```
######################## Homework 3 ####################################
# Nome:
# Cognome:
# Matricola:
# CdL:
########################################################################
####### 1 #######

#install.packages("alr3")
library("alr3")

mydf <- BigMac2003

####### 2 #######

?BigMac2003

####### 3 #######

# PREMESSA:
# Si sta valutando il potere di acquisto in termini di minuti
# lavorati per acquistare un bene X presente in tutto il mondo
# BigMac dovrebbe essere un noto indice riferito alla c.d. "Burger Economy"
# utilizzato per comparare il potere di acquisto in vari paesi


#2 Bread - Minutes of labor to purchase 1 kg of bread

colnames(mydf)[colnames(mydf)=="Bread"] <- "CostoPane"

#3 Rice Minutes of labor to purchase 1 kg of rice

colnames(mydf)[colnames(mydf)=="Rice"] <- "CostoRiso"

#5 Bus - Cost in US dollars for a one-way 10 km ticket

colnames(mydf)[colnames(mydf)=="Bus"] <- "CostoBigliettoBus"

#6 Apt Normal rent (US dollars) of a 3 room apartment

colnames(mydf)[colnames(mydf)=="Apt"] <- "Affitto"

#7 TeachGI Primary teacher's gross income, 1000s of US dollars

colnames(mydf)[colnames(mydf)=="TeachGI"] <- "InsegnanteRLordo"

#8 TeachNI Primary teacher's net income, 1000s of US dollars

colnames(mydf)[colnames(mydf)=="TeachNI"] <- "InsegnanteRNetto"

#10 TeachHours - Primary teacher's hours of work per week:

colnames(mydf)[colnames(mydf)=="TeachHours"] <- "InsegnanteOreLavoro"


####### 4 #######

summary(mydf)


####### 5 #######


# dalle statistiche generali si puo' evincere che Lima ha un TaxRate negativo
# questo elemento pare anomalo. Si e' velocemente verificato su Internet il sistema
# di tassazione del Peru' e si e' appurato che non sembrano esistere sistemi di welfare
# di negative income tax (NIT), soprattutto se consideriamo un reddito di un insegnante
# Pertanto si e' valutato di eliminare l'osservazione Lima


mydf = mydf[-which(rownames(mydf)=="Lima"),]

# Si ? notato nella variabile BigMac un valore molto alto per Nairobi, pari a 185
# si ? valutato, con qualche esitazione, di mantenere l'osservazione relativa a Nairobi




####### 6 #######

# install.packages("ggplot2")

library(ggplot2)
```

```
#### GRAFICI TUTOR INIZIO ############
#FoodIndex


sorted_foodIndex = sort(BigMac2003$FoodIndex)
citta = factor(rownames(BigMac2003[order(BigMac2003$FoodIndex),]))
citta = factor(citta, levels(citta)[order(BigMac2003$FoodIndex)])

qplot(rownames(BigMac2003), BigMac2003$FoodIndex)

qplot(citta, sorted_foodIndex) +
  geom_hline(yintercept=mean(BigMac2003$FoodIndex), color="red") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = "Citt?", y = "FoodIndex")

#GRAFICO32
ggplot(BigMac2003, aes(citta, sorted_foodIndex, colour = sort(FoodIndex-mean(BigMac2003$FoodIndex)))) +
  geom_point() +
  geom_hline(yintercept=mean(BigMac2003$FoodIndex), color="red") +
  scale_colour_gradientn(colours = c("red", "green", "blue"), breaks = c(-40, 0, 70), limits=c(-40, 70)) +
  labs(x = "Citt?", y = "FoodIndex", colour = "Scostamento\ndalla\nmedia") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))


#GRAFICO33
#Bread
minuti = factor(round(sort(BigMac2003$Bread)/10), labels = c("10", "20", "30", "40", "50", "60", "90"))

ggplot( BigMac2003 ,
        aes(x = factor (""), fill = minuti ) ) +
  geom_bar() +
  coord_polar( theta = "y") +
  labs(x = "", y = "")

#GRAFICO34
citta = factor(rownames(BigMac2003[order(BigMac2003$Bread),]))
citta = factor(citta, levels(citta)[order(BigMac2003$Bread)])
ggplot(BigMac2003, aes(citta, minuti, colour = minuti)) +
  geom_point() +
  labs(x = "Citt?", y = "Bread") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))



#GRAFICO35
#TaxRate
newTaxRate = factor(ifelse(BigMac2003$TaxRate < mean(BigMac2003$TaxRate), "Basso", "Alto"), levels = c("Basso",
"Alto"))
names(newTaxRate) = rownames(BigMac2003)
#plot(newTaxRate, rownames(BigMac2003))
#qplot(newTaxRate, rownames(BigMac2003))
citta = factor(names(sort(newTaxRate)))
citta = factor(citta, levels(citta)[order(newTaxRate)])

newTaxRate = sort(newTaxRate)

ggplot(BigMac2003, aes(newTaxRate, citta, colour = newTaxRate)) +
  geom_point() +
  labs(x = "TaxRate", y = "Citta'")




### GRAFICI TUTOR FINE


# Valuta la distribuzione della variabile BigMac
# si evidenzia una lunga coda...
ggplot2::ggplot(data = mydf) +
  ggplot2::geom_density(mapping = aes(x = mydf$BigMac), alpha = 0.4, fill = "blue") +
  xlab("BigMac") + ylab("Densita'") + ggtitle("Grafico di densita? per BigMac")



#Variabile CostoPane: grafico a torta in suddivisione in gruppi di interesse
group = cut(mydf$CostoPane, breaks = c(0,10,20,45,max(mydf$CostoPane)))
somma = as.data.frame(group)
somma[,"conta"] = 1
suddivisione = as.data.frame(aggregate(conta ~ group, data = somma, sum))
ggplot(suddivisione, aes(x="", y=suddivisione$conta, fill=suddivisione$group))+ geom_bar(width = 1, alpha = 0.7, stat
= "identity") + coord_polar("y", start=0) + ggtitle("Suddivisione di Costo Pane per categorie")

# Boxplot di tre variabili significative
# si evidenziano degli outlier... qualche caso anomalo?
boxplot(mydf$CostoRiso, main="Boxplot di Costo del Riso ")
```

```
#rappresentazione di Food Index e traccio una semplice least squares di FoodIndex
plot(mydf$FoodIndex)
abline(lsfit(1:length(mydf$FoodIndex), mydf$FoodIndex))



# Valuto i quartili della variabile CostoBigliettoBus
plot(cut(mydf$CostoBigliettoBus, breaks = quantile(mydf$CostoBigliettoBus) ))


#Variabile Riso: grafico di densita'
ggplot2::ggplot(data = mydf) +
  geom_density(  mapping= aes(x = mydf$CostoBigliettoBus), alpha=0.6, fill="Yellow") + xlab("Costo del Biglietto") +
ylab("Densita'") + ggtitle("Grafico di densita' per CostoBigliettoBus")




# Per affitto si possono analizzare tanti punti fuori da una ipotetica banda
# di controllo
plot(mydf$Affitto)
abline(sd(mydf$Affitto)+mean(mydf$Affitto), 0)
abline((-sd(mydf$Affitto))+mean(mydf$Affitto), 0)
abline(mean(mydf$Affitto), 0)

#escludo affitto per dare piu' regolarita' al grafico
boxplot(mydf[,-which(colnames(mydf)=="Affitto")], data=mydf)


#GRAFICO36
#La crescita dello stipendio nei vari paesi del mondo non ha un andamento lineare
# ipotesi: potrebbe essere una misura della diseguaglianza sui redditi
ggplot(mydf, aes(1:length(mydf$InsegnanteRLordo), cumsum(sort(mydf$InsegnanteRLordo)))) + geom_point(alpha = 3/10,
colour = "black")

#l'andamento di InsegnanteRNetto ? molto simile al lordo
ggplot(mydf, aes(1:length(mydf$InsegnanteRNetto), cumsum(sort(mydf$InsegnanteRNetto)))) + geom_point(alpha = 5/10,
colour = "blue")


# verifico graficamente la forte correlazione (?) tra reddito lordo e reddito netto
# ipotesi: uno e' la proxy dell'altro (?!)
plot(mydf$InsegnanteRLordo, mydf$InsegnanteRNetto, main = "Grafico a dispersione",col = 4)


#GRAFICO37
#Analisi sulle TASSE per paese
# l'istogramma ha dati eccessivi, ? stato quindi ruotato, lo si propone solo come esperimento
# lo si ordina dal paese con meno tasse a quello con pi? tasse
Paesi = rownames(mydf)
Tasse = mydf$TaxRate
dt = data.frame(Paesi, Tasse)
colnames(dt) = c("Paesi", "Tasse")

ggplot2::ggplot(data = dt) +
  ggplot2::geom_histogram(stat = "identity",
                          mapping = aes(x = reorder(dt$Paesi,dt$Tasse), y = dt$Tasse) ) + coord_flip()



#GRAFICO38
# Ore lavoro
# Per l'italia il CCNL insegnati prevede 25 ore di insegnamento per la primaria
# Nel data-set Milano e Roma sono a 24
Gruppi = ifelse(mydf$InsegnanteOreLavoro >= 25,">24 ore/sett.", "<25 ore/sett." )

t = as.data.frame(Gruppi)
t[,"conta"] = 1

df1 = as.data.frame(aggregate(conta ~ Gruppi, data = t, sum))
ggplot2::ggplot(data = df1) + ggtitle("Osservazioni con ore inf. o supp rispetto a Italia") +
  ggplot2::geom_bar(stat = "identity", mapping = aes(x = "", y = conta, fill=Gruppi)) + coord_polar("y", start=0)



####### 7 #######

mydf$TaxRate = ifelse(mydf$TaxRate >= mean(mydf$TaxRate),"Alto", "Basso" )

mydf$TaxRate = factor(mydf$TaxRate)

# levels(mydf$TaxRate) = c("Alto", "Basso")
```

```
####### 8 #######

# Trasformata TaxRate in variabile qualitativa, e' possibile utilizzare un istogramma
# per rappresentare la suddivisione delle Tasse nei paesi, contando quelli con tasse alte e
# quelli con tasse basse rispetto alla media.

ggplot2::ggplot(data = mydf ) +
  stat_summary( mapping = aes(x = TaxRate, y = BigMac, fill = TaxRate), fun.y = mean, geom = "bar",
                colour = "black") +
  xlab("Tasse") + ylab("Media di BigMac per raggruppamento") +
  ggtitle("Suddivisione delle tasse rispetto alla media")




####### 9 #######

# install.packages("corrplot")
library("corrplot")

corrplot::corrplot(cor(mydf[,sapply(mydf, is.numeric)]),
                   method = "number", type = "upper", order = "AOE",
                   title = "Matrice di correlazione lineare")

# Analisi delle relazioni
# Relazione forte tra
#    +0.7    CostoRiso e BigMac
#    -0.62   BigMac e InsegnanteRLordo
#    -0.55   CostoPane e InsegnanteRLordo
#    +0.7    CostoBigliettoBus e InsegnanteRLordo
#    +0.77   InsegnanteRLordo e FoodIndex
#    +0.99   InsegnanteRLordo e InsegnanteRNetto
#    +0.79   InsegnanteRNetto e FoodIndex
#    +0.72   Affitto e FoodIndex

####### 10 #######

# Analisi delle relazioni piu' significative.
# Ho deciso di mettere in relazione la variabile BigMac con la variabile CostoRiso,
# questo perche' sembra che la variabile BigMac abbia una correlazione con CostoRiso ,
# e pare invece decorrelato con InsegnanteRLordo.
# Mentre InsegnanteRLordo e InseganteRNetto sono sostanzialmente indentici,  sono
# ambedue significativamente correlati con FoodIndex e CostoBigilettoBus

#GRAFICO39
# Di seguito si rappresenta graficamente la relazione tra CostoRiso e BigMac
# parrebbe esserci una sorta di relazione inversa, all'aumento di BigMac
# diminuisce foodIndex e viceversa.
# Inoltre, si e' provato a classificare il TaxRate colorando diversamente i
# pallini. Si evidenzia una concentrazione di tasse piu' elevate in relazione
# a valori di BigMac pi? bassi e FoodIndex pi? alti.
# Ho inoltre provato ad usare uno smoother per cercare di delineare un modello di regressione,
# ma solo come tenativo sperimentale
ggplot2::ggplot(data = mydf) +
  geom_point( mapping = aes( x = CostoRiso, y = BigMac, color = TaxRate), shape = 1, size = 2, stroke = 1.25) +
  geom_smooth(mapping = aes(x = CostoRiso, y = BigMac, linetype = "r2"),
              method = "lm",
              formula = y ~ log(x), se = T,
              color = "blue") +
  xlab("Indice BigMac") + ylab("Indice FoodIndex") +
  ggtitle("Relazione tra BigMac e FoodIndex e classificazione per livelli di tassazione")



#GRAFICO40
# Altro report sperimentale: I paesi che hanno tasse alte hanno il BigMac meno costoso
ggplot(mydf, aes(x = mydf$BigMac)) +
  geom_area(aes(fill = mydf$TaxRate), stat ="bin", bins = 30,  alpha=0.6) +
  theme_classic()

####### 8 #######
### Realizzazione di un modello - Modello lineare con 2 variabili Y-> BigMac e X-> Riso ####
myBigMac2003 = mydf
### Si esamina graficamente la correlazione tra le variabili e si ipotizza una correlazione di tipo lineare
plot(BigMac2003$BigMac, BigMac2003$Riso)

### Dalla correlazione lineare rilevo il forte legame lienare tra le variabili BigMac e Riso
### Utilizzo solo queste due variabili per creare un modello di correlazione lineare attraverso lm()
linear_model1 = glm(formula = myBigMac2003$BigMac ~ myBigMac2003$CostoRiso, data = myBigMac2003)


####### 9 #######
### Analisi del modello - Modello lineare con 2 variabili Y-> BigMac e X-> Riso ####
```

```
### Con la funzione summary visualizzo i dati del modello creato
summary(linear_model1)

### Si commentano i valori rilevanti
# Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
# (Intercept)         8.1097     4.4616   1.818   0.0737 .
# myBigMac2003$Riso   1.4329     0.1779   8.056  2.1e-11 ***
# Multiple R-squared:  0.4958,  Adjusted R-squared:  0.4881

### myBigMac2003$Riso, ha un p-value molto basso 2.1e-11 Ã¨ una e una significativitÃ  a tre star ***
### Multiple R-squared:  0.4958 indica una rappresetnativitÃ  dell'informazione quasi del 50%
### Commento: gli indicatori ci indicano che la variabile indipendente Riso Ã¨ una scelta corretta, ma rappresenta
ancora troppo poco
### Nuova strategia: si cerca di introdurre una nuova variabile per accrescere la rappresentativitÃ  del modello

####### 10 #######
### Realizzazione del modello - Modello lineare con 3 variabili Y-> BigMac e X-> Riso e StipendioNetto ####

# All'analisi aggiungiamo SipendioNetto in quanto correlata con BigMac dall'analisi della correlazione lienare
# L'analisi del grafico indica una correlazione (si prova la lineare per una prima verifica)
plot(myBigMac2003$BigMac, myBigMac2003$InsegnanteRNetto)

# Si crea il modello con le tre variabili tutte di tipo lineare
linear_model2 = glm(formula = myBigMac2003$BigMac ~ myBigMac2003$CostoRiso + myBigMac2003$InsegnanteRNetto, data =
myBigMac2003)


####### 11 #######
### Analisi del modello - Modello lineare con 3 variabili Y-> BigMac e X-> Riso e StipendioNetto ####
summary(linear_model2)

### Si commentano i valori rilevanti
#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
# (Intercept)                27.1405     6.4200   4.228 7.54e-05 ***
#  myBigMac2003$Riso           1.0930     0.1847   5.919 1.33e-07 ***
#  myBigMac2003$StipendioNetto  -0.7688     0.2009  -3.828 0.000293 ***
# ---
#  Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

# Residual standard error: 20.34 on 65 degrees of freedom
# Multiple R-squared:  0.5885,  Adjusted R-squared:  0.5759


### myBigMac2003$Riso, ha un p-value molto basso 1.33e-07 e una e una significativitÃ  a tre star ***
### myBigMac2003$StipendioNetto, ha un p-value  basso 0.000293 e una e una significativitÃ  a tre star ***
### Multiple R-squared:  0.5885 indica una rappresetnativitÃ  dell'informazione quasi del 58%
### Commento: Introducendo StipendioNetto il p-value di Riso Ã¨ diminuito e la rappresentativitÃ  dell'informazione
Ã¨ aumentata al 58%
### Nuova strategia: si cerca di studiare StipendioNetto con una fomulazione quadratica per verificare come cambia il
modello


####### 12 #######
### Realizzazione del modello - Modello lineare con 3 variabili, e variabile quadratica Y-> BigMac e X-> Riso e
StipendioNetto^2 ####

# Si crea il modello con le tre variabili di cui una quadratica, si utilizza la funzione poly
linear_model3 = glm(formula = myBigMac2003$BigMac ~ myBigMac2003$CostoRiso + poly(myBigMac2003$InsegnanteRNetto,2) ,
data = myBigMac2003)


####### 13 #######
### Analisi del modello - Modello lineare con 3 variabili, e variabile quadratica Y-> BigMac e X-> Riso e
StipendioNetto^2 ####
summary(linear_model3)


### Si commentano i valori rilevanti
# Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept)                              18.6397     4.2482   4.388 4.37e-05 ***
#  myBigMac2003$Riso                         0.9044     0.1797   5.033 4.19e-06 ***
#  poly(myBigMac2003$StipendioNetto, 2)1 -100.1925    21.7491  -4.607 2.00e-05 ***
#  poly(myBigMac2003$StipendioNetto, 2)2   67.8169    19.7972   3.426  0.00108 **
#  ---
#  Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1
# Residual standard error: 18.85 on 64 degrees of freedom
# Multiple R-squared:  0.6523,  Adjusted R-squared:  0.636



# myBigMac2003$Riso, mantiene significativitÃ  con p-value elevata con tre star ***
# poly(myBigMac2003$StipendioNetto, 2)1, mantiene significativitÃ  con p-value elevata con tre star ***
```

```
# poly(myBigMac2003$StipendioNetto, 2)2, ha significativitÃ  con p-value media con due star **
# Multiple R-squared:  0.6523, Ã¨ cresciuto arrivando ad una rappresantivitÃ  al 65%
# Nuova strategia:



####### 14 #######
### Realizzazione del modello - Modello lineare con 4 variabili, e variabile quadratica Y-> BigMac e X-> Riso e
StipendioNetto^2 e FoodIndex ####

# Si introduce la variabile FoodIndex in quanto fortemente collegata con StipendioNetto
# Si crea il modello con le quattro variabili di cui una quadratica, si utilizza la funzione poly
linear_model4 = glm(formula = myBigMac2003$BigMac ~ myBigMac2003$CostoRiso + poly(myBigMac2003$InsegnanteRNetto,2) +
myBigMac2003$FoodIndex , data = myBigMac2003)


####### 15 #######
### Analisi del modello - Modello lineare con 3 variabili, e variabile quadratica Y-> BigMac e X-> Riso e
StipendioNetto^2 ####
summary(linear_model4)

### Si commentano i valori rilevanti

# Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
# (Intercept)                             32.1698    10.2132   3.150   0.0025 **
# myBigMac2003$Riso                        0.9517     0.1811   5.255 1.87e-06 ***
#  poly(myBigMac2003$StipendioNetto, 2)1 -60.3785    34.8506  -1.732   0.0881 .
# poly(myBigMac2003$StipendioNetto, 2)2  57.8527    20.7888   2.783   0.0071 **
#  myBigMac2003$FoodIndex                 -0.2325     0.1599  -1.454   0.1509
# Multiple R-squared:  0.6636,

# Commento generale In generale si ha perdita di significativitÃ  su tutte le variabili
# R-square: Ã¨ rimasto invariato al 66%
# Nuova strategia: si esamina il grafico dei residui del modello precedente linear_model3 e si rileva una forma ad
imbuto che suggerisce una eteroschedasticitÃ
plot(linear_model3)

####### 16 #######
### Realizzazione del modello - Modello lineare con 3 variabili, e variabile quadratica sqrt(Y)-> sqrt(BigMac) e X->
Riso e StipendioNetto^2  ####

# Si crea il modello con le tre variabili di cui una quadratica, si utilizza la funzione poly, con la variabile
dipendente sotto radice quadrata
# Si utilizza la radice quadra sulla variabile dipendete BigMac per aver intuito la presenza di eteroschedasticitÃ
in quanto la forma del plot dei residui Ã¨ ad imbuto
linear_model5 = glm(formula = sqrt(myBigMac2003$BigMac) ~ myBigMac2003$CostoRiso +
poly(myBigMac2003$InsegnanteRNetto,2) , data = myBigMac2003)


####### 17 #######
### Analisi del modello - Modello lineare con 3 variabili, e variabile quadratica sqrt(Y)-> sqrt(BigMac) e X-> Riso e
StipendioNetto^2 ####
summary(linear_model5)

# Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept)                            4.71092    0.24526  19.208  < 2e-16 ***
#  myBigMac2003$Riso                      0.04800    0.01038   4.626 1.87e-05 ***
#  poly(myBigMac2003$StipendioNetto, 2)1 -9.58737    1.25565  -7.635 1.42e-10 ***
#  poly(myBigMac2003$StipendioNetto, 2)2  5.58794    1.14296   4.889 7.14e-06 ***
#  ---
# Multiple R-squared:  0.7548


### Si commentano i valori rilevanti
# Tutte le variabili hanno un p-value significativo
# R-squared Ã¨ cresciuto ad un livello pari al 75% di significativitÃ
# Nuova strategia: chiusura del modello con commento del grafico

####### 16 #######
### Analisi grafica del modello  lineare con 3 variabili, e variabile quadratica  sqrt(Y)-> sqrt(BigMac) e X-> Riso e
StipendioNetto^2 ####
plot(linear_model5)

# Residui: hanno una forma ad imbuto
# I residui standardizzati si dispongono sulla retta con distorsione sugli estremi
# Distanza di cooks indica la presenza di outlier e una concentrazione delle osservazioni

####### 17 #######

library(leaps)

subsets = regsubsets(BigMac ~., mydf, method = "forward")
```

```
subsets.summary = summary(subsets)

plot(subsets)

plot(subsets.summary$rss[1:dim(array(subsets.summary$rss))])

plot(subsets.summary$rsq[1:dim(array(subsets.summary$rsq))])


plot(subsets, scale="r2")
plot(subsets, scale="adjr2")
plot(subsets, scale="Cp")
plot(subsets, scale="bic")

plot(subsets.summary$cp,xlab="Number of Variables",ylab="Cp", type="l")
which.min(subsets.summary$cp)
coef(subsets, which.min(subsets.summary$cp))

points(which.min(subsets.summary$cp), subsets.summary$cp[which.min(subsets.summary$cp)],col="red",cex=2,pch=20)




####### 18 #######

library(boot)

cverr = rep(NA,5)

linear_model1 = glm(formula = sqrt(BigMac) ~ CostoRiso, data = mydf)
linear_model2 = glm(formula = sqrt(BigMac) ~ CostoRiso + InsegnanteRNetto, data = mydf)
linear_model3 = glm(formula = sqrt(BigMac) ~ CostoRiso + poly(InsegnanteRNetto,2) , data = mydf)
linear_model4 = glm(formula = sqrt(BigMac) ~ CostoRiso + poly(InsegnanteRNetto,2) + FoodIndex , data = mydf)
linear_model5 = glm(formula = sqrt(BigMac) ~ CostoRiso + poly(InsegnanteRNetto,2) , data = mydf)


coef(linear_model1)

cverr[1] = cv.glm(mydf, linear_model1, K = 10)$delta[1]

cverr[2] = cv.glm(mydf, linear_model2, K = 10)$delta[1]

cverr[3] = cv.glm(mydf, linear_model3, K = 10)$delta[1]

cverr[4] = cv.glm(mydf, linear_model4, K = 10)$delta[1]

cverr[5] = cv.glm(mydf, linear_model5, K = 10)$delta[1]

for(i in 1:5){
  print(cverr[i])
  }
which.min(cverr)


####### 19 #######
#Intervallo di predizione e confidenza

## VALIDATION APPROACH
## Costruisco il TRAINING SET
set.seed(1)
nperc = 0.8
training.index = sample(1:nrow(mydf), nperc * nrow(mydf), replace = F)

## Creo un modello con il TRAINING SET
#NON USARE GLM
lm3_train = lm(formula = BigMac ~ CostoRiso + poly(InsegnanteRNetto,2) , data = mydf,subset = training.index)

#plot(linear_model3_train)


# Costruisco la predizione utilizzando il TEST-SET
p_mod3_test = predict(lm3_train, mydf[-training.index,],  interval="prediction")
p = as.data.frame(p_mod3_test)

p$BigMac = mydf[-training.index,]$BigMac
p$CostoRiso = mydf[-training.index,]$CostoRiso

# Verifico l'accuratezza della mia predizione
mean((mydf[-training.index,]$BigMac - p$fit) ^ 2)


# Plot verifico aderenza tra la predizione e i dati reali del test set
plot(mydf[-training.index,]$BigMac ~ p$fit )
```

```
# 2. Regression line + confidence intervals
library("ggplot2")
gg <- ggplot(p, aes( x=CostoRiso, y=BigMac)) +
  geom_point() +
  stat_smooth(method = lm)
# 3. Add prediction intervals
gg + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y = upr), color = "red", linetype = "dashed")



####### 20 #######
#LOOCV


LOOCV1 = cv.glm(mydf, linear_model1 )
LOOCV5 = cv.glm(mydf, linear_model5)


LOOCV1$delta[1]
LOOCV5$delta[1]


#LOOCV MANUALE
sum_loocv=rep(0, nrow(mydf))

for(i in 1:nrow(mydf)){
  fit = glm(formula = sqrt(BigMac) ~ CostoRiso + poly(InsegnanteRNetto,2), data = mydf[-i,])
  pred = predict(fit, newdata = mydf[i,], type = "response")

  # attenzione: la variabile di risposta in questo caso ha SQRT
  # MSE[i]
  sum_loocv[i] = (pred - sqrt(mydf$BigMac[i]))^2

}
# loocv
sum(sum_loocv)/nrow(mydf)
```