^2

rbind(TSS,SSR,SSE,s,s2,R2) # Creo matrice riassuntiva

anova(modello) #Analizzo il modello attraverso l'utilizzo della varianza

seb = s/(sqrt(sum((x-mean(x))^2))) #SE di b

#Test di verifica di ipotesi su b

b = as.numeric(modello$coefficients[2])

testT = b/seb #test

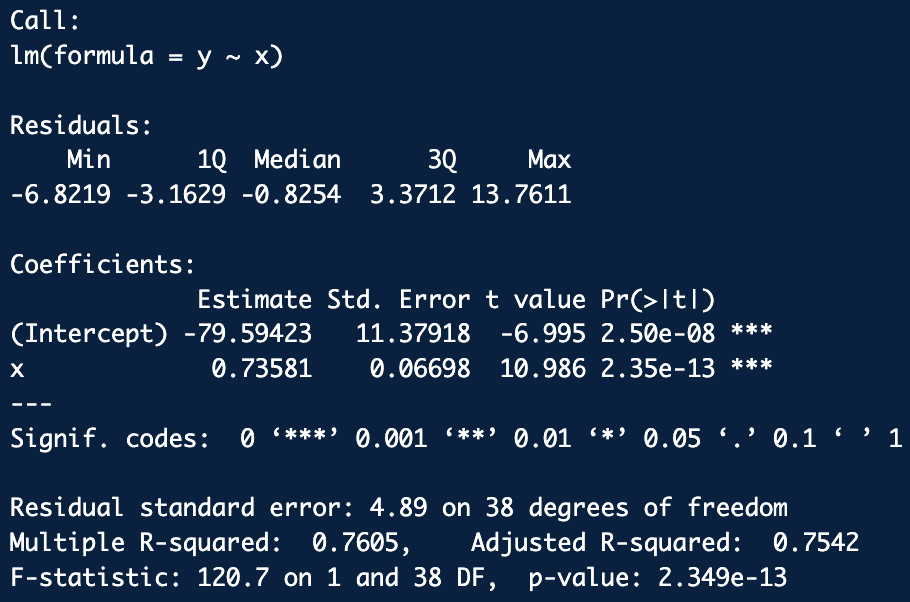
pvalue = 2\*pt(testT,length(y)-2,lower.tail=FALSE) #pvalue

quanT = qt(0.025,length(y)-2,lower.tail=FALSE) #quantile

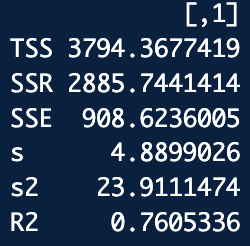
c(b-(seb\*quanT), " < b < ", b+(seb\*quanT)) #intervallo di confidenza, dove è contenuto 0.8

SOL:

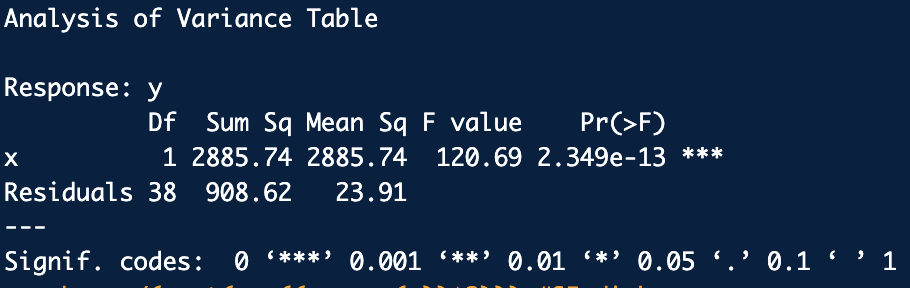
> summary(modello) #sommario del modello di regressione lineare per trovare alpha



> rbind(TSS,SSR,SSE,s,s2,R2) # Creo matrice riassuntiva



> anova(modello) #Analizzo il modello attraverso l'utilizzo della varianza



> c(b-(seb\*quanT), " < b < ", b+(seb\*quanT)) #intervallo di confidenza, dove è contenuto 0.8

[1] "0.600218102945188" " < b < " "0.871400612812053"

# Modello di regressione lineare ycen = a + b(xcen)

xcen = x - mean(x)

ycen = y - mean(y)

modello2 = lm(ycen ~ xcen) #modello di regressione lineare

summary(modello2)

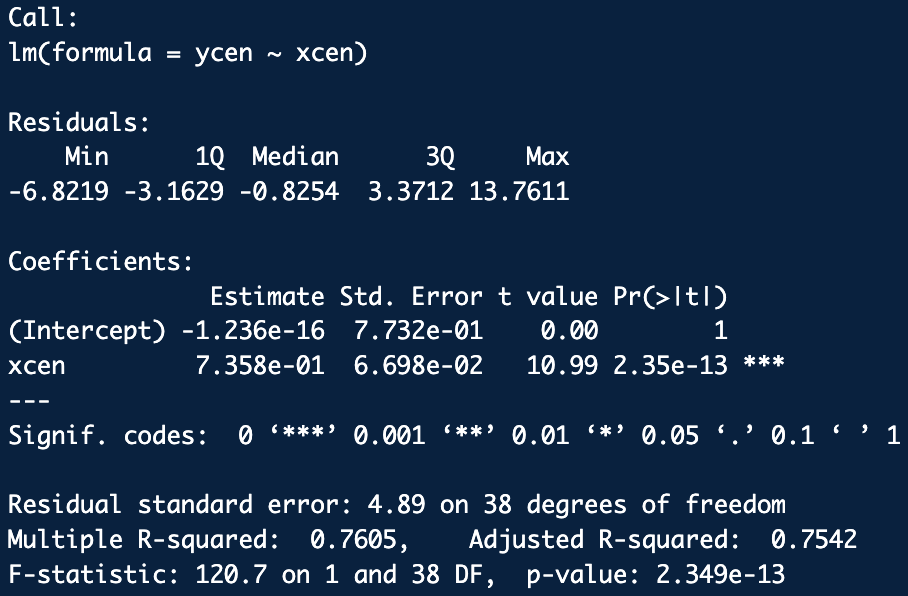
alpha1 = as.numeric(modello$coefficients[1]); alpha2 = as.numeric(modello2$coefficients[1]) #calcolo di a dei due modelli

beta1 = as.numeric(modello$coefficients[2]); beta2 = as.numeric(modello2$coefficients[2]) #calcolo di b dei due modelli

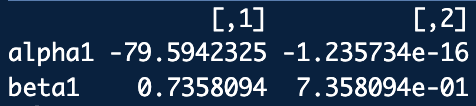
cbind(rbind(alpha1, beta1),rbind(alpha2, beta2)) #mostro i risultati a video

SOL:

> summary(modello2)



> cbind(rbind(alpha1, beta1),rbind(alpha2, beta2)) #mostro i risultati a video



# ESERCIZIO 2 ---------------------------

# Introduzione

df2 = read.csv("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/2004\ statewide\ crime.txt", header = TRUE)

# Modello di regressione multipla

#y = tasso di omicidi; x1 = tasso di povertà; x2 = diplomati all'high school

y = df2$Murder; x1 = df2$Poverty; x2 = df2$HighSch

pairs(cbind(y,x1,x2)) #mostro su dei grafici le tre variabili

lm2 = lm(y ~ x1 + x2) #modello di regressione lineare multivariato

summary(lm2)

#Partial regression plot controllando per x2

lmy = lm(y ~ x2)

lmx1 = lm(x1 ~ x2)

par(mfrow = c(1,2))

colors = rep("black", length(y)); colors[51] = "red"

plot(lmx1$res, lmy$res, xlab = "Redisui del modello 'tasso di povertà' regredito su

'percentuale di diplomati'",

ylab = "Residui del modello 'tasso di omicidi' regredito su 'percentuale di diplomati'",

main = "Partial regression plot controllando

per la percentuale di diplomati", col = colors)

abline(lm(lmy$res ~ lmx1$res)) #aggiungo la retta di regressione

text(lmx1$residuals[51],lmy$residuals[51]-2, labels = "DC") #aggiungo l'etichetta al dato del DC

#Partial regression plot controllando per x1

lmyy = lm(y ~ x1)

lmx2 = lm(x2 ~ x1)

plot(lmx2$res, lmyy$res, xlab = "Redisui del modello 'percentuale di diplomati'

regredito su 'tasso di povertà'",

ylab = "Residui del modello 'tasso di omicidi' regredito su 'tasso di povertà'",

main = "Partial regression plot controllando

per il tasso di povertà", col = colors)

abline(lm(lmyy$res ~ lmx2$res)) #aggiungo la retta di regressione

text(lmx2$residuals[51],lmyy$residuals[51]-2, labels = "DC") #aggiungo l'etichetta al dato del DC (Distretto Columbia)

par(mfrow = c(1,1))

#Commento: il grafico a sinistra rappresenta la relazione tra il tasso di omicidi e il

#tasso di povertà esludendo gli effetti della percentuale di diplomati, mentre

#il grafico di destra rappresenta la relazione tra la percentuale di diplomati e il tasso

#di povertà escludendo gli effetti del tasso di povertà.

#Si può concludere che il tasso di omicidio e il tasso di povertà hanno una

#correlaazione maggiore di quella che hanno il tasso di omicidi e la percentuale di diplomati

#Inoltre, in entrambi i casi emerge come il Distretto della Columbia (in rosso) sia un outlier

#Analisi del modello di regressione multipla

summary(lm2)

alpha = as.numeric(lm2$coefficients[1]); beta1 = as.numeric(lm2$coefficients[2]);

beta2 = as.numeric(lm2$coefficients[3])

cbind(alpha, beta1, beta2)

#Commento: per ogni incremento unitario del tasso di povertà, il tasso di omicidi medio aumenta di 1.6.

#Per ogni incremento unitario della percentuale di diplomati, il tasso di omicidio medio aumenta di 0.58

#È strano questo ultimo risultato ottenuto perchè ci si aspetterebbe che all'aumentare del livello di scolarizzazione, il tasso di omicidio

#cali. Inoltre, è anche strano vedere che il tasso di omicidi aumenti più che proporzionalmente al tasso di povertà (vorrebbe dire che per

#ogni punto percentuale del tasso di povertà in più ci sono 1.6 omicidi in più).

#Per questo motivo si stima nuovamente il modello con l'esclusione del dato del District of Columbia che sembra essere un outlier

# Modello di regressione multipla con l'esclusione del dato del DC

NoDC = df2$Murder<df2$Murder[51]

yNoDC = df2$Murder[NoDC]; x1NoDC = df2$Poverty[NoDC]; x2NoDC = df2$HighSch[NoDC]

lm2NoDC = lm(yNoDC ~ x1NoDC + x2NoDC) #modello di regressione lineare multivariato (senza DC)

summary(lm2NoDC)

#Analisi del modello di regressione multipla (senza DC)

summary(lm2NoDC)

alphaNoDC = as.numeric(lm2NoDC$coefficients[1]); beta1NoDC =

as.numeric(lm2NoDC$coefficients[2]); beta2NoDC =

as.numeric(lm2NoDC$coefficients[3])

rbind(cbind("Intercetta","b - tasso di povertà senza DC", "b - tasso di diplomati senza

DC"),cbind(as.numeric(alphaNoDC), as.numeric(beta1NoDC),

as.numeric(beta2NoDC)))

#Commento: In questo caso i dati sembrano più ragionevoli, in quanto all'aumentare di una unità del tasso di povertà, la media del tasso di omicidio

#aumenta di 0.30 (meno che proporzionale), mentre all'aumentare del tasso di scolarizzazione diminuisce il tasso di omicidi medio (b = - 0.19)

#L'intercetta alpha (18.91) rappresenta il tasso di omicidi frizionale, ovvero il tasso di omicidi registrato nella società quando il tasso di povertà è nullo, ma anche la percentuale di diplomati all'high school

# Riassunto dei quattro partial regression plot

par(mfrow = c(2,2))

#Partial regression plot con DC

plot(lmx1$res, lmy$res, xlab = "Redisui del modello 'tasso di povertà' regredito su

'percentuale di diplomati'",

ylab = "Residui del modello 'tasso di omicidi'

regredito su 'percentuale di diplomati'",

main = "Partial regression plot controllando

per la percentuale di diplomati", col = colors)

abline(lm(lmy$res ~ lmx1$res)) #aggiungo la retta di regressione

text(lmx1$residuals[51],lmy$residuals[51]-2, labels = "DC") #aggiungo l'etichetta al dato del DC

plot(lmx2$res, lmyy$res, xlab = "Redisui del modello 'percentuale di diplomati'

regredito su 'tasso di povertà'",

ylab = "Residui del modello 'tasso di omicidi'

regredito su 'tasso di povertà'",

main = "Partial regression plot controllando

per il tasso di povertà", col = colors)

abline(lm(lmyy$res ~ lmx2$res)) #aggiungo la retta di regressione

text(lmx2$residuals[51],lmyy$residuals[51]-2, labels = "DC") #aggiungo l'etichetta al dato del DC

#Partial regression plot senza DC

lmyNoDC = lm(yNoDC ~ x2NoDC)

lmx1NoDC = lm(x1NoDC ~ x2NoDC)

plot(lmx1NoDC$res, lmyNoDC$res, xlab = "Redisui del modello 'tasso di povertà'

regredito su 'percentuale di diplomati'",

ylab = "Residui del modello 'tasso di omicidi'

regredito su 'percentuale di diplomati'",

main = "Partial regression plot controllando per

la percentuale di diplomati (senza DC)")

abline(lm(lmyNoDC$res ~ lmx1NoDC$res)) #aggiungo la retta di regressione

lmyyNoDC = lm(yNoDC ~ x1NoDC)

lmx2NoDC = lm(x2NoDC ~ x1NoDC)

plot(lmx2NoDC$res, lmyyNoDC$res, xlab = "Redisui del modello 'percentuale di diplomati'

regredito su 'tasso di povertà'",

ylab = "Residui del modello 'tasso di omicidi'

regredito su 'tasso di povertà'",

main = "Partial regression plot controllando

per il tasso di povertà (senza DC)")

abline(lm(lmyyNoDC$res ~ lmx2NoDC$res)) #aggiungo la retta di regressione

par(mfrow = c(1,1))

#Commento: dall'analisi dei partial regression plot senza l'outlier DC si evince che la relazione tra tasso di povertà e tasso di omicidi non è così forte

#come sembrava inzialmente (includendo DC), mentre la relazione tra tasso di omicidi e percentuali di diplomati allhigh school è pressoché costante con o senza il dato

#del Distretto della Columbia in valore nominale, ma cambia di segno.

SOLUTION:

> # Introduzione

> df2 = read.csv("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/2004\ statewide\ crime.txt", header = TRUE)

> # Modello di regressione multipla

> #y = tasso di omicidi; x1 = tasso di povertà; x2 = diplomati all'high school

> y = df2$Murder; x1 = df2$Poverty; x2 = df2$HighSch

> pairs(cbind(y,x1,x2)) #mostro su dei grafici le tre variabili

> lm2 = lm(y ~ x1 + x2) #modello di regressione lineare multivariato

> summary(lm2)

Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min 1Q Median 3Q Max

-7.9194 -2.0351 -0.1987 1.4771 24.0338

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -60.4982 24.6153 -2.458 0.0176 \*

x1 1.6049 0.3010 5.332 2.58e-06 \*\*\*

x2 0.5877 0.2605 2.256 0.0287 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.764 on 48 degrees of freedom

Multiple R-squared: 0.4049, Adjusted R-squared: 0.3801

F-statistic: 16.33 on 2 and 48 DF, p-value: 3.891e-06

> #Partial regression plot controllando per x2

> lmy = lm(y ~ x2)

> lmx1 = lm(x1 ~ x2)

> par(mfrow = c(1,2))

> colors = rep("black", length(y)); colors[51] = "red"

> plot(lmx1$res, lmy$res, xlab = "Redisui del modello 'tasso di povertà' regredito su

+ 'percentuale di diplomati'",

+ ylab = "Residui del modello 'tasso di omicidi' regredito su 'percentuale di diplomati'",

+ main = "Partial regression plot controllando

+ per la percentuale di diplomati", col = colors)

> abline(lm(lmy$res ~ lmx1$res)) #aggiungo la retta di regressione

> text(lmx1$residuals[51],lmy$residuals[51]-2, labels = "DC") #aggiungo l'etichetta al dato del DC

> #Partial regression plot controllando per x1

> lmyy = lm(y ~ x1)

> lmx2 = lm(x2 ~ x1)

> plot(lmx2$res, lmyy$res, xlab = "Redisui del modello 'percentuale di diplomati'

+ regredito su 'tasso di povertà'",

+ ylab = "Residui del modello 'tasso di omicidi' regredito su 'tasso di povertà'",

+ main = "Partial regression plot controllando

+ per il tasso di povertà", col = colors)

> abline(lm(lmyy$res ~ lmx2$res)) #aggiungo la retta di regressione

> text(lmx2$residuals[51],lmyy$residuals[51]-2, labels = "DC") #aggiungo l'etichetta al dato del DC (Distretto Columbia)

> par(mfrow = c(1,1))

> #Analisi del modello di regressione multipla

> summary(lm2)

Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min 1Q Median 3Q Max

-7.9194 -2.0351 -0.1987 1.4771 24.0338

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -60.4982 24.6153 -2.458 0.0176 \*

x1 1.6049 0.3010 5.332 2.58e-06 \*\*\*

x2 0.5877 0.2605 2.256 0.0287 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.764 on 48 degrees of freedom

Multiple R-squared: 0.4049, Adjusted R-squared: 0.3801

F-statistic: 16.33 on 2 and 48 DF, p-value: 3.891e-06

> alpha = as.numeric(lm2$coefficients[1]); beta1 = as.numeric(lm2$coefficients[2]);

> beta2 = as.numeric(lm2$coefficients[3])

> cbind(alpha, beta1, beta2)

alpha beta1 beta2

[1,] -60.49824 1.604876 0.5876648

> # Modello di regressione multipla con l'esclusione del dato del DC

> NoDC = df2$Murder<df2$Murder[51]

> yNoDC = df2$Murder[NoDC]; x1NoDC = df2$Poverty[NoDC]; x2NoDC = df2$HighSch[NoDC]

> lm2NoDC = lm(yNoDC ~ x1NoDC + x2NoDC) #modello di regressione lineare multivariato (senza DC)

> summary(lm2NoDC)

Call:

lm(formula = yNoDC ~ x1NoDC + x2NoDC)

Residuals:

Min 1Q Median 3Q Max

-3.7641 -1.3474 -0.2136 1.2162 6.3625

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 18.9123 12.4373 1.521 0.1351

x1NoDC 0.3035 0.1644 1.846 0.0712 .

x2NoDC -0.1960 0.1298 -1.510 0.1378

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.136 on 47 degrees of freedom

Multiple R-squared: 0.3382, Adjusted R-squared: 0.31

F-statistic: 12.01 on 2 and 47 DF, p-value: 6.123e-05

> #Analisi del modello di regressione multipla (senza DC)

> summary(lm2NoDC)

Call:

lm(formula = yNoDC ~ x1NoDC + x2NoDC)

Residuals:

Min 1Q Median 3Q Max

-3.7641 -1.3474 -0.2136 1.2162 6.3625

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 18.9123 12.4373 1.521 0.1351

x1NoDC 0.3035 0.1644 1.846 0.0712 .

x2NoDC -0.1960 0.1298 -1.510 0.1378

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.136 on 47 degrees of freedom

Multiple R-squared: 0.3382, Adjusted R-squared: 0.31

F-statistic: 12.01 on 2 and 47 DF, p-value: 6.123e-05

> alphaNoDC = as.numeric(lm2NoDC$coefficients[1]); beta1NoDC =

+ as.numeric(lm2NoDC$coefficients[2]); beta2NoDC =

+ as.numeric(lm2NoDC$coefficients[3])

> rbind(cbind("Intercetta","b - tasso di povertà senza DC", "b - tasso di diplomati senza

+ DC"),cbind(as.numeric(alphaNoDC), as.numeric(beta1NoDC),

+ as.numeric(beta2NoDC)))

[,1] [,2] [,3]

[1,] "Intercetta" "b - tasso di povertà senza DC" "b - tasso di diplomati senza\nDC"

[2,] "18.9122949220433" "0.303500053376645" "-0.195951337202001"

> #Commento: In questo caso i dati sembrano più ragionevoli, in quanto all'aumentare di una unità del tasso di povertà, la media del tasso di omicidio

> #aumenta di 0.30 (meno che proporzionale), mentre all'aumentare del tasso di scolarizzazione diminuisce il tasso di omicidi medio (b = - 0.19)

> #L'intercetta alpha (18.91) rappresenta il tasso di omicidi frizionale, ovvero il tasso di omicidi registrato nella società quando il tasso di povertà è nullo, ma anche la percentuale di diplomati all'high school

> # Riassunto dei quattro partial regression plot

> par(mfrow = c(2,2))

> #Partial regression plot con DC

> plot(lmx1$res, lmy$res, xlab = "Redisui del modello 'tasso di povertà' regredito su

+ 'percentuale di diplomati'",

+ ylab = "Residui del modello 'tasso di omicidi'

+ regredito su 'percentuale di diplomati'",

+ main = "Partial regression plot controllando

+ per la percentuale di diplomati", col = colors)

> abline(lm(lmy$res ~ lmx1$res)) #aggiungo la retta di regressione

> text(lmx1$residuals[51],lmy$residuals[51]-2, labels = "DC") #aggiungo l'etichetta al dato del DC

> plot(lmx2$res, lmyy$res, xlab = "Redisui del modello 'percentuale di diplomati'

+ regredito su 'tasso di povertà'",

+ ylab = "Residui del modello 'tasso di omicidi'

+ regredito su 'tasso di povertà'",

+ main = "Partial regression plot controllando

+ per il tasso di povertà", col = colors)

> abline(lm(lmyy$res ~ lmx2$res)) #aggiungo la retta di regressione

> text(lmx2$residuals[51],lmyy$residuals[51]-2, labels = "DC") #aggiungo l'etichetta al dato del DC

> #Partial regression plot senza DC

> lmyNoDC = lm(yNoDC ~ x2NoDC)

> lmx1NoDC = lm(x1NoDC ~ x2NoDC)

> plot(lmx1NoDC$res, lmyNoDC$res, xlab = "Redisui del modello 'tasso di povertà'

+ regredito su 'percentuale di diplomati'",

+ ylab = "Residui del modello 'tasso di omicidi'

+ regredito su 'percentuale di diplomati'",

+ main = "Partial regression plot controllando per

+ la percentuale di diplomati (senza DC)")

> abline(lm(lmyNoDC$res ~ lmx1NoDC$res)) #aggiungo la retta di regressione

> lmyyNoDC = lm(yNoDC ~ x1NoDC)

> lmx2NoDC = lm(x2NoDC ~ x1NoDC)

> plot(lmx2NoDC$res, lmyyNoDC$res, xlab = "Redisui del modello 'percentuale di diplomati'

+ regredito su 'tasso di povertà'",

+ ylab = "Residui del modello 'tasso di omicidi'

+ regredito su 'tasso di povertà'",

+ main = "Partial regression plot controllando

+ per il tasso di povertà (senza DC)")

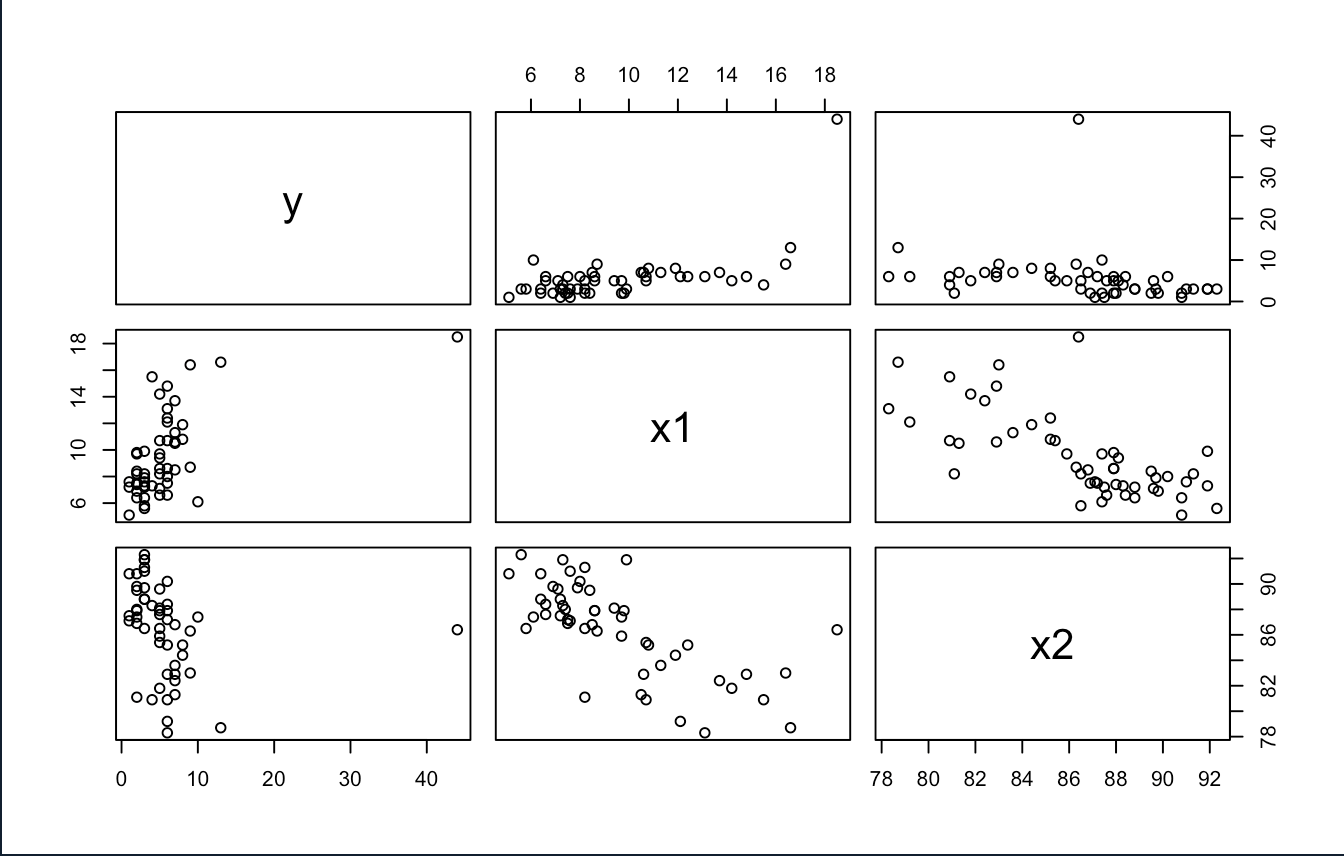
> abline(lm(lmyyNoDC$res ~ lmx2NoDC$res)) #aggiungo la retta di regressione

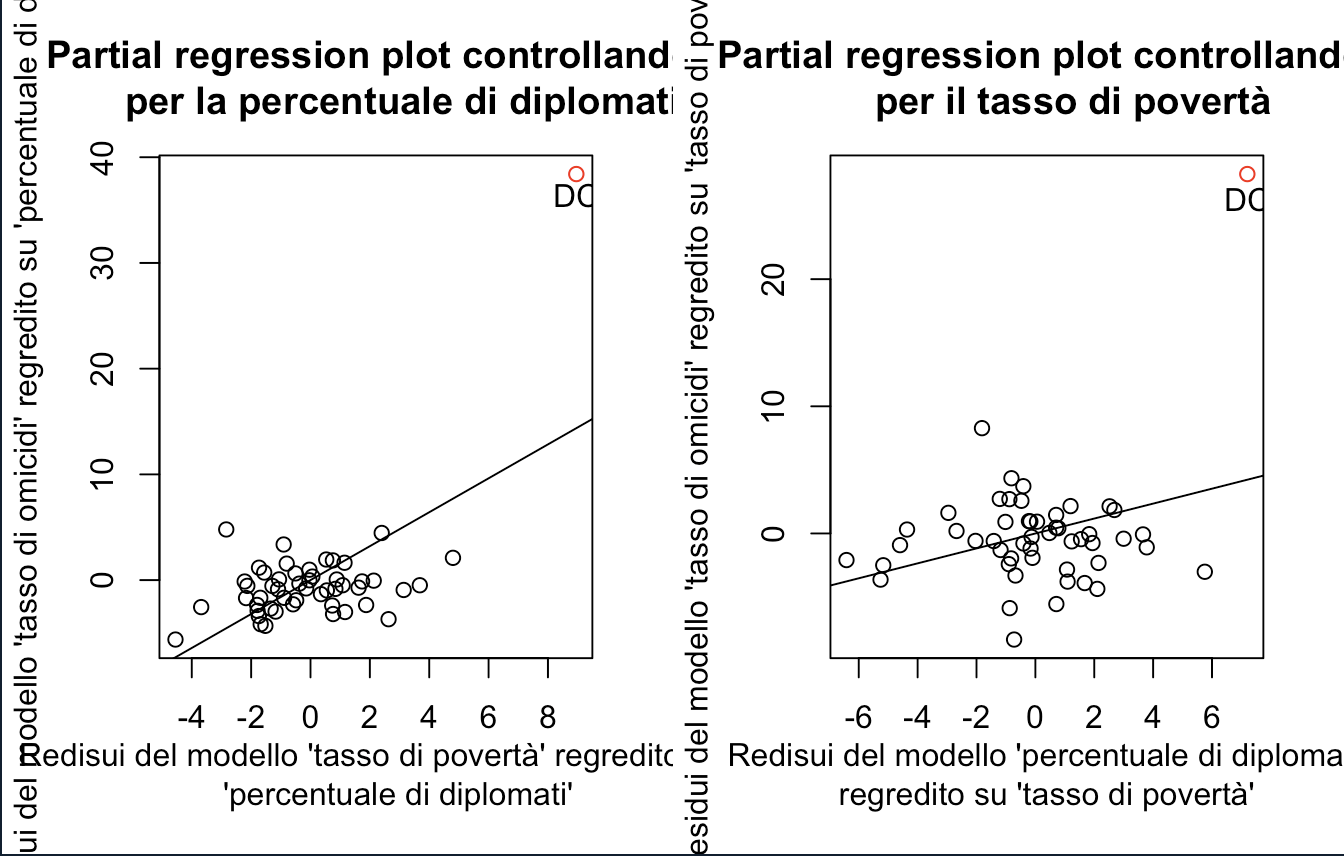
> par(mfrow = c(1,1))

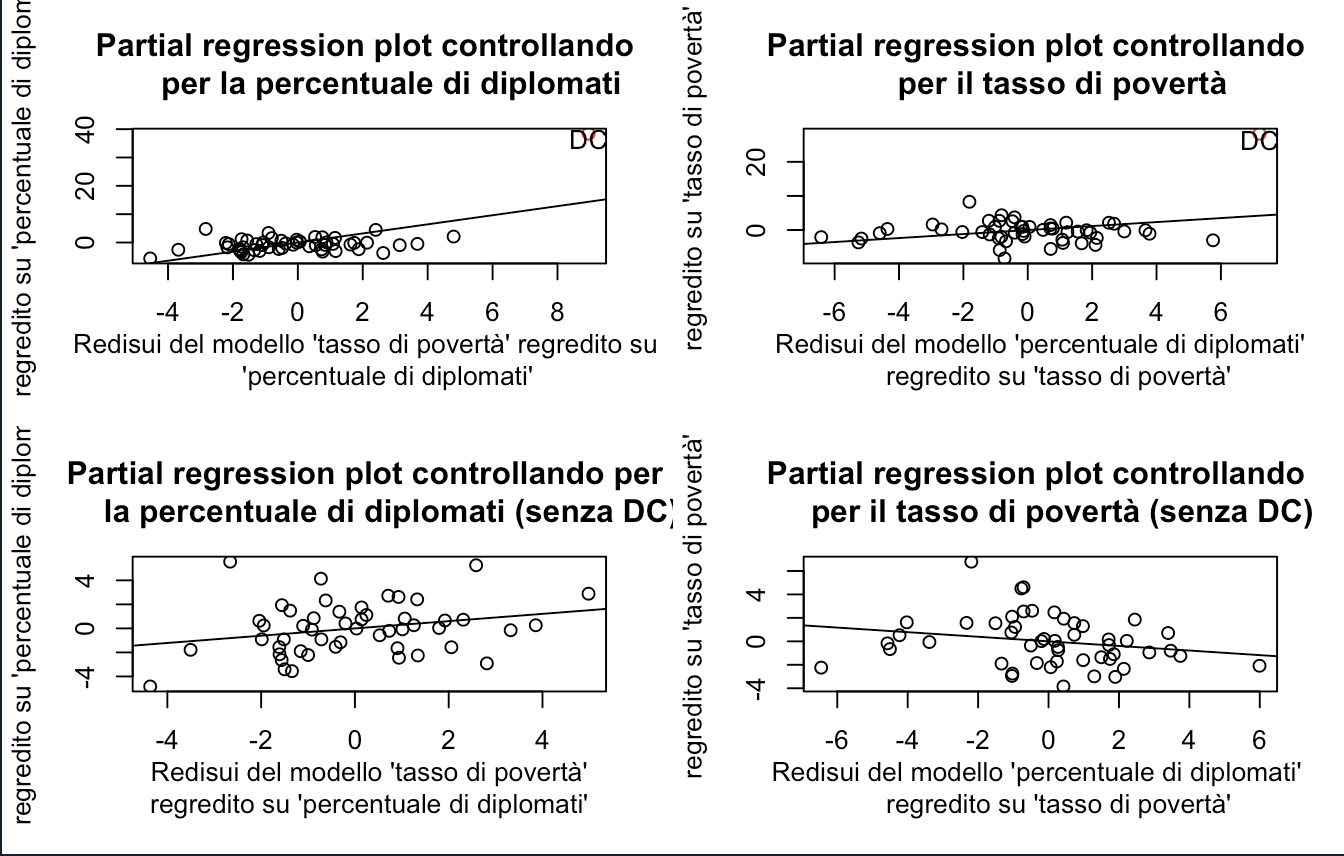
> # ESERCIZIO 3 --------------------------

> # Svolgimento esercizio 30

> df = read.csv("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/lavoro.txt", header = TRUE)







# ESERCIZIO 3 --------------------------

# Svolgimento esercizio 30

df = read.csv("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/lavoro.txt", header = TRUE)

str(df) #Controllo che le colonne della tabella siano definite

# Non sono definite quindi le separo

# Utilizzando il parametro "sep"

df <- read.table("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/lavoro.txt", header=TRUE, sep="\t")

str(df) # Controllo nuovamente che le colonne della tabella siano definite separatamente

# RICHIESTA 1

#Regressione di y (average\_score) su x1 (sex) e x2 (years\_service)

avgPunteggio = df$Average\_Score; sex = df$Sex; AnniLavoro = df$Years\_Service; race = df$Race; #dichiaro le variabili

fit30 = lm(avgPunteggio ~ sex + AnniLavoro); summary(fit30) #modello di regressione lineare

yTeorica = df$Average\_Score;

yPrevista = fit30$fitted.values

cor(yTeorica, yPrevista);

plot(yTeorica, yPrevista) #correlazione multipla del modello

#Commento: la correlazione elevata (che si nota anche dal grafico) evidenzia come il modello di regressione lineare stimato (fit30) sia un buon stimatore della variabile y = punteggio medio

#Test globale di indipendenza H0: bSex = bAnniLavoro = 0

(statisticaF = as.numeric(summary(fit30)$fstatistic[1]))

qf(0.05, 2, length(sex)-2-1, lower.tail = FALSE) #valore critico

pf(statisticaF, 2, length(sex)-2-1, lower.tail = FALSE) #pvalue (calcolato "a mano")

#Rifiuto H0: bSex = bAnniLavoro = 0, quindi almeno uno tra bSex e bAnniLavoro è diverso da 0

#Test H0: bSex = 0 e test H0: bAnniLavoro = 0

(testTsex = summary(fit30)$coefficients[8]); (testAnnilavoro =

summary(fit30)$coefficients[9])

qt(0.025, length(sex)-2-1, lower.tail = FALSE) #valore critico

(pvalueSex = summary(fit30)$coefficients[11]); (pvalueAnniLavoro =

summary(fit30)$coefficients[12])

#rifiuto H0: bSex = 0; H0: bAnniLavoro = 0 perchè testTsex < -ValoreCritico e testAnnilavoro > ValoreCritico (e entrambi hanno un pvalue prossimo allo 0)

# Commento: Il coordinatore può usare gli anni di lavoro e il sesso del dipendente per stimare il punteggio medio

# Rappresentazione grafica dei dati

#Regressione di y (average\_score) su x1 (sex), x2 (years\_service) e x3 (race) fit3 = lm(avgPunteggio ~ sex + AnniLavoro + race); summary(fit3) #modello di regressione lineare

#Rappresentazione grafica

#Gestione dei colori: le donne sono in rosa, gli uomini in blu

colors = rep(NA, length(df$Employee)) #vettore dei colori

counterS = 1 #contatore per ciclo while

#ciclo while per assegnare i colori

while(counterS <= length(df$Employee)){

if(sex[counterS] == "Male"){

colors[counterS] = "blue"

counterS = counterS + 1

} else {

colors[counterS] = "red"

counterS = counterS + 1

}

}

#Gestione dei punti: i bianchi sono rappresentati da dei triangoli, i non bianchi sono da dei cerchi

point = rep(NA, length(df$Employee)) #vettore dei colori

counterR = 1 #contatore per ciclo while

#ciclo while per assegnare i colori

while(counterR <= length(df$Employee)){

if(race[counterR] == "White"){

point[counterR] = 2 #2 è il codice per il triangoli

counterR = counterR + 1

} else {

point[counterR] = 1 #1 è il codice del cerchio

counterR = counterR + 1

}

}

#Grafico (con legenda):

plot(df$Employee, avgPunteggio, pch = point, col = colors, xlab = "Codice

dell'impiegato", ylab = "Punteggio medio")

legend("bottomleft", legend = c("Maschi - bianchi", "Maschi - non bianchi", "Femmine

- bianche", "Femmine - non bianche"), col = c("blue", "blue", "red", "red"), pch = c(2,

1, 2, 1), lwd = 1, bty = "n")

#Commento: dal grafico si evince come le donne abbiano ottenuto un punteggio medio maggiore di quello degli uomini.

#Nel gruppo delle donne non si nota nessuna associazione palese tra la razza e il punteggio medio,

#mentre nel tra gli uomini si vede facilmente come i maschi bianchi abbiano ottenuto un punteggio medio maggiore dei maschi non bianchi

# RICHIESTA 2

# Analisi dell'interazione

#Regressione di y (average\_score) su x1 (sex), x2 (years\_service) e x3 (race), consideraziondo l interazione x2x1 e x2x3

fit3.1 = lm(avgPunteggio ~ sex + AnniLavoro + race + AnniLavoro\*sex + AnniLavoro\*race); summary(fit3.1) #modello di regressione lineare

#In automatico R mi imposta male = 1, female = 0 e white = 1, non white = 0

#testF

anova(fit30, fit3.1)

testX3 = anova(fit30, fit3.1)$F[2]; pValueX3 = anova(fit30, fit3.1)$`Pr(>F)`[2] #Valore del test e del pvalue

cbind(testX3, pValueX3)

#Non ci sono abbastanza evidenze statistiche per rifiutare H0: b interazione annilavoro\*sex = b interazione annilavoro\*race =

#Dal test globale di indipendenza si evince quindi che il modello con le interazioni non siano significative

#Prendo il valore della statistica test t e il p-value della stima di b per le due interpolazioni

tAS = summary(fit3.1)$coeff[17]; pAS = summary(fit3.1)$coeff[23]

tAR = summary(fit3.1)$coeff[18]; pAR = summary(fit3.1)$coeff[24]

#Risultato del test (facendo riferimento alla statistica test t)

ValoreCritico = qt(0.025, length(AnniLavoro)-5-1, lower.tail = FALSE)

cbind(tAS,pAS,tAR,pAR,ValoreCritico) #tabella riassuntila

if(tAS < ValoreCritico | tAS > -ValoreCritico){

print("Risultato test su b di anni di lavoro-sesso: non ci sono abbastanza evidenze

statistiche per rifiutare H0: bAS = 0, quindi l'interpolazione tra anni di lavoro e sesso

non è statisticamente significativa (livello di significativà al 5%)")

} else {

print("Risultato test su b di anni di lavoro-sesso: si rifiuta H0: bAS = 0, quindi

l'interpolazione tra anni di lavoro e sesso è statisticamente significativa (livello di

significativà al 5%)")

}

if(tAR < ValoreCritico | tAR > -ValoreCritico){

print("Risultato test su b di anni di lavoro-sesso: non ci sono abbastanza evidenze

statistiche per rifiutare H0: bAR = 0, quindi l'interpolazione tra anni di lavoro e razza

non è statisticamente significativa (livello di significativà al 5%)")

} else {

print("Risultato test su b di anni di lavoro-sesso: si rifiuta H0: bAR = 0, quindi

l'interpolazione tra anni di lavoro e razza è statisticamente significativa (livello di

significativà al 5%)")

}

# RICHIESTE 3 & 4

summary(fit30) #Modello senza le due interpolazioni non significative

# Grafici di diagnostica del modello

#imposto per il sesso male = 1 e female = o

i = 1; sex.num = rep(NA, length(sex))

while(i <= length(sex)){

if(sex[i] == "Male"){

sex.num[i] = 1

i = i+1

} else {

sex.num[i] = 0

i = i+1

}

}

#imposto per la razza white = 1 e non white = o

j = 1; race.num = rep(NA, length(race))

while(j <= length(race)){

if(race[j] == "White"){

race.num[j] = 1

j = j+1

} else {

race.num[j] = 0

j = j+1

}

}

pairs(cbind(avgPunteggio, AnniLavoro, sex.num, race.num)) #riporto le relazioni sul grafico 2 a 2

cor(cbind(avgPunteggio, AnniLavoro, sex.num, race.num)) #calcolo le correlazioni tra le variabili 2 a 2

#Commento: dall'analisi dei grafici e della matrice delle correlazioni si evince come ci sia una correlazione negativa elevata tra il punteggio

#medio e il sesso, ovvero il punteggio medio aumenta quando il sesso passa da 1 a 0, quindi da maschi a femmine (evidenza già riscontrata nel grafico precedente)

#Inoltre, tra il punteggio medio è correlato positivamente a circa 0.6 con gli anni di lavoro. Le altre variabili hanno una correlazione bassa

SOLUTION:

> # ESERCIZIO 3 --------------------------

> # Svolgimento esercizio 30

> df = read.csv("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/lavoro.txt", header = TRUE)

> str(df) #Controllo che le colonne della tabella siano definite

'data.frame': 20 obs. of 1 variable:

$ Employee.Average\_Score.Years\_Service.Sex.Race: chr "1\t7.6\t5\tFemale\tNonwhite" "2\t9\t30\tFemale\tNonwhite" "3\t8\t12\tFemale\tNonwhite" "4\t6.8\t7\tFemale\tNonwhite" ...

> # Non sono definite quindi le separo

> # Utilizzando il parametro "sep"

> df <- read.table("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/lavoro.txt", header=TRUE, sep="\t")

> str(df) # Controllo nuovamente che le colonne della tabella siano definite separatamente

'data.frame': 20 obs. of 5 variables:

$ Employee : int 1 2 3 4 5 6 7 8 9 10 ...

$ Average\_Score: num 7.6 9 8 6.8 7.4 9.8 7.4 8.9 8.1 8.6 ...

$ Years\_Service: int 5 30 12 7 7 27 4 6 7 11 ...

$ Sex : chr "Female" "Female" "Female" "Female" ...

$ Race : chr "Nonwhite" "Nonwhite" "Nonwhite" "Nonwhite" ...

> # RICHIESTA 1

> #Regressione di y (average\_score) su x1 (sex) e x2 (years\_service)

> avgPunteggio = df$Average\_Score; sex = df$Sex; AnniLavoro = df$Years\_Service; race = df$Race; #dichiaro le variabili

> fit30 = lm(avgPunteggio ~ sex + AnniLavoro); summary(fit30) #modello di regressione lineare

Call:

lm(formula = avgPunteggio ~ sex + AnniLavoro)

Residuals:

Min 1Q Median 3Q Max

-1.23832 -0.49061 -0.05023 0.49141 1.49221

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.03542 0.35862 19.618 4.10e-13 \*\*\*

sexMale -2.59099 0.36058 -7.186 1.52e-06 \*\*\*

AnniLavoro 0.09695 0.02228 4.351 0.000435 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7861 on 17 degrees of freedom

Multiple R-squared: 0.8394, Adjusted R-squared: 0.8205

F-statistic: 44.44 on 2 and 17 DF, p-value: 1.771e-07

> yTeorica = df$Average\_Score;

> yPrevista = fit30$fitted.values

> cor(yTeorica, yPrevista);

[1] 0.9162022

> plot(yTeorica, yPrevista) #correlazione multipla del modello

> #Test globale di indipendenza H0: bSex = bAnniLavoro = 0

> (statisticaF = as.numeric(summary(fit30)$fstatistic[1]))

[1] 44.43527

> qf(0.05, 2, length(sex)-2-1, lower.tail = FALSE) #valore critico

[1] 3.591531

> pf(statisticaF, 2, length(sex)-2-1, lower.tail = FALSE) #pvalue (calcolato "a mano")

[1] 1.771036e-07

> #Rifiuto H0: bSex = bAnniLavoro = 0, quindi almeno uno tra bSex e bAnniLavoro è diverso da 0

> #Test H0: bSex = 0 e test H0: bAnniLavoro = 0

> (testTsex = summary(fit30)$coefficients[8]); (testAnnilavoro =

[1] -7.185669

+ summary(fit30)$coefficients[9])

[1] 4.350611

> qt(0.025, length(sex)-2-1, lower.tail = FALSE) #valore critico

[1] 2.109816

> (pvalueSex = summary(fit30)$coefficients[11]); (pvalueAnniLavoro =

[1] 1.524473e-06

+ summary(fit30)$coefficients[12])

[1] 0.0004349557

> #Rappresentazione grafica

> #Gestione dei colori: le donne sono in rosa, gli uomini in blu

> colors = rep(NA, length(df$Employee)) #vettore dei colori

> counterS = 1 #contatore per ciclo while

> #ciclo while per assegnare i colori

> while(counterS <= length(df$Employee)){

+ if(sex[counterS] == "Male"){

+ colors[counterS] = "blue"

+ counterS = counterS + 1

+ } else {

+ colors[counterS] = "red"

+ counterS = counterS + 1

+ }

+ }

> #Gestione dei punti: i bianchi sono rappresentati da dei triangoli, i non bianchi sono da dei cerchi

> point = rep(NA, length(df$Employee)) #vettore dei colori

> counterR = 1 #contatore per ciclo while

> #ciclo while per assegnare i colori

> while(counterR <= length(df$Employee)){

+ if(race[counterR] == "White"){

+ point[counterR] = 2 #2 è il codice per il triangoli

+ counterR = counterR + 1

+ } else {

+ point[counterR] = 1 #1 è il codice del cerchio

+ counterR = counterR + 1

+ }

+ }

> #Grafico (con legenda):

> plot(df$Employee, avgPunteggio, pch = point, col = colors, xlab = "Codice

+ dell'impiegato", ylab = "Punteggio medio")

> legend("bottomleft", legend = c("Maschi - bianchi", "Maschi - non bianchi", "Femmine

+ - bianche", "Femmine - non bianche"), col = c("blue", "blue", "red", "red"), pch = c(2,

+ 1, 2, 1), lwd = 1, bty = "n")

> # RICHIESTA 2

> # Analisi dell'interazione

> #Regressione di y (average\_score) su x1 (sex), x2 (years\_service) e x3 (race), consideraziondo l interazione x2x1 e x2x3

> fit3.1 = lm(avgPunteggio ~ sex + AnniLavoro + race + AnniLavoro\*sex + AnniLavoro\*race); summary(fit3.1) #modello di regressione lineare

Call:

lm(formula = avgPunteggio ~ sex + AnniLavoro + race + AnniLavoro \*

sex + AnniLavoro \* race)

Residuals:

Min 1Q Median 3Q Max

-0.80493 -0.31983 -0.03909 0.24599 0.94029

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.621494 0.328461 20.159 9.66e-12 \*\*\*

sexMale -2.700704 0.380441 -7.099 5.35e-06 \*\*\*

AnniLavoro 0.082729 0.022993 3.598 0.00291 \*\*

raceWhite 1.306699 0.382630 3.415 0.00419 \*\*

sexMale:AnniLavoro 0.008501 0.033229 0.256 0.80180

AnniLavoro:raceWhite -0.013545 0.031301 -0.433 0.67180

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5107 on 14 degrees of freedom

Multiple R-squared: 0.9442, Adjusted R-squared: 0.9243

F-statistic: 47.37 on 5 and 14 DF, p-value: 2.785e-08

> #In automatico R mi imposta male = 1, female = 0 e white = 1, non white = 0

> #testF

> anova(fit30, fit3.1)

Analysis of Variance Table

Model 1: avgPunteggio ~ sex + AnniLavoro

Model 2: avgPunteggio ~ sex + AnniLavoro + race + AnniLavoro \* sex + AnniLavoro \*

race

Res.Df RSS Df Sum of Sq F Pr(>F)

1 17 10.5044

2 14 3.6509 3 6.8534 8.7601 0.001605 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> testX3 = anova(fit30, fit3.1)$F[2]; pValueX3 = anova(fit30, fit3.1)$`Pr(>F)`[2] #Valore del test e del pvalue

> cbind(testX3, pValueX3)

testX3 pValueX3

[1,] 8.760121 0.00160535

> #Prendo il valore della statistica test t e il p-value della stima di b per le due interpolazioni

> tAS = summary(fit3.1)$coeff[17]; pAS = summary(fit3.1)$coeff[23]

> tAR = summary(fit3.1)$coeff[18]; pAR = summary(fit3.1)$coeff[24]

> #Risultato del test (facendo riferimento alla statistica test t)

> ValoreCritico = qt(0.025, length(AnniLavoro)-5-1, lower.tail = FALSE)

> cbind(tAS,pAS,tAR,pAR,ValoreCritico) #tabella riassuntila

tAS pAS tAR pAR ValoreCritico

[1,] 0.255832 0.8018012 -0.4327306 0.6718023 2.144787

> if(tAS < ValoreCritico | tAS > -ValoreCritico){

+ print("Risultato test su b di anni di lavoro-sesso: non ci sono abbastanza evidenze

+ statistiche per rifiutare H0: bAS = 0, quindi l'interpolazione tra anni di lavoro e sesso

+ non è statisticamente significativa (livello di significativà al 5%)")

+ } else {

+

+ print("Risultato test su b di anni di lavoro-sesso: si rifiuta H0: bAS = 0, quindi

+ l'interpolazione tra anni di lavoro e sesso è statisticamente significativa (livello di

+ significativà al 5%)")

+ }

[1] "Risultato test su b di anni di lavoro-sesso: non ci sono abbastanza evidenze\nstatistiche per rifiutare H0: bAS = 0, quindi l'interpolazione tra anni di lavoro e sesso\nnon è statisticamente significativa (livello di significativà al 5%)"

> if(tAR < ValoreCritico | tAR > -ValoreCritico){

+ print("Risultato test su b di anni di lavoro-sesso: non ci sono abbastanza evidenze

+ statistiche per rifiutare H0: bAR = 0, quindi l'interpolazione tra anni di lavoro e razza

+ non è statisticamente significativa (livello di significativà al 5%)")

+ } else {

+ print("Risultato test su b di anni di lavoro-sesso: si rifiuta H0: bAR = 0, quindi

+ l'interpolazione tra anni di lavoro e razza è statisticamente significativa (livello di

+ significativà al 5%)")

+ }

[1] "Risultato test su b di anni di lavoro-sesso: non ci sono abbastanza evidenze\nstatistiche per rifiutare H0: bAR = 0, quindi l'interpolazione tra anni di lavoro e razza\nnon è statisticamente significativa (livello di significativà al 5%)"

> # RICHIESTE 3 & 4

> summary(fit30) #Modello senza le due interpolazioni non significative

Call:

lm(formula = avgPunteggio ~ sex + AnniLavoro)

Residuals:

Min 1Q Median 3Q Max

-1.23832 -0.49061 -0.05023 0.49141 1.49221

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.03542 0.35862 19.618 4.10e-13 \*\*\*

sexMale -2.59099 0.36058 -7.186 1.52e-06 \*\*\*

AnniLavoro 0.09695 0.02228 4.351 0.000435 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7861 on 17 degrees of freedom

Multiple R-squared: 0.8394, Adjusted R-squared: 0.8205

F-statistic: 44.44 on 2 and 17 DF, p-value: 1.771e-07

> # Grafici di diagnostica del modello

> #imposto per il sesso male = 1 e female = o

> i = 1; sex.num = rep(NA, length(sex))

> while(i <= length(sex)){

+ if(sex[i] == "Male"){

+ sex.num[i] = 1

+ i = i+1

+ } else {

+ sex.num[i] = 0

+ i = i+1

+ }

+ }

> #imposto per la razza white = 1 e non white = o

> j = 1; race.num = rep(NA, length(race))

> while(j <= length(race)){

+ if(race[j] == "White"){

+ race.num[j] = 1

+ j = j+1

+ } else {

+ race.num[j] = 0

+ j = j+1

+ }

+ }

> pairs(cbind(avgPunteggio, AnniLavoro, sex.num, race.num)) #riporto le relazioni sul grafico 2 a 2

> cor(cbind(avgPunteggio, AnniLavoro, sex.num, race.num)) #calcolo le correlazioni tra le variabili 2 a 2

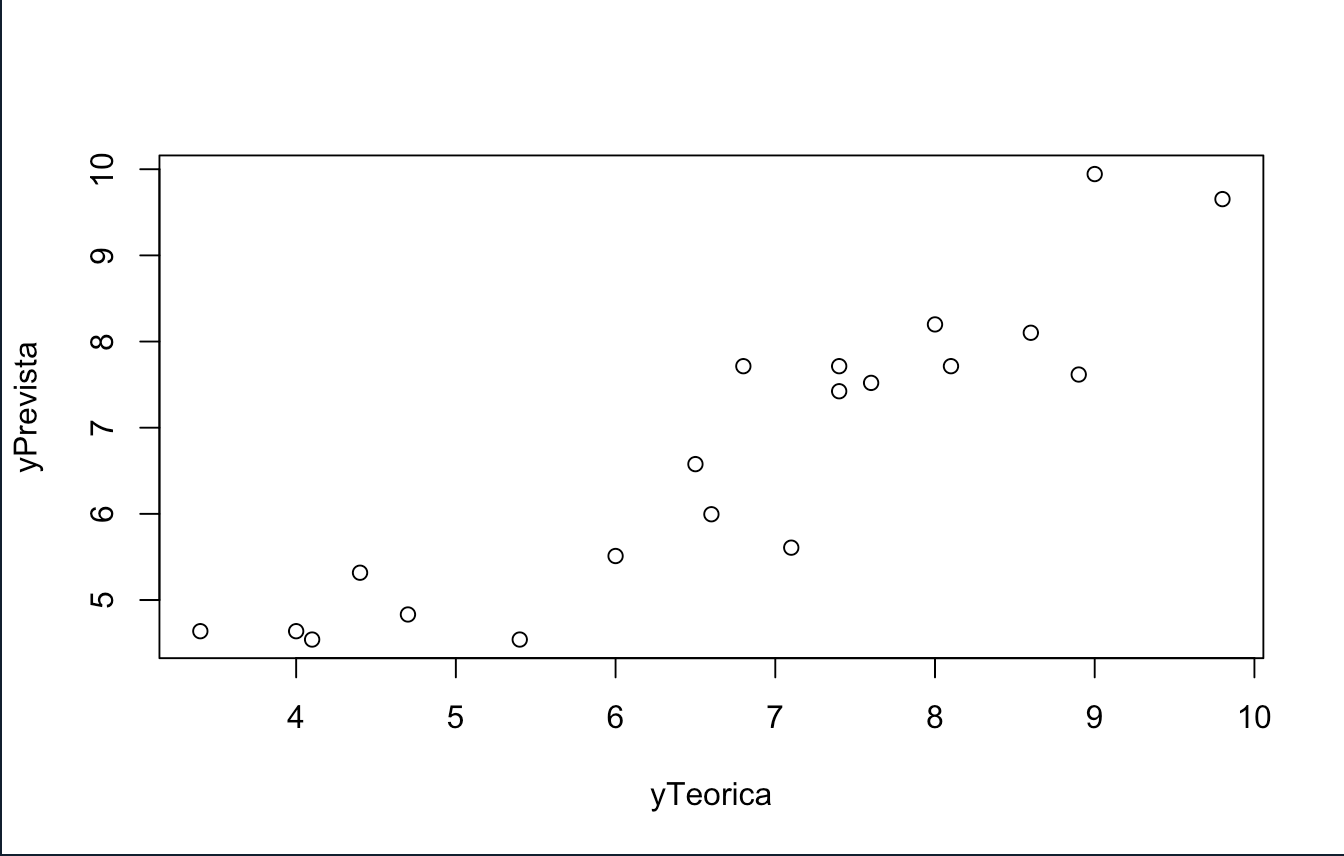
avgPunteggio AnniLavoro sex.num race.num

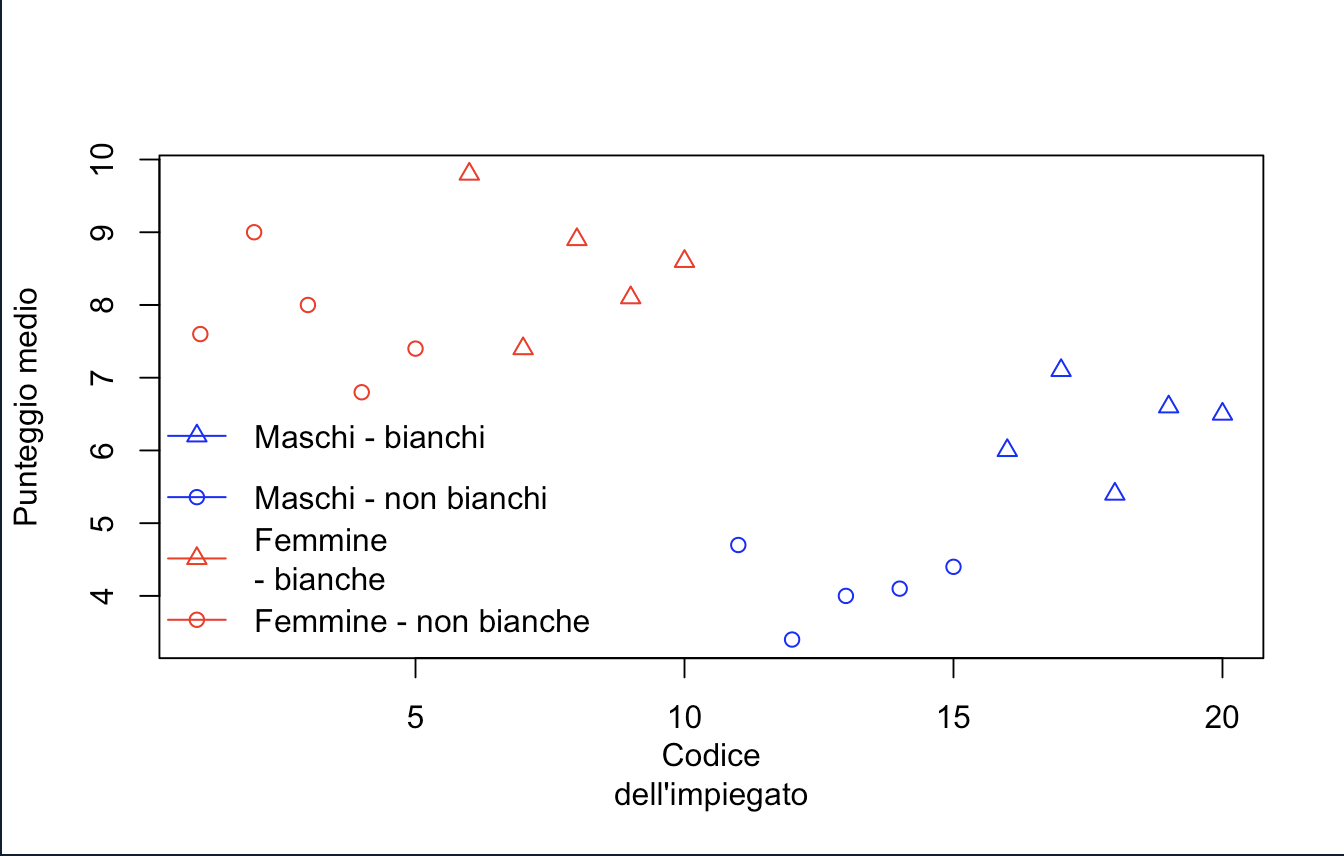
avgPunteggio 1.0000000 0.5930592 -0.8128000 0.4146939

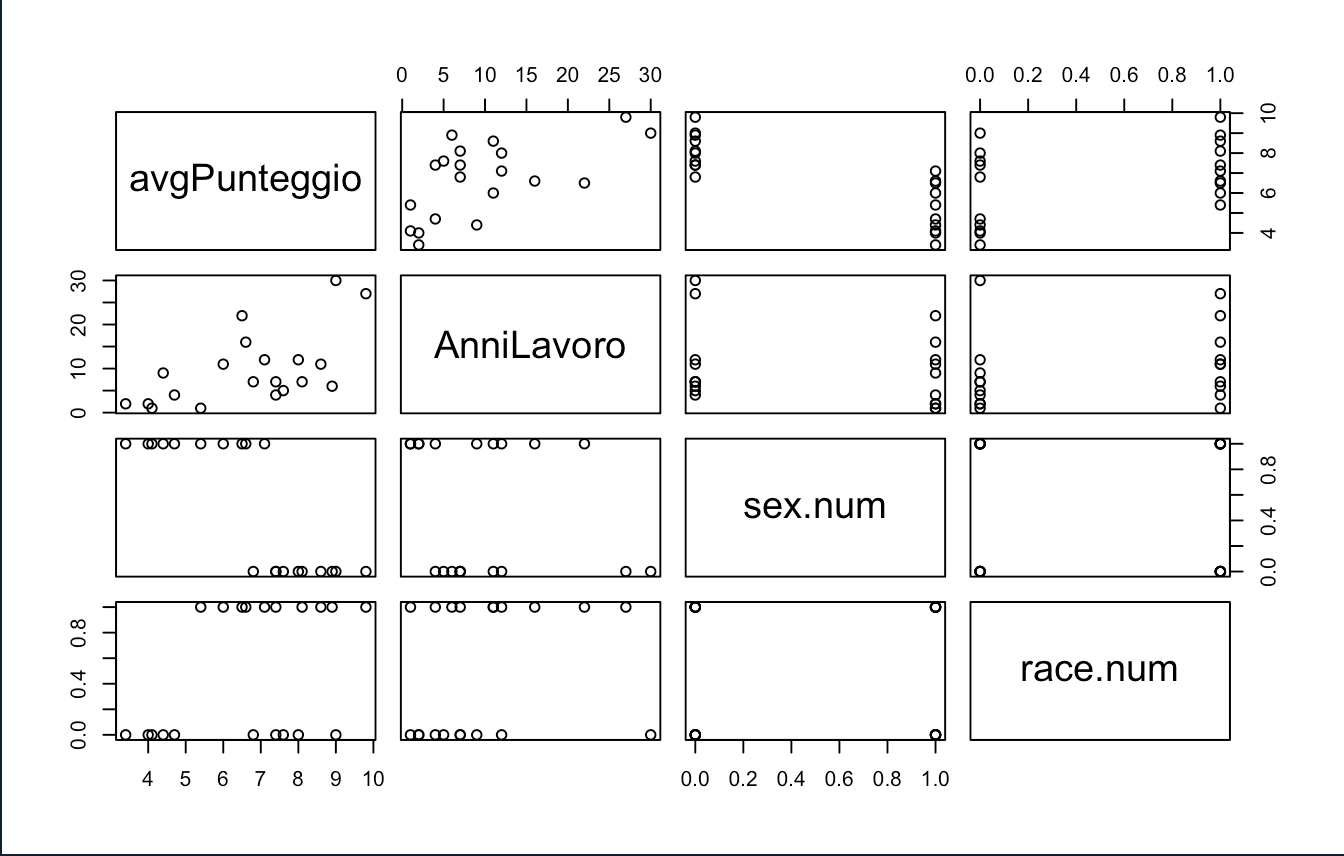
AnniLavoro 0.5930592 1.0000000 -0.2224767 0.2348365

sex.num -0.8128000 -0.2224767 1.0000000 0.0000000

race.num 0.4146939 0.2348365 0.0000000 1.0000000







# ESERCIZIO 4 --------------------------

# Introduzione

#Obiettivo: verificare se il valore della merce consegnata corrisponda effettivamente a quello dichiarato

#$prima = merce di prima caterogia (1) o di seconda categoria (0)

#$ExtraUE = merce che arriva da extraUE (1) o no (0)

#$dichiarato = valore dichiarato del lotto

#$irr = 1 se il valore si discosta da quello dichiarato e 0 se è lo stesso

df4 = read.csv("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/dati\_controlli.csv",header=TRUE)

lotto = df4$lotto; qualità = df4$prima; origine = df4$ExtraUE; valoreD = df4$dichiarato; esitoControllo = df4$Irr

# Relaione tra la variabile Irr (esitoControllo) e le variabili esplicative della merce

logit4 = glm(esitoControllo ~ qualità + origine + valoreD, family = binomial);

summary(logit4)

# Probabilità che un lotto proveniente da UE (origine = 0), con qualità = 1 e valoreD = 200€ sia irregolare (esitoControllo = 1)

(lnOdds4 = as.numeric(logit4$coefficients[1]) + (as.numeric(logit4$coefficients[2])\*1)

+ (as.numeric(logit4$coefficients[3])\*0) + (as.numeric(logit4$coefficients[4])\*200))

(p4 = exp(lnOdds4)/(1+exp(lnOdds4))) #P(Y = 1 | x1 = 1, x2 = 0, x3 = 200)

# Modello logistico che considera anche l'interazione tra qualità e origine

logit4.1 = glm(esitoControllo ~ qualità + origine + valoreD + qualità\*origine, family = binomial); summary(logit4.1)

#Test del rapporto delle verosomiglianze

anova(logit4, logit4.1, test = "Chisq")

testX4 = anova(logit4, logit4.1, test = "Chisq")$Deviance[2]; pValueX4 =anova(logit4,

logit4.1, test = "Chisq")$`Pr(>Chi)`[2] #Valore del test e del pvalue

cbind(testX4, pValueX4)

#Si conclude quindi che è preferibile il modello senza l'interazione, dato che test LR = 5.003 con pvalue = 0.025 (-> rifiuto H0: b interazione = 0)

#Tabella riassuntiva

cbind(rbind("Nome variabile", "Qualità","Origine","Valore dichiarato"),rbind("b",as.numeric(logit4$coefficients[2]),as.numeric(logit4$coefficients[3]),as.numeric(logit4$coefficients[4])),rbind("exp(b)",as.numeric(exp(logit4$coefficients[2])),as.numeric(exp(logit4$coefficients[3])),as.numeric(exp(logit4$coefficients[4]))))

#Analizzando i dati, un consiglio utile agli addetti sarebbe quello di controllare con maggiore attenzionele merci con un alto valore dichiarato, dato che per ogni incremento unitario

#del valore dichiarato, l'odds ratio stimato aumenta di 0.98, ovvero aumenta la probabilità che Irr = 1 (aumenta il numeratore dell'odds, diminuisce il denominatore)

#Anche la provenienza del lotto non è irrilevante dato che l'odds ratio aumenta di 0,89 per ogni incremento unitario della variabile ExtraUE

#In conclusione si consiglia agli addetti di fare maggiore attenzione ai lotti provenienti da paesi non dell'UE e con un alto valore dichiarato

SOLUTION:

> # ESERCIZIO 4 --------------------------

>

> # Introduzione

> #Obiettivo: verificare se il valore della merce consegnata corrisponda effettivamente a quello dichiarato

> #$prima = merce di prima caterogia (1) o di seconda categoria (0)

> #$ExtraUE = merce che arriva da extraUE (1) o no (0)

> #$dichiarato = valore dichiarato del lotto

> #$irr = 1 se il valore si discosta da quello dichiarato e 0 se è lo stesso

> df4 = read.csv("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 2/ESERCITAZIONE\ FINALE/homework\ SPEI\ a.a.\ 2023-2024/dati\_controlli.csv",header=TRUE)

> lotto = df4$lotto; qualità = df4$prima; origine = df4$ExtraUE; valoreD = df4$dichiarato; esitoControllo = df4$Irr

>

> # Relaione tra la variabile Irr (esitoControllo) e le variabili esplicative della merce

> logit4 = glm(esitoControllo ~ qualità + origine + valoreD, family = binomial);

> summary(logit4)

Call:

glm(formula = esitoControllo ~ qualità + origine + valoreD,

family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.1501 -0.5590 -0.3531 -0.1467 3.1049

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.934337 0.280931 3.326 0.000881 \*\*\*

qualità -1.069282 0.368416 -2.902 0.003703 \*\*

origine -0.110042 0.132893 -0.828 0.407643

valoreD -0.014272 0.001516 -9.412 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1913.9 on 2869 degrees of freedom

Residual deviance: 1662.3 on 2866 degrees of freedom

AIC: 1670.3

Number of Fisher Scoring iterations: 7

>

> # Probabilità che un lotto proveniente da UE (origine = 0), con qualità = 1 e valoreD = 200€ sia irregolare (esitoControllo = 1)

> (lnOdds4 = as.numeric(logit4$coefficients[1]) + (as.numeric(logit4$coefficients[2])\*1)

+ + (as.numeric(logit4$coefficients[3])\*0) + (as.numeric(logit4$coefficients[4])\*200))

[1] -2.989325

> (p4 = exp(lnOdds4)/(1+exp(lnOdds4))) #P(Y = 1 | x1 = 1, x2 = 0, x3 = 200)

[1] 0.04791045

>

> # Modello logistico che considera anche l'interazione tra qualità e origine

> logit4.1 = glm(esitoControllo ~ qualità + origine + valoreD + qualità\*origine, family = binomial); summary(logit4.1)

Call:

glm(formula = esitoControllo ~ qualità + origine + valoreD +

qualità \* origine, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.1600 -0.5571 -0.3513 -0.1222 3.3075

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.958541 0.281176 3.409 0.000652 \*\*\*

qualità -1.854169 0.609737 -3.041 0.002358 \*\*

origine -0.172665 0.136462 -1.265 0.205766

valoreD -0.014281 0.001516 -9.417 < 2e-16 \*\*\*

qualità:origine 1.499095 0.707329 2.119 0.034059 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1913.9 on 2869 degrees of freedom

Residual deviance: 1657.3 on 2865 degrees of freedom

AIC: 1667.3

Number of Fisher Scoring iterations: 7

> #Test del rapporto delle verosomiglianze

> anova(logit4, logit4.1, test = "Chisq")

Analysis of Deviance Table

Model 1: esitoControllo ~ qualità + origine + valoreD

Model 2: esitoControllo ~ qualità + origine + valoreD + qualità \* origine

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 2866 1662.3

2 2865 1657.3 1 5.0028 0.02531 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> testX4 = anova(logit4, logit4.1, test = "Chisq")$Deviance[2]; pValueX4 =anova(logit4,

+ logit4.1, test = "Chisq")$`Pr(>Chi)`[2] #Valore del test e del pvalue

> cbind(testX4, pValueX4)

testX4 pValueX4

[1,] 5.002843 0.02530572

> #Si conclude quindi che è preferibile il modello senza l'interazione, dato che test LR = 5.003 con pvalue = 0.025 (-> rifiuto H0: b interazione = 0)

>

> #Tabella riassuntiva

> cbind(rbind("Nome variabile", "Qualità","Origine","Valore dichiarato"),rbind("b",as.numeric(logit4$coefficients[2]),as.numeric(logit4$coefficients[3]),as.numeric(logit4$coefficients[4])),rbind("exp(b)",as.numeric(exp(logit4$coefficients[2])),as.numeric(exp(logit4$coefficients[3])),as.numeric(exp(logit4$coefficients[4]))))

[,1] [,2] [,3]

[1,] "Nome variabile" "b" "exp(b)"

[2,] "Qualità" "-1.06928182301928" "0.34325494671944"

[3,] "Origine" "-0.110041879273984" "0.895796619198911"

[4,] "Valore dichiarato" "-0.0142719053496638" "0.985829455514321"

> #Analizzando i dati, un consiglio utile agli addetti sarebbe quello di controllare con maggiore attenzionele merci con un alto valore dichiarato, dato che per ogni incremento unitario

> #del valore dichiarato, l'odds ratio stimato aumenta di 0.98, ovvero aumenta la probabilità che Irr = 1 (aumenta il numeratore dell'odds, diminuisce il denominatore)

> #Anche la provenienza del lotto non è irrilevante dato che l'odds ratio aumenta di 0,89 per ogni incremento unitario della variabile ExtraUE

> #In conclusione si consiglia agli addetti di fare maggiore attenzione ai lotti provenienti da paesi non dell'UE e con un alto valore dichiarato

# ESERCIZIO 5 --------------------------

# Introduzione

library(readxl)

#Mi creo due database separati in xlsx in due fogli separati per importare i dati più comodamente

m = read\_xlsx("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 1/ESERCIZIO\ DA\ FARE\ A\ CASA\ Anno\ 1/homework\ ADES\ 2022-2023/dati\_covid.xlsx", sheet = 1)

f = read\_xlsx("/Users/marcocaurla/Documents/Informatica/DSL/R/ANNO\ 1/ESERCIZIO\ DA\ FARE\ A\ CASA\ Anno\ 1/homework\ ADES\ 2022-2023/dati\_covid.xlsx", sheet = 2)

#Creo i vettori per la distribuzione dell'età per il numero dei casi e dei morti di ogni sesso

maschi <- data.frame(

Classe\_di\_eta = c("0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", "80-89", "90 e oltre"),

N\_deceduti = c(7, 15, 61, 222, 948, 3808, 10785, 24300, 31274, 9870),

N\_casi = c(444638, 667351, 697747, 675651, 791996, 778178, 469786, 305004, 163360, 32024),

Letalita = c(0.0016, 0.0022, 0.0087, 0.0329, 0.1197, 0.4893, 2.2957, 7.9671, 19.1442, 30.8206)

)

femmine <- data.frame(

Classe\_di\_eta = c("0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", "80-89", "90 e oltre"),

N\_deceduti = c(10, 13, 34, 128, 429, 1535, 4278, 11978, 26480, 18144),

N\_casi = c(437817, 644080, 749093, 743312, 903908, 827714, 466019, 309141, 229827, 93904),

Letalita = c(0.0023, 0.0020, 0.0045, 0.0172, 0.0475, 0.1855, 0.9180, 3.8746, 11.5217, 19.4900)

)

# I.

#Confronto della letalità tra maschi e femmine con un test t

test\_letalita <- t.test(maschi$Letalita, femmine$Letalita, var.equal = TRUE)

print(test\_letalita)

#Commento: si conclude che p1 è diversa da p2 (al 95%) dato che il test riporta un p-value basso (bassa probabilità di

#trovare un valore più estremo di quello osservato). A supporto di ciò, l'intervallo di confidenza non contiene lo zero,

#quindi p1 - p1 != 0

# II.

#Costruzione dell'intervallo di confidenza per la differenza delle letalità

diff\_letalita <- maschi$Letalita - femmine$Letalita

mean\_diff <- mean(diff\_letalita)

se\_diff <- sd(diff\_letalita) / sqrt(length(diff\_letalita))

intervallo\_confidenza <- mean\_diff + c(-1, 1) \* qt(0.975, df=length(diff\_letalita)-1) \* se\_diff

print(intervallo\_confidenza)

# III.

#Calcolare la media dell'età al decesso per maschi e femmine

# Si presume che l'età media per ogni gruppo di età sia il punto medio dell'intervallo

# Calcolare il punto medio dell'intervallo di età per maschi e femmine

punti\_medi <- c(4.5, 14.5, 24.5, 34.5, 44.5, 54.5, 64.5, 74.5, 84.5, 94.5) # Punto medio per ogni classe di età

maschi$eta\_media\_decesso <- maschi$N\_deceduti \* punti\_medi

femmine$eta\_media\_decesso <- femmine$N\_deceduti \* punti\_medi

# Calcolare la media ponderata dell'età al decesso

media\_eta\_decesso\_maschi <- sum(maschi$eta\_media\_decesso) / sum(maschi$N\_deceduti)

media\_eta\_decesso\_femmine <- sum(femmine$eta\_media\_decesso) / sum(femmine$N\_deceduti)

print(media\_eta\_decesso\_maschi)

print(media\_eta\_decesso\_femmine)

# IV.

# Introduzione delle librerie necessarie

library(DescTools)

# Il fenomeno del decesso è, a livello di popolazione, lo stesso nelle diverse classi di età?

# Creiamo una tabella di contingenza per l'età al decesso per maschi e femmine

TabMorti <- table(maschi$Classe\_di\_eta, maschi$N\_deceduti) + table(femmine$Classe\_di\_eta, femmine$N\_deceduti)

# Eseguiamo Test Esatto di Fisher per l'indipendenza

test\_morti = fisher.test(TabMorti)

print(test\_morti)

#Definisco la funzione vCramer e prevede che tu passi solo la tabella di contingenza come argomento, e calcola il numero totale di osservazioni all'interno della funzione stessa,

vCramer <- function(TabellaDiContingenza) {

test <- chisq.test(TabellaDiContingenza)

chiquadro <- as.numeric(test$statistic)

n <- sum(TabellaDiContingenza) # Numero totale di osservazioni

minDim <- min(nrow(TabellaDiContingenza)-1, ncol(TabellaDiContingenza)-1)

v <- sqrt(chiquadro / (n\*minDim))

return(v)

}

# Calcolo dell'indice di Cramer

vCramer\_morti <- vCramer(TabMorti)

print(vCramer\_morti)

# Output dei risultati

print(test\_morti)

print(vCramer\_morti)

print(CramerV(TabMorti)) # Conferma con il pacchetto DescTools

# Commento: il test eseguito porta a rifiutare l'ipotesi nulla: non c'è indipendenza

# Questo lo si evince dal valore del p-value (minore del livello di significatività) e dall'indice di Cramer

#V.

# Supponendo che N\_deceduti\_0\_9\_maschi e N\_deceduti\_0\_9\_femmine siano il numero di decessi

# e che N\_sopravvissuti\_0\_9\_maschi e N\_sopravvissuti\_0\_9\_femmine siano il numero di sopravvissuti nella fascia di età "0-9"

# Dati fittizi per esemplificare

N\_deceduti\_0\_9\_maschi <- 7

N\_deceduti\_0\_9\_femmine <- 10

N\_sopravvissuti\_0\_9\_maschi <- 444638 - 7 # Numero totale di casi maschi - numero di decessi

N\_sopravvissuti\_0\_9\_femmine <- 437817 - 10 # Numero totale di casi femmine - numero di decessi

# Creazione della tabella di contingenza 2x2

TabMortiPrimaEta <- matrix(c(N\_deceduti\_0\_9\_maschi, N\_sopravvissuti\_0\_9\_maschi,

N\_deceduti\_0\_9\_femmine, N\_sopravvissuti\_0\_9\_femmine),

nrow = 2,

dimnames = list(c("Maschi", "Femmine"), c("Decessi", "Sopravvissuti")))

# Esecuzione del Test Esatto di Fisher

test\_morti\_prima\_eta <- fisher.test(TabMortiPrimaEta)

print(test\_morti\_prima\_eta)

#Commento: il test eseguito porta a rifiutare l'ipotesi nulla: non c'è indipendenza

#Questo lo si evince dal bassissimo valore del p-value (ca. 0) e dall'elevato valore della statistica test X-sqared.