Università degli Studi di Trento
Dipartimento di Economia e Management
Corso di Economia e Management
Anno Accademico 2022-2023

Statistica, probabilità e inferenza

Docente: Espa Giuseppe, Santi Flavio

Homework

De Massari Thomas – numero di matricola: 226091

Indice

Esercizio 1 3
Esercizio 2 6
Esercizio 3 12
Esercizio 4 21
Esercizio 5 24

Esercizio 1

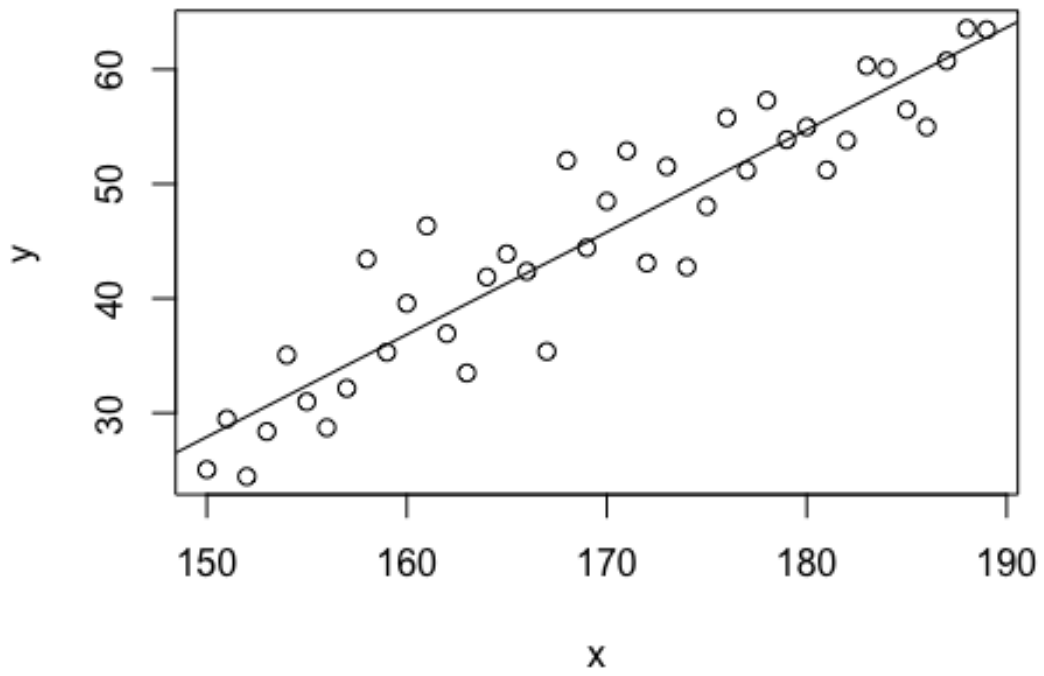
```
# Introduzione
x = c(150:189) #crea un vettore x contenente i numeri 150,151,...,188,189
alpha = -90; beta = 0.8 #definizione di alpha e beta
ym = alpha + (beta*x) #vettore contenente i valori alpha+(beta*x)

set.seed(226091) #innesco
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
```

```
## [1] 0.9318666
```

```
plot(x, y) #scatterplot

# Modello di regressione lineare y = a +b(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)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min       1Q     Median     3Q     Max
## -7.7342 -2.7321 -0.1133  2.6949  8.5887
##
## Coefficients:
##                  Estimate     Std. Error   t value  Pr(>|t|)
## (Intercept) -106.16199    9.59096  -11.07 1.88e-13 ***
## x                   0.89385    0.05645     15.83  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.121 on 38 degrees of freedom
## Multiple R-squared:  0.8684, Adjusted R-squared:  0.8649
## F-statistic: 250.7 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
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
cbind(TSS,SSR,SSE,s,s2,R2)
```

```
##         TSS             SSR         SSE         s             s2         R2
## [1,] 4903.977 4258.492 645.4846 4.121461 16.98644 0.8683753
```

```
anova(modello)
```

```
## Analysis of Variance Table
##
## Response: y
##               Df  Sum Sq  Mean Sq  F value    Pr(>F)
## x              1  4258.5   4258.5     250.7   < 2.2e-16 ***
## Residuals  38   645.5     17.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
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
```

```
## [1] "0.779566093932779"  < b < "1.00813261488655"
```

```r
# 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)
```

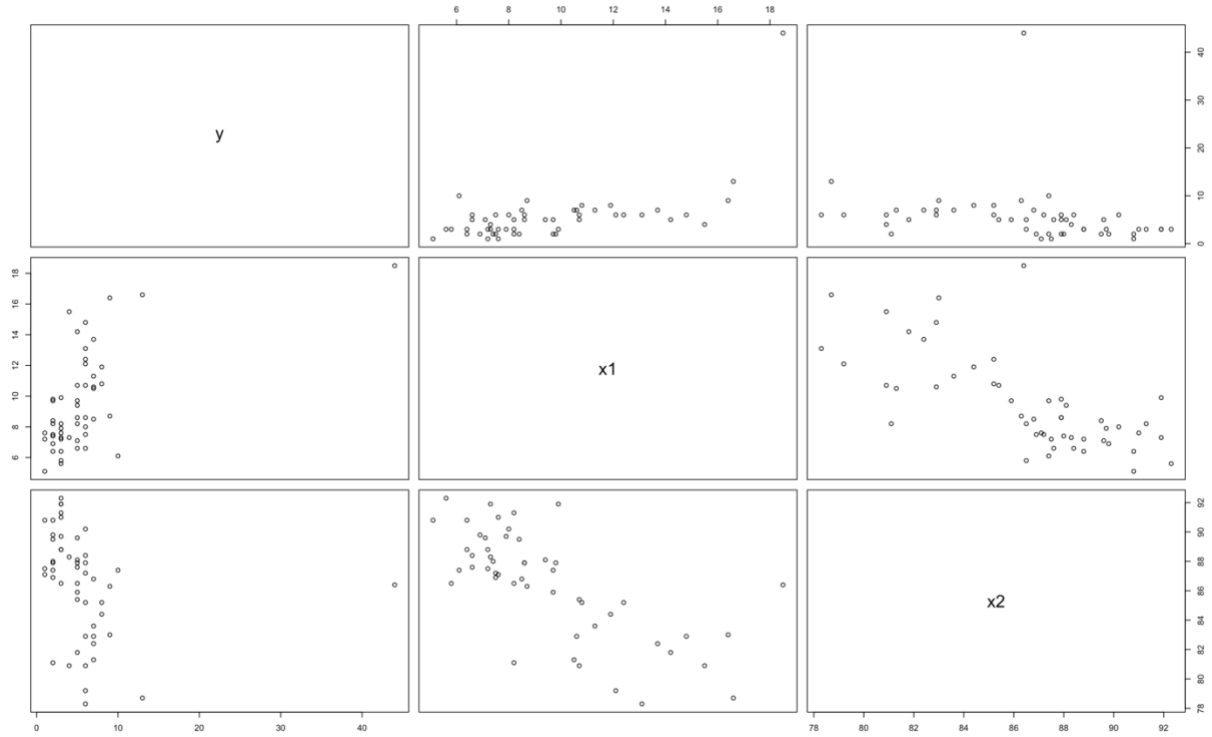
```
##
## Call:
## lm(formula = ycen ~ xcen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7342 -2.7321 -0.1133  2.6949  8.5887
##
## Coefficients:
##              Estimate.   Std. Error    t value   Pr(>|t|)
## (Intercept) 2.868e-15   6.517e-01    0.00       1
## xcen          8.938e-01  5.645e-02   15.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.121 on 38 degrees of freedom
## Multiple R-squared:  0.8684, Adjusted R-squared:  0.8649
## F-statistic: 250.7 on 1 and 38 DF,  p-value: < 2.2e-16
```

```r
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
```

```
##                    [,1]          [,2]
## alpha1 -106.1619945 2.868295e-15
## beta1     0.8938494     8.938494e-01
```

Esercizio 2

```
# Introduzione
df2 = read.csv("/Users/thomasdemassari/Documents/Università/Statistica, probabilità e inferenza/Homework/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 (diplomati all'high school)
lmy = lm(y ~ x2)
lmx1 = lm(x1 ~ x2)
PR.x2 = lm(lmy$res ~ lmx1$res); summary(PR.x2) #regressione parziale controllando per tasso di scolarizzazione
```

```
##
## Call:
## lm(formula = lmy$res ~ lmx1$res)
##
## 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) 1.121e-16    6.602e-01    0.000       1
## lmx1$res    1.605e+00   2.979e-01   5.388  2.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.715 on 49 degrees of freedom
## Multiple R-squared:  0.372,  Adjusted R-squared:  0.3592
## F-statistic: 29.03 on 1 and 49 DF,  p-value: 2.023e-06
```

```
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 (tasso di povertà)
lmyy = lm(y ~ x1)
lmx2 = lm(x2 ~ x1)
```

```
PR.x1 = lm(lmyy$res ~ lmx2$res); summary(PR.x1) #regressione parziale controllando per il tasso di povertà
```

```
##
## Call:
## lm(formula = lmyy$res ~ lmx2$res)
##
## 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) 6.218e-17  6.602e-01   0.000    1.000
## lmx2$res    5.877e-01  2.578e-01   2.279    0.027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.715 on 49 degrees of freedom
## Multiple R-squared:  0.09588,    Adjusted R-squared:  0.07742
## F-statistic: 5.196 on 1 and 49 DF,  p-value: 0.02703
```

```
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
par(mfrow = c(1,1))
#Commento: il grafico a sinistra rappresenta la relazione tra il tasso di omicidi e il tasso di povertà escludendo 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 sia il tasso di povertà, che il tasso di scolarizzazione, sono correlati positivamente con il tasso di omicidi
#Inoltre, in entrambi i casi emerge come il Distretto della Columbia (in rosso) sia un outlier
```

```
#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
```

```
#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 di un 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)
```

```
##
## 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
## ---
```

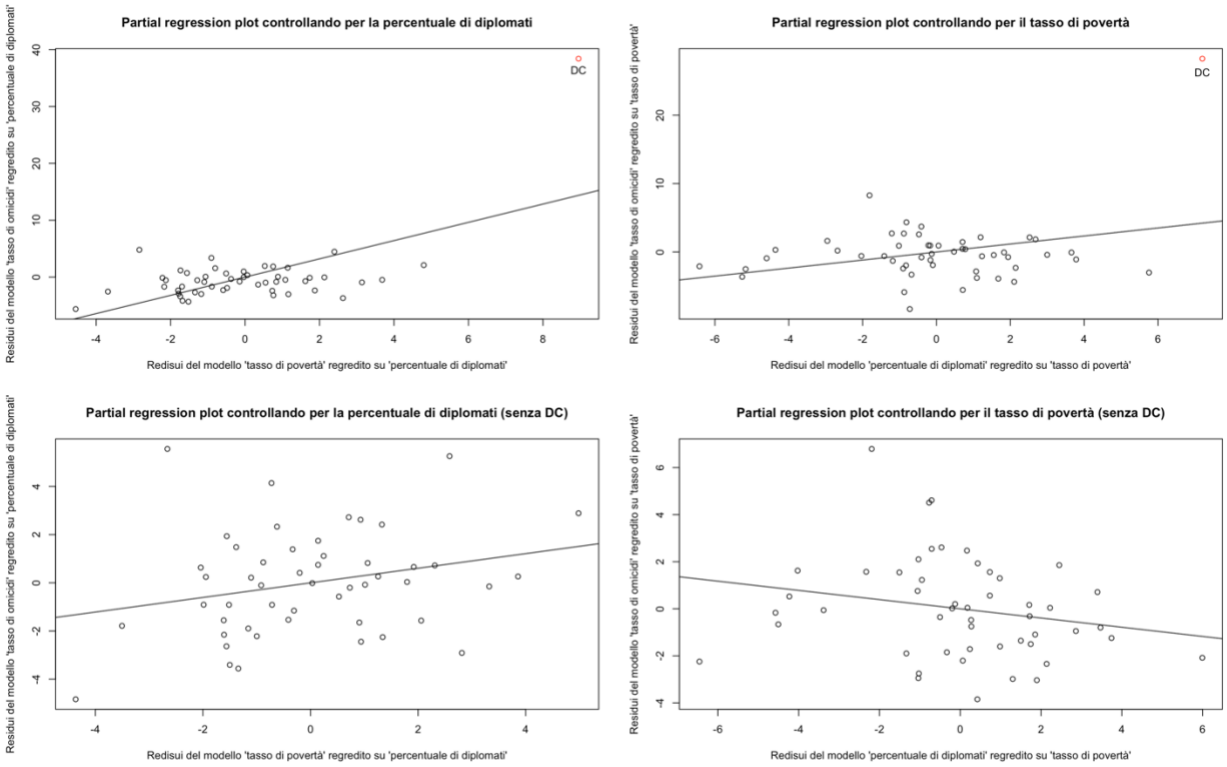
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.136 on 47 degrees of freedom
## Multiple R-squared:  0.3382, Adjusted R-squared:   0.31
## F-statistic: 12.01 on 2 and 47 DF,  p-value: 6.123e-05
```

```
#Analisi del modello di regressione multipla (senza DC)
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(round(alphaNoDC, 4), round(beta1NoDC, 4), round(beta2NoDC, 4)))
```

```
##              [,1]       [,2]                                  [,3]
## [1,] "Intercetta" "b - tasso di povertà senza DC"  "b - tasso di diplomati senza DC"
## [2,] "18.9123"    "0.3035"                          "-0.196"
```

```
#Si nota come cambiano sensibilmente i parametri a e b(i) stimati. Ciò vuol dire che il dato DC è un
outlier
#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
#D'altro canto, osservando il p-value, si può notare come i coefficienti non siano significati (non ci sono
evidenze statistiche contro H0: b = 0).

# Riassunto dei quattro partial regression plot
par(mfrow = c(2,2))
#Partial regression plot con DC
plot(lmx1$res, lmy$res, xlab = "Residui 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 = "Residui 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 = "Residui del modello 'tasso di povertà' regredito su
```

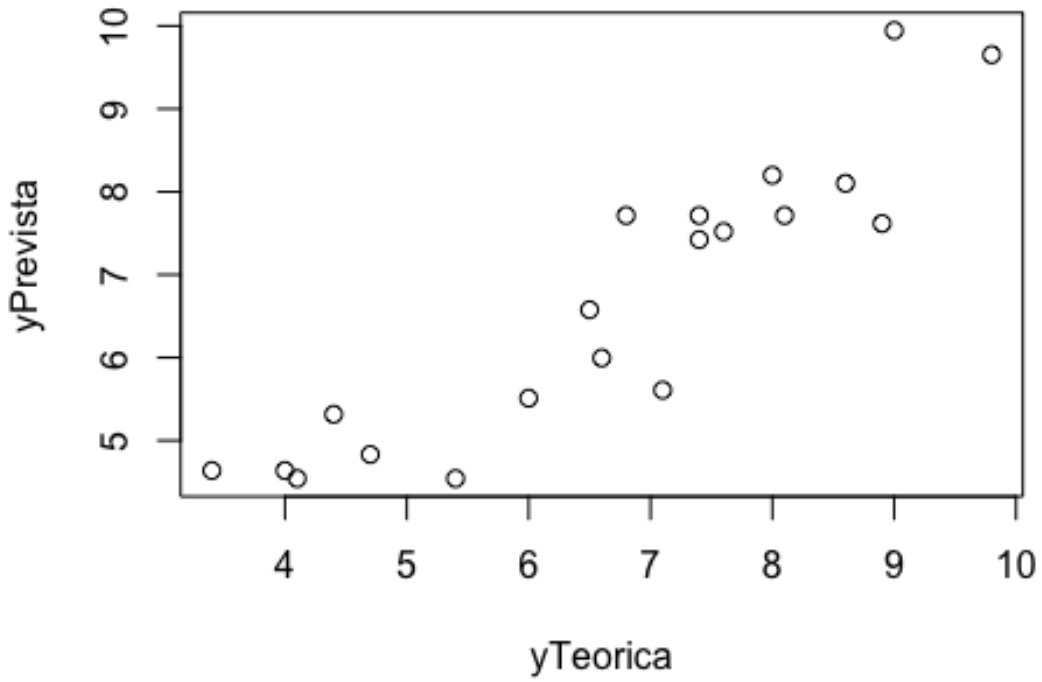
```
'percentuale di diplomati'",
    ylab = "Residui del modello 'tasso di omicidi' regredito su 'percentuale di diplomati'",
    main = "Partial regression plot controllando per la percentuale di diplomati (senza DC)")
abline(lm(lmyNoDC$res ~ lmx1NoDC$res)) #aggiungo la retta di regressione

lmyyNoDC = lm(yNoDC ~ x1NoDC)
lmx2NoDC = lm(x2NoDC ~ x1NoDC)
plot(lmx2NoDC$res, lmyyNoDC$res, xlab = "Residui 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 inizialmente (includendo DC), mentre la relazione tra tasso di omicidi e percentuali di
diplomati all'high school è invertita di segno
#Questo è coerente con quanto dimostrato precedentemente, ovvero che all'aumentare del tasso di
scolarizzazione cala il tasso di omicidi
```

Esercizio 3

```
# Svolgimento esercizio 30
df = read.csv("/Users/thomasdemassari/Documents/Università/Statistica, probabilità e inferenza/Homework/lavoro.csv", header = TRUE)

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

## [1] 0.9162022
```



#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]))

[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)$coefficients[8]); (testAnnilavoro = summary(fit30)$coefficients[9])

[1] -7.185669

[1] 4.350611

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

```
## [1] 2.109816
```

```
(pvalueSex = summary(fit30)$coefficients[11]); (pvalueAnniLavoro = summary(fit30)$coefficients[12])
```

```
## [1] 1.524473e-06
```

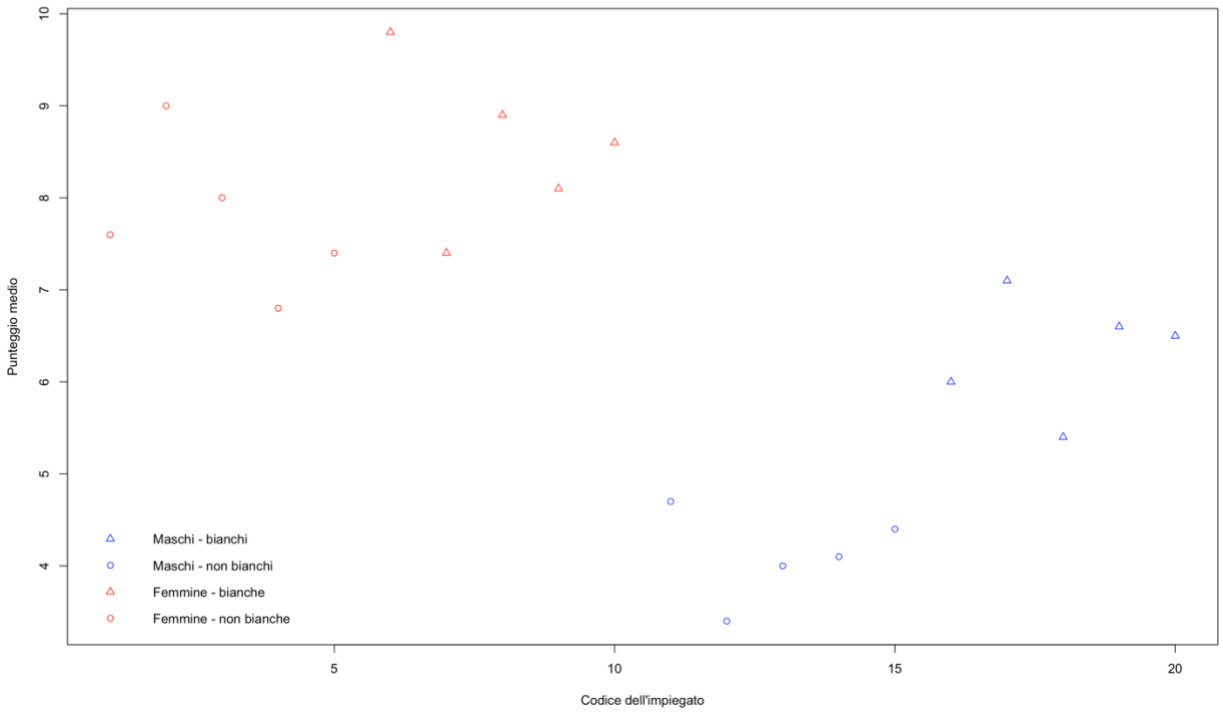
```
## [1] 0.0004349557
```

```
#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
```

```
##
## Call:
## lm(formula = avgPunteggio ~ sex + AnniLavoro + race)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.76326 -0.29823 -0.03107  0.25044  0.96493
##
## Coefficients:
##                  Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  6.64754    0.23099     28.779   3.30e-15 ***
## sexMale     -2.65702    0.22114    -12.015   2.02e-09 ***
## AnniLavoro   0.07860    0.01406     5.591    4.06e-05 ***
## raceWhite    1.20130    0.22180     5.416    5.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 16 degrees of freedom
## Multiple R-squared:  0.9433, Adjusted R-squared:  0.9327
## F-statistic: 88.77 on 3 and 16 DF,  p-value: 3.462e-10
```

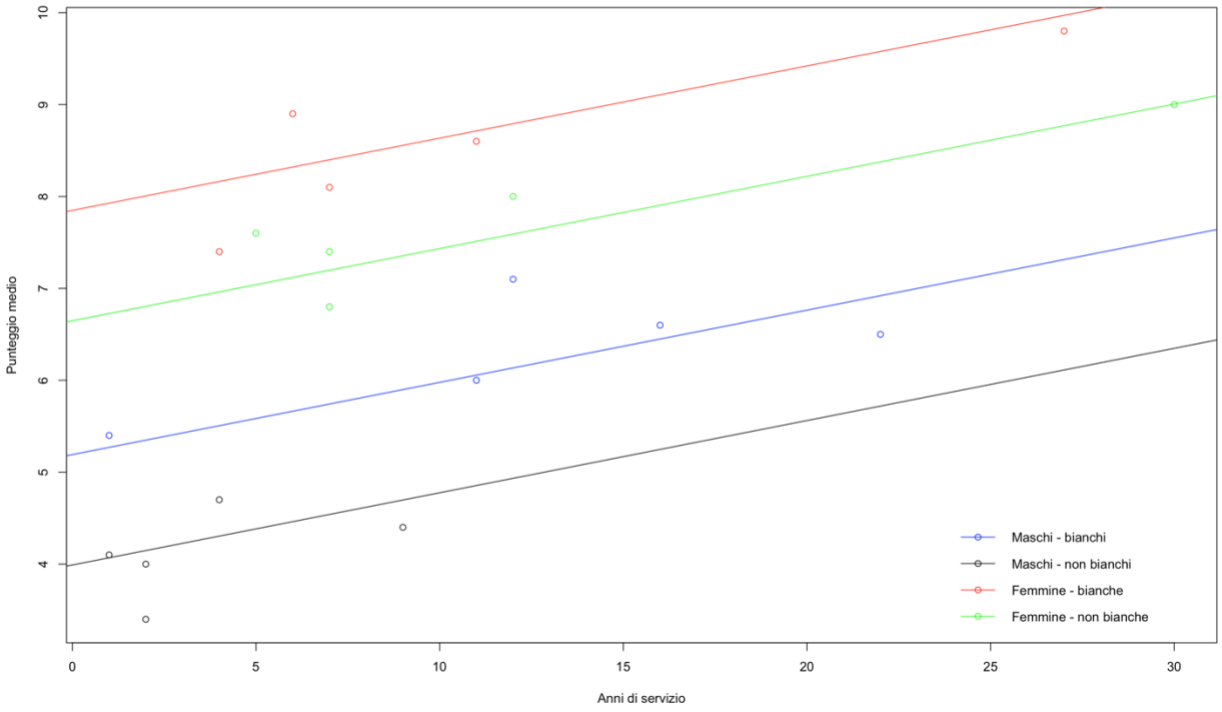
```
#Rappresentazione grafica con scatter plot
#Gestione dei colori: le donne sono in rosa, gli uomini in blu
colors = rep(NA, length(df$Employee)) #vettore dei colori
counterS = 1 #contatore per ciclo while
#ciclo while per assegnare i colori
while(counterS <= length(df$Employee)){
  if(sex[counterS] == "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), bty = "n")
```



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

```r
#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

#Rappresentazione grafica con includendo anche gli anni di servizio
#Coloro i punti
#Masch bianchi: blu; maschi non bianchi: neri; femmine bianche = rosso; femmine non bianche: nere
pointAS = rep(NA, length(df$Employee)) #vettore dei colori
counterAS = 1 #contatore per ciclo while
#ciclo while per assegnare i colori
while(counterAS <= length(df$Employee)){
 if(sex[counterAS] == "Male" & race[counterAS] == "White"){
  pointAS[counterAS] = "blue"
  counterAS = counterAS + 1
 } else {
  if(sex[counterAS] == "Male" & race[counterAS] == "Nonwhite"){
   pointAS[counterAS] = "black"
   counterAS = counterAS + 1
  } else {
   if(sex[counterAS] == "Female" & race[counterAS] == "White"){
    pointAS[counterAS] = "red"
    counterAS = counterAS + 1
   } else {
    if(sex[counterAS] == "Female" & race[counterAS] == "Nonwhite"){
     pointAS[counterAS] = "green"
     counterAS = counterAS + 1
    } else {
     pointAS[counterAS] = NA
     counterAS = counterAS + 1
    }
   }
  }
 }
}
plot(AnniLavoro, avgPunteggio, xlab="Anni di servizio", ylab="Punteggio medio", col = pointAS)
#Salvo i coefficienti in variabili per comodità
a = as.numeric(fit3$coeff[1]); bSex = as.numeric(fit3$coeff[2]); bAS = as.numeric(fit3$coeff[3]); bRace =
as.numeric(fit3$coeff[4])
#Aggiungo le rette di regressione parziale, controllando per sesso e razza
maschio = bSex*1; femmina = bSex*0; bianco = bRace*1; nonbianco = bRace*0
abline(a+maschio+bianco, bAS, col = "blue") #maschi bianchi
abline(a+maschio+nonbianco, bAS, col = "black") #maschi non bianchi
abline(a+femmina+bianco, bAS, col = "red") #femmine bianche
abline(a+femmina+nonbianco, bAS, col = "green") #femmine non bianche
legend("bottomright", legend = c("Maschi - bianchi", "Maschi - non bianchi", "Femmine - bianche",
"Femmine - non bianche"), col = c("blue", "black", "red", "green"), bty = "n", lwd = 1, pch = 1)
```



#Si nota che la relazione tra anni di servizio e punteggio medio sia crescente.
#Inoltre, si confermano le intuizioni di prima, ovvero che i maschi bianchi hanno un punteggio superiore a quello dei maschi non bianchi e che le donne registrano, per ogni livello di razza, un punteggio medio superiore.
#Da questo grafico è anche possibile trarre delle conclusioni sul gruppo femminile: le donne bianche hanno un punteggio medio maggiore di quello di tutte le altre categorie

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(fit3, fit3.1)
```

```
## Analysis of Variance Table
##
## Model 1: avgPunteggio ~ sex + AnniLavoro + race
## Model 2: avgPunteggio ~ sex + AnniLavoro + race + AnniLavoro * sex + AnniLavoro *
##     race
##   Res.Df    RSS      Df   Sum of Sq      F      Pr(>F)
## 1     16    3.7074
## 2     14.    3.6509  2     0.056448 0.1082 0.8982
```

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

```
##        testX3        pValueX3
## [1,] 0.1082279 0.8981665
```

```
#Non ci sono abbastanza evidenze statistiche per rifiutare H0: b interazione annilavoro*sex = 0
#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
```

```
##            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 significativa 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 significativa al 5%)")
}
```

```
## [1] "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 significativa 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 significativa 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 significativa al 5%)")
}
```

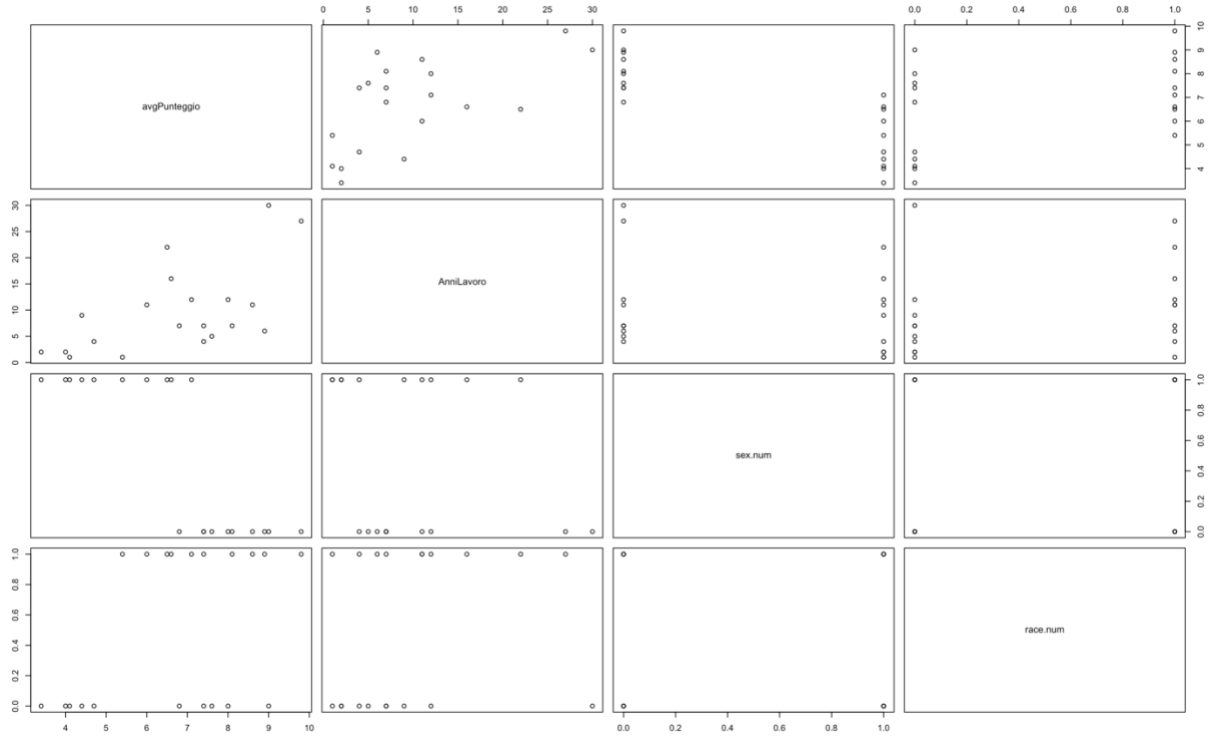
```
## [1] "Risultato test su b di anni di lavoro-sesso: non ci sono abbastanza evidenze statistiche per rifiutare H0: bAR = 0, quindi l'interpolazione tra anni di lavoro e razza non è statisticamente significativa (livello di significativa al 5%)"
```

```
summary(fit3) #Modello senza le due interpolazioni non significative
```

```
##
## Call:
## lm(formula = avgPunteggio ~ sex + AnniLavoro + race)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.76326 -0.29823 -0.03107  0.25044  0.96493
##
## Coefficients:
##                 Estimate   Std. Error t value   Pr(>|t|)
## (Intercept)  6.64754    0.23099    28.779   3.30e-15 ***
## sexMale     -2.65702    0.22114   -12.015   2.02e-09 ***
## AnniLavoro   0.07860    0.01406   5.591    4.06e-05 ***
## raceWhite    1.20130    0.22180   5.416    5.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 16 degrees of freedom
## Multiple R-squared:  0.9433, Adjusted R-squared:  0.9327
## F-statistic: 88.77 on 3 and 16 DF,  p-value: 3.462e-10
```

```
# Grafici di diagnostica del modello
#imposto per il sesso male = 1 e female = 0
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 = 0
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
```

#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

Esercizio 4

```
# Introduzione
#Obiettivo: verificare se il valore della merce consegnata corrisponda effettivamente a quello dichiarato
#$prima = merce di prima categoria (1) o di seconda categoria (0)
#$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/thomasdemassari/Documents/Università/Statistica, probabilità e inferenza/Homework/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(esitoControllo = 1 | qualità = 1, origine = 0, valoreD = 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 verosimiglianze
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 con 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","Interazione qualità*origine"),

rbind("b",as.numeric(logit4.1$coefficients[2]),as.numeric(logit4.1$coefficients[3]),as.numeric(logit4.1$coefficients[4]),as.numeric(logit4.1$coefficients[5])),

rbind("exp(b)",as.numeric(exp(logit4.1$coefficients[2])),as.numeric(exp(logit4.1$coefficients[3])),as.numeric(exp(logit4.1$coefficients[4])),as.numeric(exp(logit4.1$coefficients[5]))),
    rbind("p-value",as.numeric(summary(logit4.1)$coefficients[17]),as.numeric(summary(logit4.1)$coefficients[18]),as.numeric(summary(logit4.1)$coefficients[19]),as.numeric(summary(logit4.1)$coefficients[20])))
```

```
##      [,1]                          [,2]                  [,3]                  [,4]
## [1,] "Nome variabile"              "b"                   "exp(b)"              "p-value"
## [2,] "Qualità"                    "-1.85416948114116"   "0.156582933766922" "0.00235847748247447"
## [3,] "Origine"                    "-0.172665002397984"  "0.841419441131491" "0.205766067398754"
## [4,] "Valore dichiarato"          "-0.0142810041820008" "0.9858204856582"   "4.62759397075101e-21"
## [5,] "Interazione qualità*origine" "1.49909464004373"    "4.47763336473107" "0.0340587806319537"
```

```
#Analizzando i dati, un consiglio utile agli addetti sarebbe quello di controllare con maggiore attenzione le 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)
#Non è significativa invece la provenienza del lotto, dato il pvalue > 0.05 che porta a non rifiutare H0: b-origine = 0.
#In secondo luogo gli addetti dovrebbero controllare anche la qualità della merce di prima qualità, dato che per ogni incremento unitario della variabile "prima" (quindi passa da 0 a 1) l'odds ratio aumenta di 0.34.
```

```
#Inoltre, dall'analisi di b per l'interpolazione qualità*origine (b = 1.499) si può concludere che se si controlla per l'origine e la qualità contemporaneamente
#la probabilità che il valore si discosti da quello dichiarato è quasi del 50%.
#In conclusione si consiglia agli addetti di fare maggiore attenzione ai lotti con alto valore dichiarato e quelli di prima qualità
```

Esercizio 5

```
# Introduzione
#Mi creo due database in csv per importare i dati
m = read.csv("/Users/thomasdemassari/Documents/Università/Statistica, probabilità e inferenza/Homework/COVID - maschi.csv")
f = read.csv("/Users/thomasdemassari/Documents/Università/Statistica, probabilità e inferenza/Homework/COVID - femmine.csv")

#Creo i vettori per la distribuzione dell'età per il numero dei casi e dei morti di ogni sesso
i=2; sex=m;
mMorti =
as.numeric(c(rep(sex[1,1],sex[1,i]),rep(sex[2,1],sex[2,i]),rep(sex[3,1],sex[3,i]),rep(sex[4,1],sex[4,i]),rep(sex[5,1],sex[5,i]),

rep(sex[6,1],sex[6,i]),rep(sex[7,1],sex[7,i]),rep(sex[8,1],sex[8,i]),rep(sex[9,1],sex[9,i]),rep(sex[10,1],sex[10,i])))
sex=f
fMorti =
as.numeric(c(rep(sex[1,1],sex[1,i]),rep(sex[2,1],sex[2,i]),rep(sex[3,1],sex[3,i]),rep(sex[4,1],sex[4,i]),rep(sex[5,1],sex[5,i]),

rep(sex[6,1],sex[6,i]),rep(sex[7,1],sex[7,i]),rep(sex[8,1],sex[8,i]),rep(sex[9,1],sex[9,i]),rep(sex[10,1],sex[10,i])))
sex=m
i = 3
mCasi =
as.numeric(c(rep(sex[1,1],sex[1,i]),rep(sex[2,1],sex[2,i]),rep(sex[3,1],sex[3,i]),rep(sex[4,1],sex[4,i]),rep(sex[5,1],sex[5,i]),

rep(sex[6,1],sex[6,i]),rep(sex[7,1],sex[7,i]),rep(sex[8,1],sex[8,i]),rep(sex[9,1],sex[9,i]),rep(sex[10,1],sex[10,i])))
sex=f
fCasi =
as.numeric(c(rep(sex[1,1],sex[1,i]),rep(sex[2,1],sex[2,i]),rep(sex[3,1],sex[3,i]),rep(sex[4,1],sex[4,i]),rep(sex[5,1],sex[5,i]),

rep(sex[6,1],sex[6,i]),rep(sex[7,1],sex[7,i]),rep(sex[8,1],sex[8,i]),rep(sex[9,1],sex[9,i]),rep(sex[10,1],sex[10,i])))
RiassuntoFemmine = cbind(table(fMorti),table(fCasi)); colnames(RiassuntoFemmine) <- c("Morti","Casi"); RiassuntoFemmine
```

```
##    Morti   Casi
## 5     10 413787
## 15    13 644080
## 25    34 694093
## 35   128 743312
## 45   429 903908
## 55  1535 827714
## 65  4278 466019
## 75 11978 309141
## 85 26480 229827
## 95 18144  93094
```

```
RiassuntoMaschi = cbind(table(mMorti),table(mCasi)); colnames(RiassuntoMaschi) <- c("Morti","Casi");
RiassuntoMaschi
```

```
##    Morti   Casi
## 5      7    444638
## 15    15  667351
## 25    61  697747
## 35   222  675651
## 45   948  791996
## 55  3808  778178
## 65 10785 469786
## 75 24300 305004
## 85 31274 163360
## 95  9870  32024
```

```
# Test sulla letalità maschi (1) - femmine (2)
#H0: p1 - p2 = 0; alpha = 0.05
p1 = m[11,2]; p2 = f[11,2] #successi maschi e femmine
n1 = m[11,3]; n2 = f[11,3] #popolazione maschi e femmine
#se = sqrt(((p1*(1-p1))/n1)+((p2*(1-p2))/n2))
(test = prop.test(c(p1,p2),c(n1,n2),correct=FALSE, conf = 0.95)) #test sotto H0: p1-p2 = 0
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(p1, p2) out of c(n1, n2)
## X-squared = 3539.3, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.004194730 0.004481792
## sample estimates:
##     prop 1      prop 2
## 0.01617475 0.01183649
```

```
test$conf.int #Intervallo di confidenza al 95%
```

```
## [1] 0.004194730 0.004481792
## attr(,"conf.level")
## [1] 0.95
```

#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

```
# Test sull'età media al contagio
#Calcolo media dell'età al contagio e della sua varianza
mimc = mean(mCasi); mifc = mean(fCasi)
t.test(mCasi, fCasi) #test t con H0: mimc = mifc
```

```
##
##  Welch Two Sample t-test
##
## data:  mCasi and fCasi
## t = -122.56, df = 10338159, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.720456 -1.666294
## sample estimates:
## mean of x   mean of y
##  40.21378   41.90715
```

```
t.test(mCasi, fCasi)$conf.int #intervallo di confidenza al 95%
```

```
## [1] -1.720456 -1.666294
## attr(,"conf.level")
## [1] 0.95
```

#Commento: il test eseguito porta a rifiutare l'ipotesi nulla mimc = mifc con una confidenza del 95%. #Questo lo si evince dal bassissimo valore del p-value (0) e dal fatto che l'intervallo di confidenza non contiene lo zero.

```
# Test sull'età media al decesso
#Calcolo media dell'età al contagio e della sua varianza
mimm = mean(mMorti); mifm = mean(fMorti)
t.test(mMorti, fMorti) #test t con H0: mimm = mifm
```

```
##
##  Welch Two Sample t-test
##
## data:  mMorti and fMorti
## t = -85.692, df = 138083, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.072101 -4.845268
## sample estimates:
```

```
## mean of x   mean of y
##  78.49822  83.45690
```

```
t.test(mMorti, fMorti)$conf.int #intervallo di confidenza al 95%
```

```
## [1] -5.072101 -4.845268
## attr(,"conf.level")
## [1] 0.95
```

```
#Commento: il test eseguito porta a rifiutare l'ipotesi nulla mimm = mifm con una confidenza del 95%.
#Questo lo si evince dal bassissimo valore del p-value (ca. 0) e dal fatto che l'intervallo di confidenza non contiene lo zero.
```

```
# Analisi per sesso dell'età al contagio
#Funzione per calcolare l'indice di Cramer
vCramer = function(TabellaDiContingenza, NumeroDiOsservazioni){
 test = chisq.test(TabellaDiContingenza)
 chiquadro = as.numeric(test$statistic)
 riga = nrow(TabellaDiContingenza)-1
 colonna = ncol(TabellaDiContingenza)-1
 numeratore = (chiquadro)/NumeroDiOsservazioni
 denominatore = min(riga,colonna)
 sqrt(numeratore/denominatore)
}
```

```
#Il fenomeno del contagio è, a livello di popolazione, lo stesso nelle diverse classi di età?
TabContagi = rbind(table(mCasi),table(fCasi)); rownames(TabContagi) <- c("Maschi","Femmine");
TabContagi
```

```
##                5         15      25         35       45       55       65       75       85       95
## Maschi  444638 667351 697747 675651 791996 778178 469786 305004 163360 32024
## Femmine 413787 644080 694093 743312 903908 827714 466019 309141 229827 93094
```

```
chisq.test(TabContagi) #test H0: c'è indipendenza
```

```
##
##  Pearson's Chi-squared test
##
## data:  TabContagi
## X-squared = 46145, df = 9, p-value < 2.2e-16
```

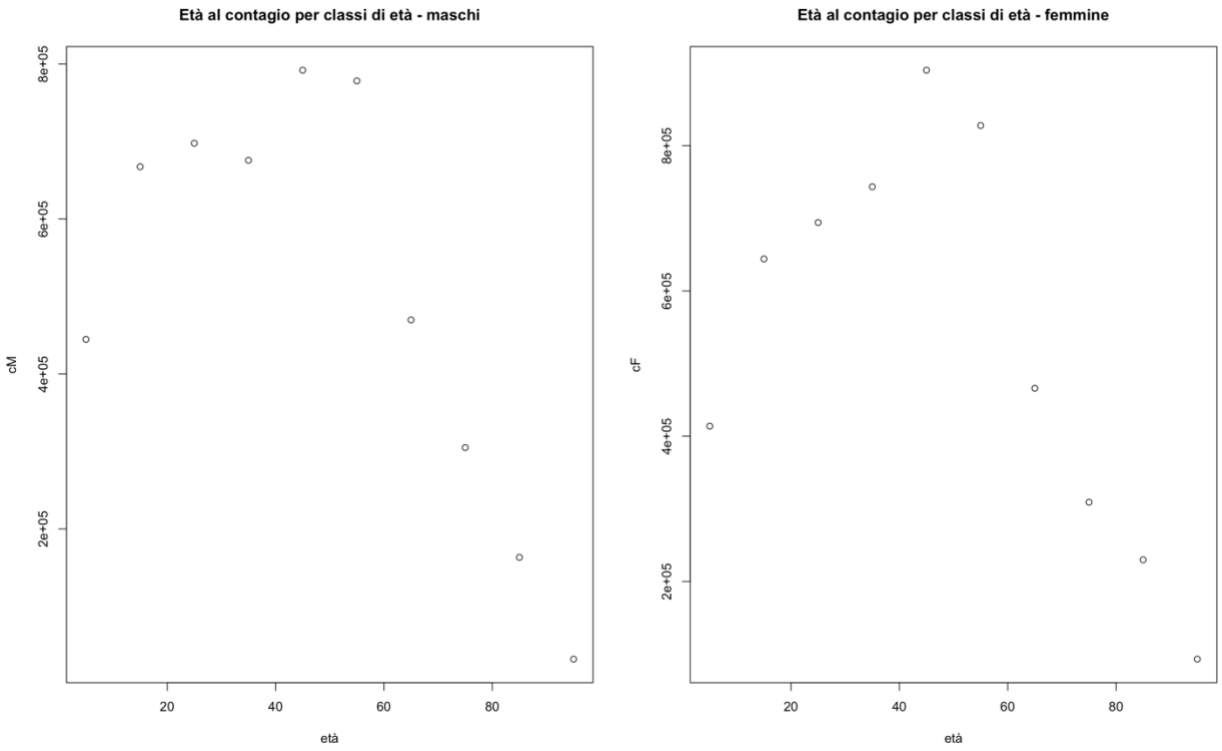
```
vCramer(TabContagi,sum(m[11,3],f[11,3])) #calcolo l'indice di Cramer con la mia funzione
```

```
## [1] 0.06676967
```

```
library(DescTools); CramerV(TabContagi) #calcolo dell'indice di Cramer con il pacchetto DescTools (per verificare i risultati)
```

```
## [1] 0.06676967
```

```
#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.
cM = as.numeric(TabContagi[1,]); cF = as.numeric(TabContagi[2,]); età = c(5, 15, 25, 35, 45, 55, 65, 75, 85, 95)
par(mfrow = c(1,2))
plot(età, cM, main = "Età al contagio per classi di età - maschi"); plot (età, cF, main = "Età al contagio per classi di età - femmine")
```



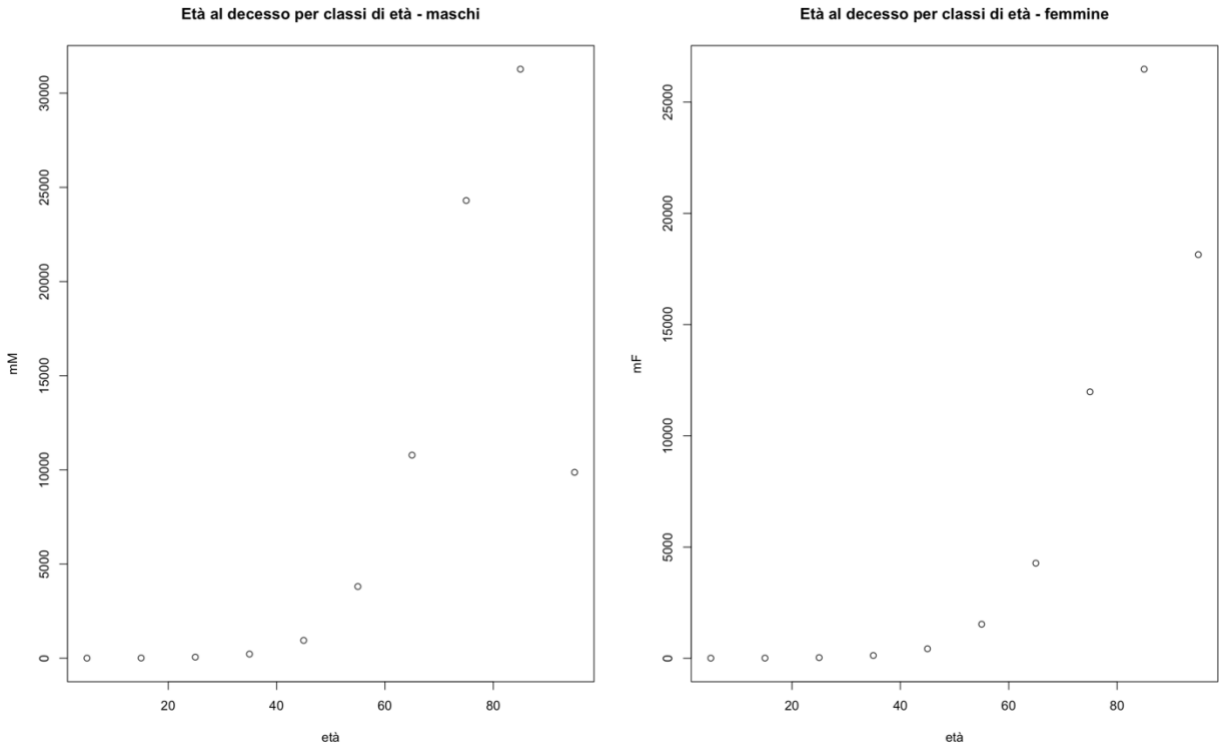

```
par(mfrow = c(1,1))
#Si può concludere quindi che l'età al contagio è maggiore per le classi di età più giovani  (per entrambi i sessi)

# Analisi per sesso dell'età al decesso
#Il fenomeno del decesso è, a livello di popolazione, lo stesso nelle diverse classi di età?
TabMorti = rbind(table(mMorti),table(fMorti)); rownames(TabMorti) <- c("Maschi","Femmine");
TabMorti
```

```
##              5 15 25   35  45    55      65      75     85       95
## Maschi    7 15 61 222 948 3808 10785 24300 31274  9870
## Femmine 10 13 34 128 429 1535  4278 11978 26480 18144
```

```
chisq.test(TabMorti) #test H0: c'è indipendenza
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  TabMorti
## X-squared = 8865.4, df = 9, p-value < 2.2e-16

vCramer(TabMorti, sum(m[11,2],f[11,2])) #calcolo l'indice di Cramer con la mia funzione

## [1] 0.2478485

CramerV(TabMorti) #calcolo dell'indice di Cramer con il pacchetto DescTools per verificare i risultati

## [1] 0.2478485

#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-squared.
mM = as.numeric(TabMorti[1,]); mF = as.numeric(TabMorti[2,])
par(mfrow = c(1,2))
plot(età, mM, main = "Età al decesso per classi di età - maschi"); plot (età, mF, main = "Età al decesso per classi di età - femmine")
```




```
par(mfrow = c(1,1))
#Si può concludere quindi che il numero di morti aumenta all'aumentare dell'età  (per entrambi i sessi)
```