Esame29042021

Alessandro Risaro

29/4/2021

Pulizia dell' environment, caricamento delle librerie e di una funzione utile per svolgere il white test (meno restrittiva di quella fornita da R)

```
rm(list=ls())

library(car)
library(describedata)
library(skedastic)
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.5
```

```
library(klaR)
```

```
## Warning: package 'MASS' was built under R version 4.0.5
```

```
library(olsrr)
library(sandwich)
library(DataCombine)
library(systemfit)
library(lmtest)
library(ggplot2)

white.test<-function(lmod){

  u2<-lmod$residuals^2

  y<-lmod$fitted

  R2u<-summary(lm(u2~y+I(y^2)))$r.squared

  LM<-length(y)*R2u

  p.val<-1-pchisq(LM,2)

  data.frame("Test Statistic"=LM, "P"=p.val)

}
```

Importazione dei dati

```
data<-read.csv(file.choose(), sep=',')
```

Andiamo a verificare che non vi siano valori nulli e nel caso eliminiamo tutte le righe che hanno almeno una covariata nulla

```
data<-na.omit(data)
```

Non vi sono valori nulli procediamo quindi con l'analisi 6.13 Ordiniamo i dati secondo la colonna time

```
data<-data[order(data$time),]
```

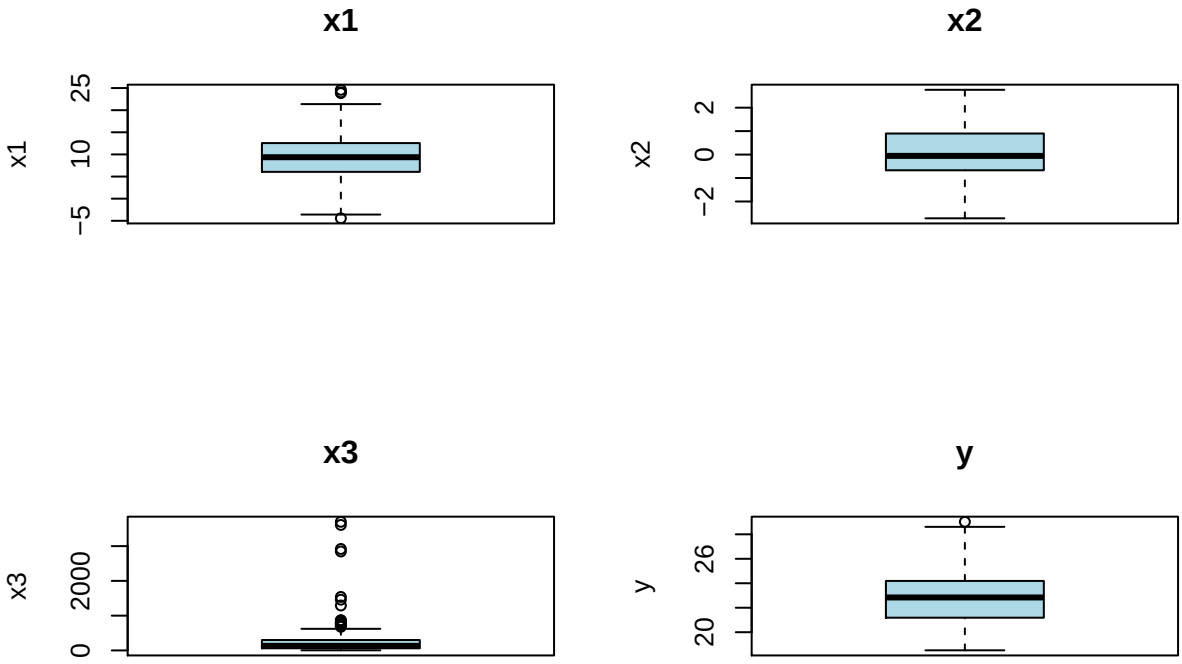
Statistiche descrittive

```
summary(data)
```

```
##       time             x1               x2                x3
##  Min.   :  1.00   Min.   :-4.440   Min.   :-2.71699   Min.   :   3.281
##  1st Qu.: 50.75   1st Qu.: 6.059   1st Qu.:-0.66916   1st Qu.:  61.562
##  Median :100.50   Median : 9.366   Median :-0.05528   Median : 131.239
##  Mean   :100.50   Mean   : 9.533   Mean   : 0.01307   Mean   : 276.579
##  3rd Qu.:150.25   3rd Qu.:12.549   3rd Qu.: 0.89788   3rd Qu.: 297.468
##  Max.   :200.00   Max.   :24.603   Max.   : 2.76130   Max.   :3699.064
##        y
##  Min.   :18.52
##  1st Qu.:21.20
##  Median :22.84
##  Mean   :22.91
##  3rd Qu.:24.18
##  Max.   :29.02
```

Vediamo che le variabili x1, x2 e x3 presentano dei range diversi, in particolare la variabile x3 presenta dei valori di massimo molto elevati, questo può segnalare la presenza di outliers.

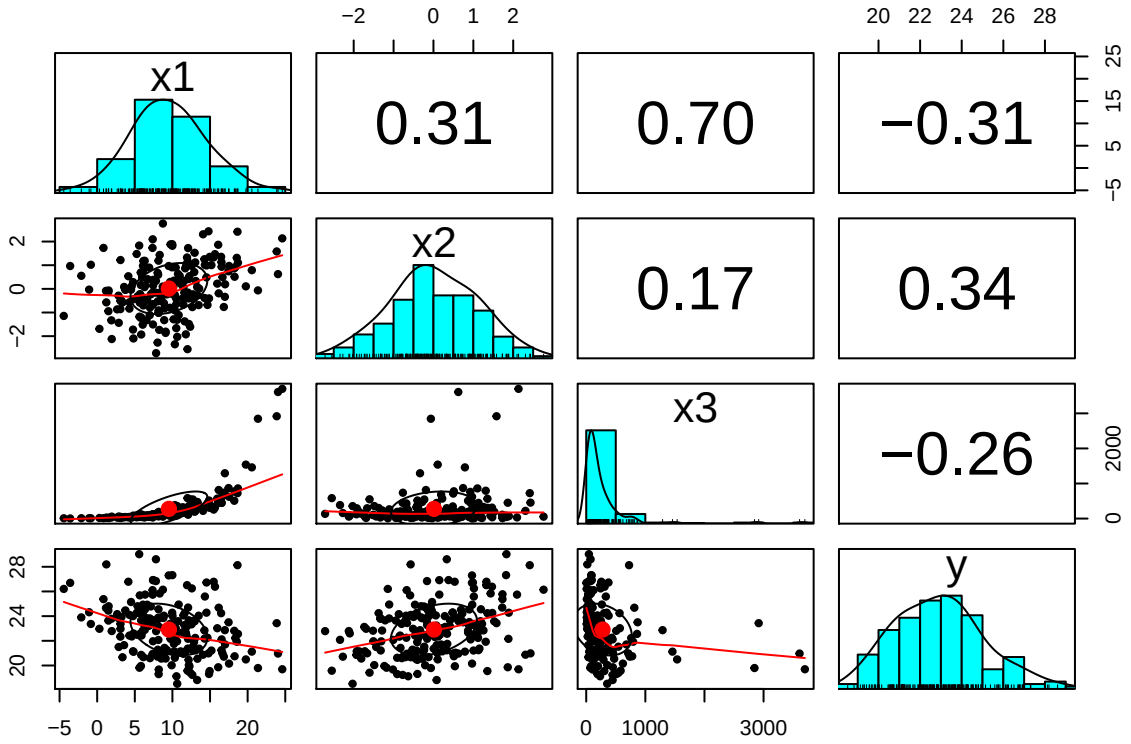
```
var_num<-c("x1","x2","x3","y")
par(mfrow=c(2,2))
for( i in var_num){ boxplot(data[,i], main=i,
col="lightblue", ylab=i)}
```



Anche guardando i boxplot vediamo la presenza nella variabile x3 di numerosi outliers

Andiamo a vedere la distribuzione e le correlazioni tra le variabili

```
pairs.panels(data[,var_num])
```

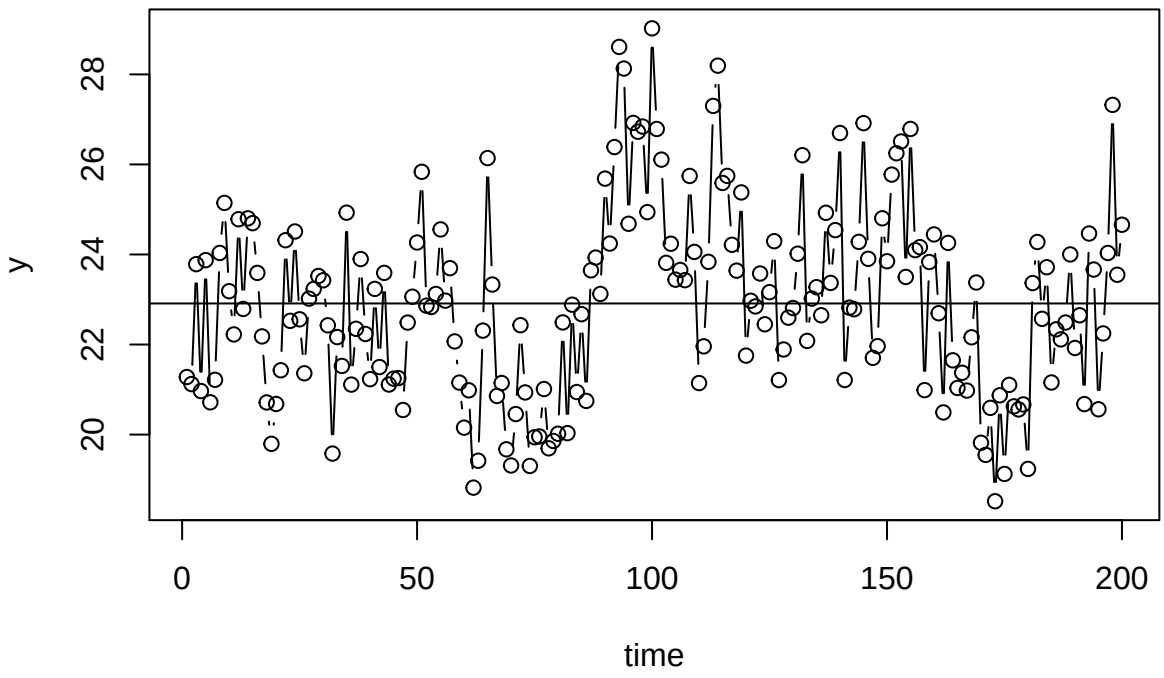


Notiamo una correlazione positiva di 0.7 tra le variabili x1 e x3, la quale però non è abbastanza elevata (>0.9) da poter indicare la presenza di multicollinearità delle variabili, le altre correlazioni sembrano essere modeste.

Le variabili x1 e x2 sembrano distribuirsi normalmente, notiamo una leggera assimmetria positiva nella variabile y ed un' elecata assimmetria positiva nella variabile x3 data dalla presenza di outliers

Andiamo a vedere l'andamento della variabile dipendente y nel tempo

```
plot(data$time,data$y,ylab="y",xlab='time',type='b')
abline(h=mean(data$y))
```



Notiamo che non vi è la presenza di trend crescenti o decrescenti, ma i valori sembrano muoversi intorno alla media

Modello lineare di y (variabile dipendente) su x1, x2 e log(x3) (variabili indipendenti): test di ipotesi ed interpretazione dei coefficienti

```
m1<-lm(y~x1+x2+I(log(x3)), data)
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + I(log(x3)), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6859 -1.1202 -0.0515  0.8360  4.9887
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.06768    1.51116  17.912  < 2e-16 ***
## x1           0.02248    0.13660   0.165  0.86943
## x2           0.70371    0.19458   3.617  0.00038 ***
## I(log(x3))  -0.89438    0.56447  -1.584  0.11470
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.748 on 196 degrees of freedom
## Multiple R-squared:  0.3183, Adjusted R-squared:  0.3078
## F-statistic:  30.5 on 3 and 196 DF,  p-value: 3.162e-16
```

Vediamo che il modello nel suo complesso è significaitvo, infatti la statistica F cade nella regione di rifiuto dell' ipotesi nulla di non significatività di tutti i coefficienti. L'unica variabile significativa risulta essere la variabile x2 ed ha un coefficente di 0.70371 il quale significa che ad un aumento unitario di x1 segue un aumento di y di 0.70371, stesso discorso vale per la variabile x1 (anche se non risulta significativa), mentre per la variabile log(x3) (anch'essa non significativa) l'interpretazione del coefficente è differente infatti si interpreta come un incremento del 1% di x3 porta ad un incremento di -0.0089438 della variabile y. L'adattamento ai dati del modello risulta essere abbastanza debole, infatti abbiamo un indice di determinazione R^2 aggiustato per il numero di variabili esplicative di 0.3078

A partire dal modello stimato nel punto precedente diagnosticare multicollinearità, omoschedasticità, normalità dei residui ed outlier. Si risolvano eventuali violazioni

Iniziamo diagnosticando un'eventuale presenza di multicollinearità con opportuni indici

- VIF

```
ols_vif_tol(m1)
```

```
##    Variables  Tolerance       VIF
## 1         x1 0.03149537 31.750696
## 2         x2 0.34794905  2.873984
## 3 I(log(x3)) 0.03470776 28.812002
```

Vediamo che il VIF considerando la soglia 10 evidenzia la presenza di multicollinearità tra le variabili x1 e log(x3), mentre non sembrano esserci problemi di multicollinearità per la variabile x2

- Condition Index

```
ols_eigen_cindex(m1)
```

```
##    Eigenvalue Condition Index    intercept           x1           x2
## 1 2.885319483        1.000000 0.0007539820 7.926600e-04 0.0004437957
## 2 1.007306244        1.692451 0.0000617798 3.105362e-05 0.3387415037
## 3 0.106144918        5.213716 0.0307002236 3.294456e-02 0.0437347227
## 4 0.001229354       48.446060 0.9684840146 9.662317e-01 0.6170799779
##     I(log(x3))
## 1 2.256514e-04
## 2 1.019403e-05
## 3 9.465887e-08
## 4 9.997641e-01
```

Abbiamo un autovalore associato ad un condition index maggiore della soglia 10, e vediamo che spiega più del 95% della variabilità delle variabili x1 e log(x3), quindi anche il condition index evidenzia la presenza di multicollinearità tra le variabili x1 e log(x3).

Procediamo quindi a togliere una delle due variabili nel nostro caso scegliamo log(x3) dal modello

```r
m2<-lm(y~x1+x2, data)
summary(m2)
```

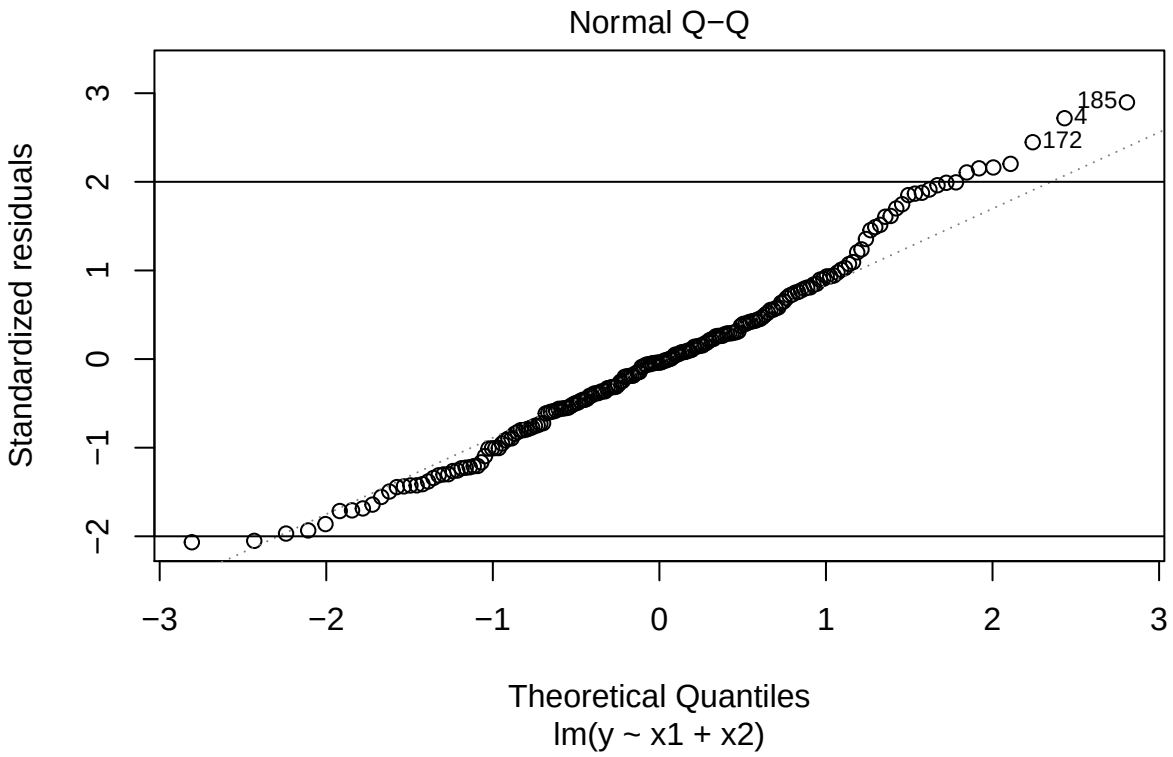
```
##
## Call:
## lm(formula = y ~ x1 + x2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5983 -1.0597 -0.0690  0.9671  5.0657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.71253    0.27347  90.367  < 2e-16 ***
## x1          -0.19014    0.02562  -7.422 3.38e-12 ***
## x2           0.94538    0.12129   7.794 3.65e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.754 on 197 degrees of freedom
## Multiple R-squared:  0.3095, Adjusted R-squared:  0.3025
## F-statistic: 44.16 on 2 and 197 DF,  p-value: < 2.2e-16
```

Notiamo che adesso il modello rimane significativo nel suo complesso, in particolare anche la variabile x1 risulta ora essere significativa. Mentre la bontà di adattamento del modello non sembra essere cambiata

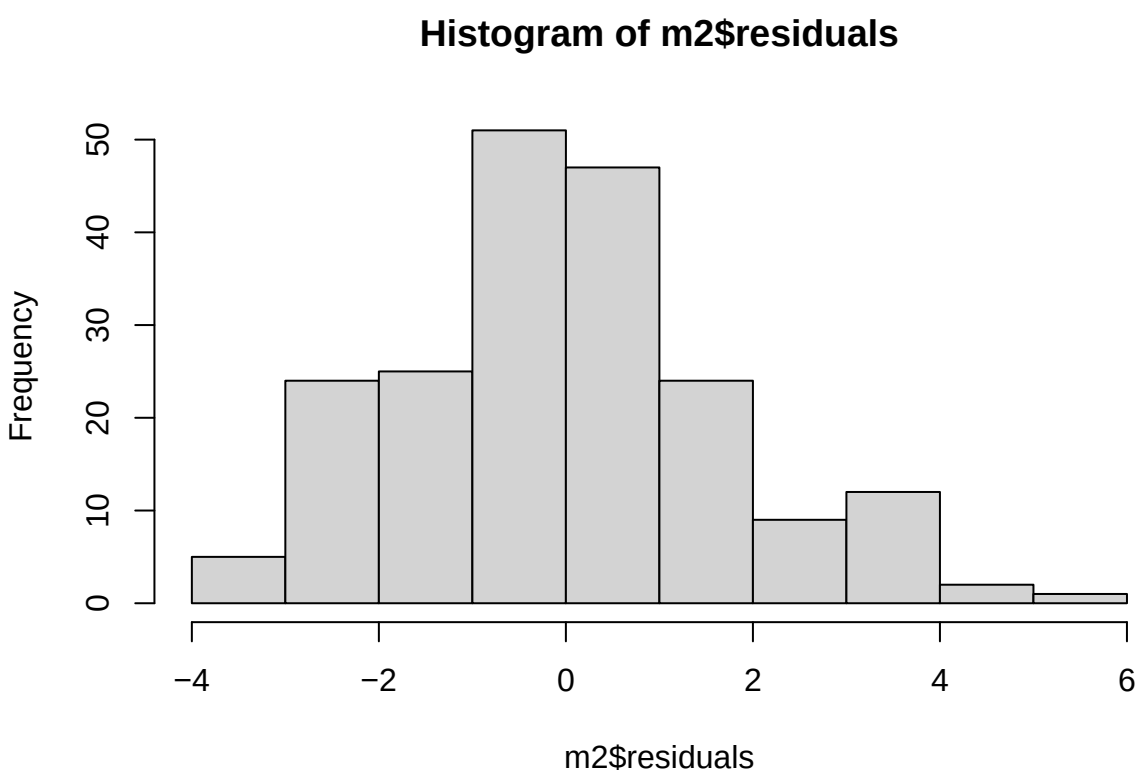
Andiamo ora ad indagare la normalità dei residui

La normalità dei residui viene in un primo luogo indagata graficamente, attraverso l'uso di un QQ-plot e anche dell'istogramma dei residui

```r
plot(m2, which=2)
abline(h=-2)
abline(h=2)
```




```
hist(m2$residuals)
```



Andiamo a verificare la normalità tramite l'uso di alcuni test

```
ols_test_normality(m2)
```

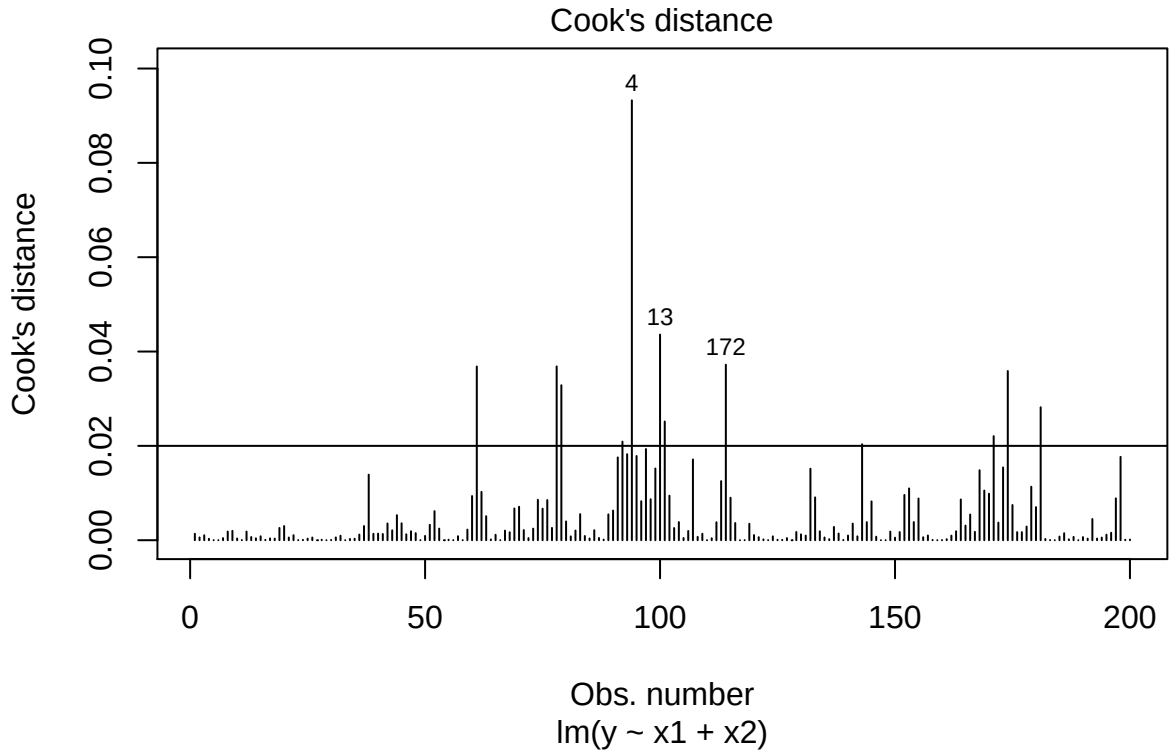
```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9832         0.0175
## Kolmogorov-Smirnov        0.0622         0.4221
## Cramer-von Mises         14.1514         0.0000
## Anderson-Darling          0.8314         0.0315
## -----------------------------------------------
```

Dai grafici dei residui si può vedere come la loro distribuzione si avvicina a quella della normale standardizzata eccetto degli scostamenti sulla coda destra (probabilmente dovuti ad outlier). Il test di Shapiro Wilk ha p-value=0.175, quindi potrebbe portare al rifiuto dell'ipotesi di normalità, però va detto che questo test non è molto affidabile quando vi è la presenza di outlier, mentre il test di Kolmogorov-Smirnov evidenzia la presenza di normalita. Va detto che essendo n=200 gli errori standard calcolati sono validi asintoticamente in quanto vale il teorema del limite centrale, quindi scegliamo di non attuare correzioni.

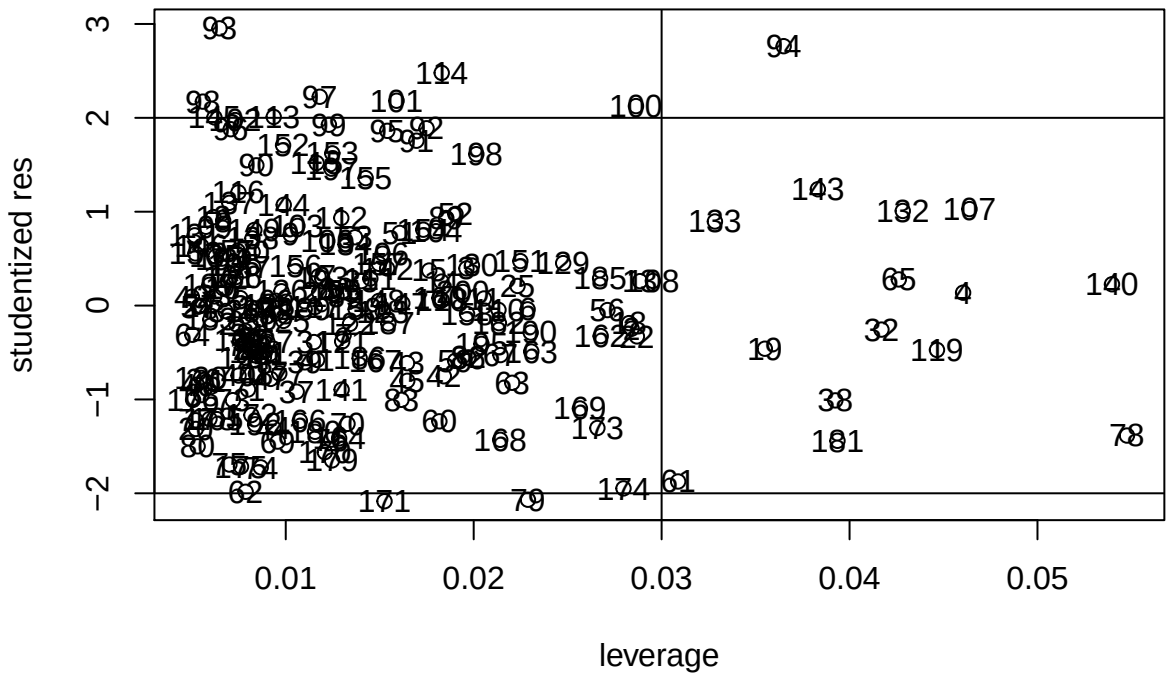
Andiamo ora a verificare la presenza di outliers e di valori influenti

```
k=length(coef(m2))
n=nrow(data)
```

```
plot(m2,which=4)
abline(h=4/n)
```




```
plot(hatvalues(m2),rstudent(m2),xlab='leverage',ylab='studentized res')
abline(h=2)
abline(h=-2)
abline(v=2*k/n)
text(hatvalues(m2),rstudent(m2))
```



La cook's distance è una misura di influenza, ovvero di quanto un'osservazione influisca sulle stime dei parametri, vediamo che l'osservazione 4 ha un valore molto influente, notiamo anche che le osservazioni 13 e 172 superano abbondantemente le soglie I residui studentizzati anch'essi misura di influenza, indica in particolare la 140 e la 78 come misure influenti I levarage indicano la presenza di alcuni outlier

Procediamo quindi a rimuovere gli outlier usando come metrica i residui studentizzati e i valori di leverage

```
data2<-data[hatvalues(m2)<=2*k/n & abs(rstudent(m2))<2,]
```

Notiamo che abbiamo eliminato ben 25 osservazioni

Vediamo se questo ha portato ad un miglioramento del nostro modello

```
m3<-lm(y~x1+x2, data2)
summary(m3)
```

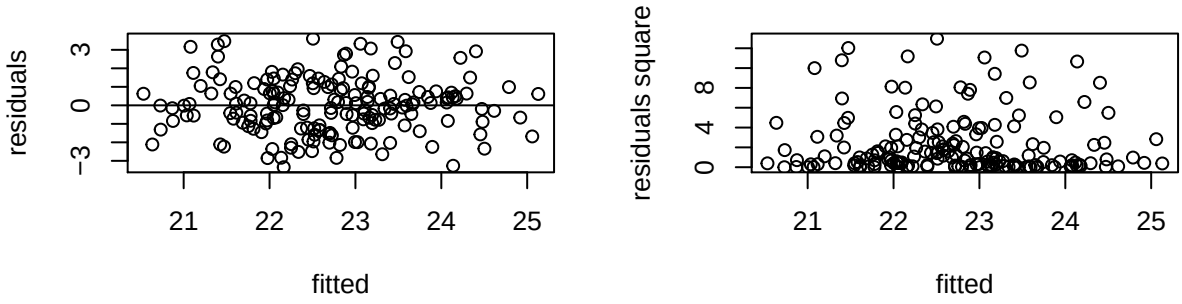
```
##
## Call:
## lm(formula = y ~ x1 + x2, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3434 -0.9706  0.0542  0.9356  3.6005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 24.45482    0.28606  85.489  < 2e-16 ***
## x1          -0.17809    0.02749  -6.477 9.44e-10 ***
## x2           0.94927    0.12426   7.639 1.45e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.506 on 172 degrees of freedom
## Multiple R-squared:  0.3058, Adjusted R-squared:  0.2977
## F-statistic: 37.88 on 2 and 172 DF,  p-value: 2.336e-14
```

Vediamo che sembra essere tutto rimasto uguale al modello precendente, infatti il modello risulta significativo, entrambe le variabili sono significative e il coefficente di determinazione aggiustato sembra essere rimasto più o meno uguale

Procediamo indagando la presenza di omoschedasticità dei residui sia graficamente che attraverso il test di white

```
par(mfrow=c(2,2))
plot(m3$fitted,m3$residuals,xlab='fitted',ylab='residuals')
abline(h=0)
plot(m3$fitted,(m3$residuals)^2,xlab='fitted',ylab='residuals square')
```




```
white.test(m3)
```

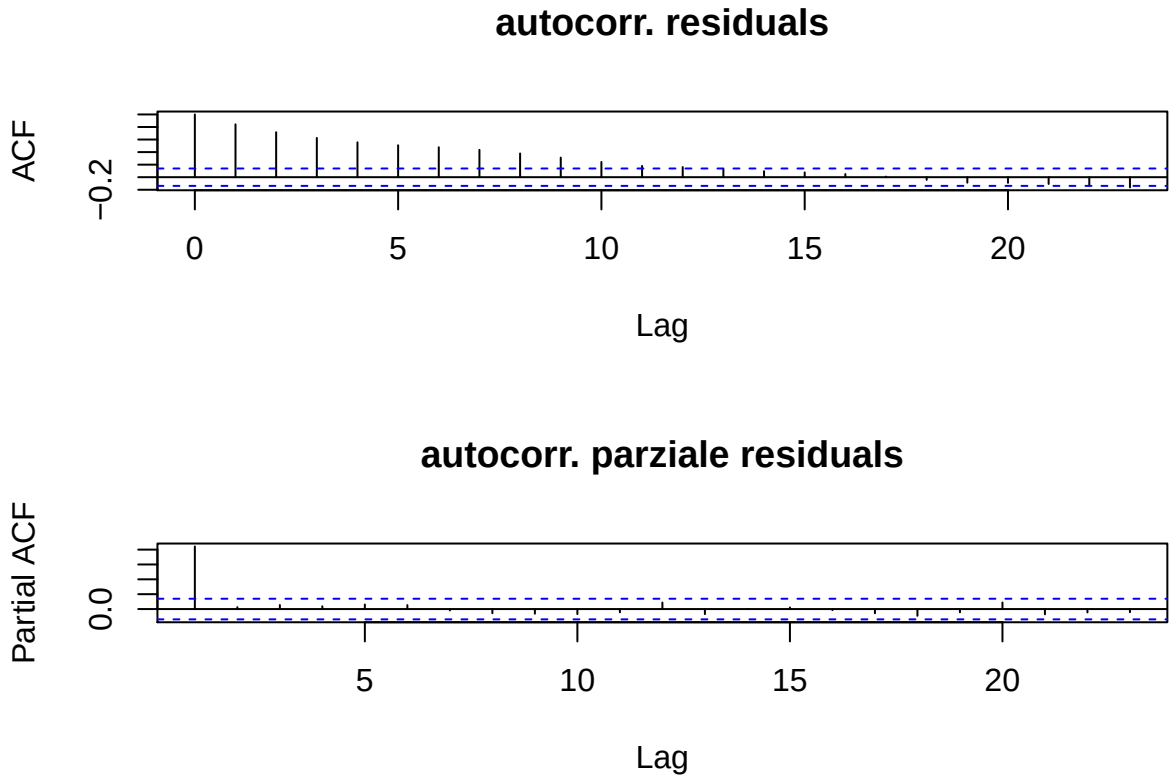
```
##   Test.Statistic        P
## 1      0.7507225 0.687041
```

Graficamente non si può dire moltissimo se non che non sembra esserci presenza di una particolare forma a ventaglio dei residui, mentre il white test porta a non rifiutare l'ipotesi nulla di omoschedasticità dei residui. Quindi possiamo concludere che non vi sia eteroschedasticità

Senza rimuovere osservazioni e a partire dalle variabili incluse nel modello stimato al termine del punto precedente , studiare e risolvere la potenziale autocorrelazione utilizzando test ed almeno un grafico diagnostico.

Andiamo a vedere la presnza di correlazioni dei residui attraverso l'uso dei correlogrammi

```
par(mfrow=c(2,1))
acf(m2$residuals,main='autocorr. residuals')
pacf(m2$residuals,main='autocorr. parziale residuals')
```



Notiamo che dal primo correlogramma emerge la presenza di autocorrelazioni significative fino al decimo ordine, mentre per quanto riguarda il secondo grafico emerge la presenza di autocorrelazione dei residui di primo ordine, in particolare le statistiche di durbin Watson tendenti a 0 segnalano la presenza di una correlazione positiva tra i residui

Andiamo avedere con il test di darbinwatson se si ottengono ulteriori conferme riguardo la presenza di autocorrelazione (studiamo la presenza di autocorrelazione fino al quindicesimo ordine)

```
durbinWatsonTest(m2, max.lag=15)
```

```
##  lag Autocorrelation D-W Statistic p-value
```

```
##    1      0.84246351     0.3117179   0.000
##    2      0.71720770     0.5616804   0.000
##    3      0.62598705     0.7305112   0.000
##    4      0.55596888     0.8598509   0.000
##    5      0.50920478     0.9518115   0.000
##    6      0.47752573     1.0143303   0.000
##    7      0.43594489     1.0957253   0.000
##    8      0.37888203     1.2052920   0.000
##    9      0.31299188     1.3261575   0.000
##   10      0.24294558     1.4649026   0.000
##   11      0.18000515     1.5903437   0.026
##   12      0.16144519     1.6229720   0.080
##   13      0.12597828     1.6923262   0.196
##   14      0.09620399     1.7513644   0.360
##   15      0.07648394     1.7885261   0.564
##  Alternative hypothesis: rho[lag] != 0
```

Notiamo che anche il test di DurbinWatson ci segnala la presenza di autocorrelazione dei residui fino all'undicesimo ordine

Andiamo a cercare di risolvere l'autocorrelazione di primo ordine attraverso la procedura di di cochrane-orcutt, la quale si divide in diversi passi:

- Ottenere i residui di m2 creare i residui ritardati attraverso la funzione slide,

```
data$u_hat<-m2$residuals
data<-slide(data=data,Var='u_hat',TimeVar='time',NewVar='u_hat_lag')
```

```
##
## Lagging u_hat by 1 time units.
```

- Andiamo a svolgere la regressione dei residui su i residui ritardati, in quanto vogliamo andare a ricavare il coefficente di autocorrelazione

```
aux<-lm(u_hat~u_hat_lag,data)
summary(aux)
```

```
##
## Call:
## lm(formula = u_hat ~ u_hat_lag, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97025 -0.58003  0.05489  0.58255  2.35994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.008121   0.066611   0.122    0.903
## u_hat_lag   0.842555   0.038165  22.077   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9397 on 197 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7122, Adjusted R-squared:  0.7107
## F-statistic: 487.4 on 1 and 197 DF,  p-value: < 2.2e-16
```

- Memorizziamo il coefficente di autocorrelazione

```
rho<-aux$coefficients[2]
rho
```

```
## u_hat_lag
## 0.8425552
```

- Tramite la funzione slide andiamo a creare tutte le variabili ritardate

```
data<-slide(data=data,Var='y',TimeVar='time',NewVar='y_lag')
```

```
##
## Lagging y by 1 time units.
```

```
data<-slide(data=data,Var='x1',TimeVar='time',NewVar='x1_lag')
```

```
##
## Lagging x1 by 1 time units.
```

```
data<-slide(data=data,Var='x2',TimeVar='time',NewVar='x2_lag')
```

```
##
## Lagging x2 by 1 time units.
```

- Andiamo a calcolare le variabili trasformate tramite il coefficente di autocorrelazione

```
data$y_t<-data$y-rho*data$y_lag
data$x1_t<-data$x1-rho*data$x1_lag
data$x2_t<-data$x2-rho*data$x2_lag
data$interc_t<-1-rho
```

- Stimiamo il modello delle variabili trasformate (metto lo zero perchè devo togliere la costante del modello e al suo posto sarà inserito 1-rho) e lo confrontiamo con quello del modello con errori correlati

```
m2_t<-lm(y_t~0+interc_t+x1_t+ x2_t,data)
summary(m2_t)
```

```
##
## Call:
## lm(formula = y_t ~ 0 + interc_t + x1_t + x2_t, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.92368 -0.60019  0.02343  0.58831  2.33102
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## interc_t 24.79894    0.43424   57.11   <2e-16 ***
## x1_t     -0.19396    0.01022  -18.97   <2e-16 ***
## x2_t      0.97845    0.05186   18.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.941 on 196 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9468, Adjusted R-squared:  0.946
## F-statistic:  1162 on 3 and 196 DF,  p-value: < 2.2e-16
```

Andiamo a vedere che il modello nel suo complesso risulta essere significativo, entrambe le variabili x1 e x2 sono significative e vi è stato un aumento della bontà di adattamento, infatti si è passati da un R^2 aggiustato di circa 0.3 ad un R^2 aggiustato di circa 0.95

```
durbinWatsonTest(m2_t, max.lag=15)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.015731459      2.028123   0.744
##    2    -0.032258886      2.033594   0.766
##    3    -0.007593326      1.981864   0.986
##    4    -0.020029177      1.989661   0.902
##    5     0.007448101      1.920812   0.750
##    6     0.058878327      1.805288   0.306
##    7     0.098683891      1.700929   0.080
##    8     0.056557335      1.753859   0.248
##    9     0.026949442      1.778999   0.300
##   10     0.008203675      1.813492   0.478
##   11    -0.100387219      2.016180   0.382
##   12     0.054844832      1.698751   0.152
##   13    -0.005976656      1.818347   0.642
##   14    -0.021387805      1.847520   0.866
##   15     0.023729678      1.744636   0.368
##  Alternative hypothesis: rho[lag] != 0
```

Andiamo a vedere che l'autocorrelazione è stata risolta anche per tutti gli altri ordini