Es2

Niccolò Puccinelli

2022-05-10

```
data <- read.csv("companies.csv", sep=";")
View(data)
```

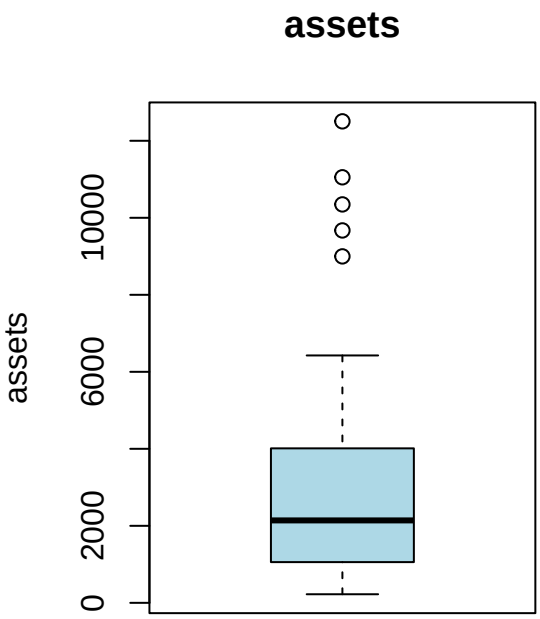
Il dataset presenta variabili numeriche, eccezion fatta per la "company" e "sector", di tipo qualitativo nominale. Occupiamoci anzitutto delle statistiche descrittive.

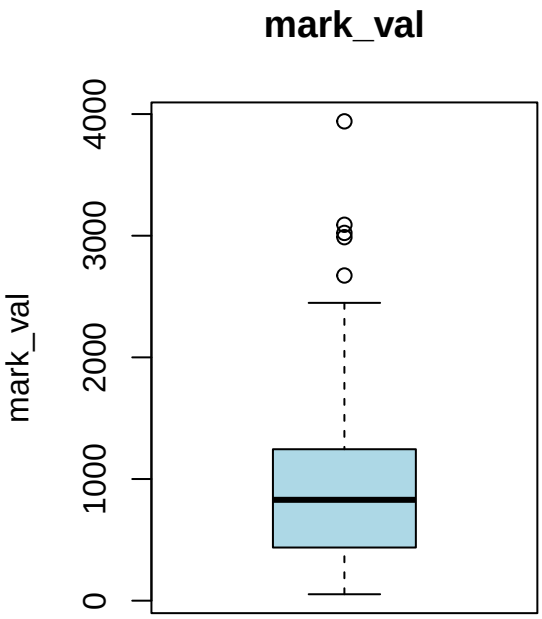
Statistiche descrittive

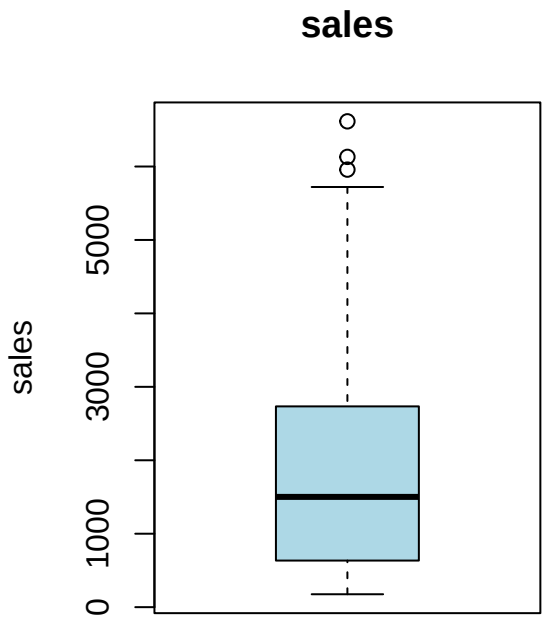
```
var_num<-c("assets", "mark_val", "sales", "profits", "cash", "employ")

summary(data[, var_num])
```

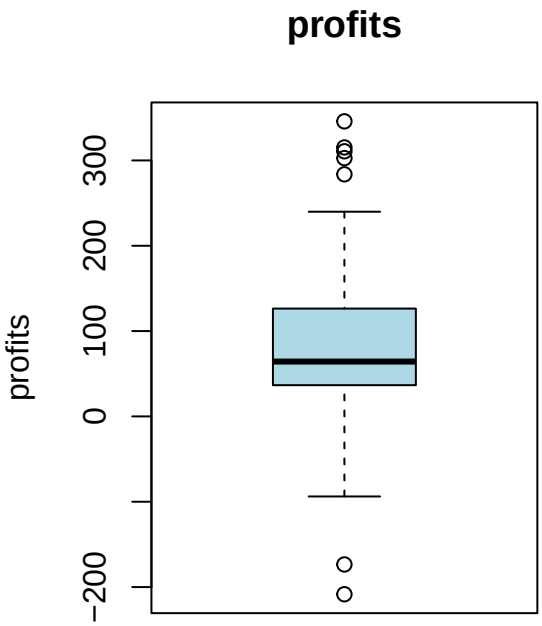
```
##      assets         mark_val          sales           profits
##  Min.   :  223   Min.   :  53.0   Min.   : 176.0   Min.   :-208.40
##  1st Qu.: 1075   1st Qu.: 440.5   1st Qu.: 653.2   1st Qu.:  37.20
##  Median : 2140   Median : 829.0   Median :1501.5   Median :  64.30
##  Mean   : 2997   Mean   :1036.7   Mean   :2006.5   Mean   :  86.01
##  3rd Qu.: 3976   3rd Qu.:1182.5   3rd Qu.:2698.0   3rd Qu.: 124.00
##  Max.   :12505   Max.   :3940.0   Max.   :6615.0   Max.   : 345.80
##       cash             employ
##  Min.   :-108.10   Min.   :  0.600
##  1st Qu.:  69.03   1st Qu.:  3.475
##  Median : 119.25   Median :  8.100
##  Mean   : 169.96   Mean   : 18.859
##  3rd Qu.: 228.38   3rd Qu.: 23.775
##  Max.   : 682.50   Max.   :118.100
```

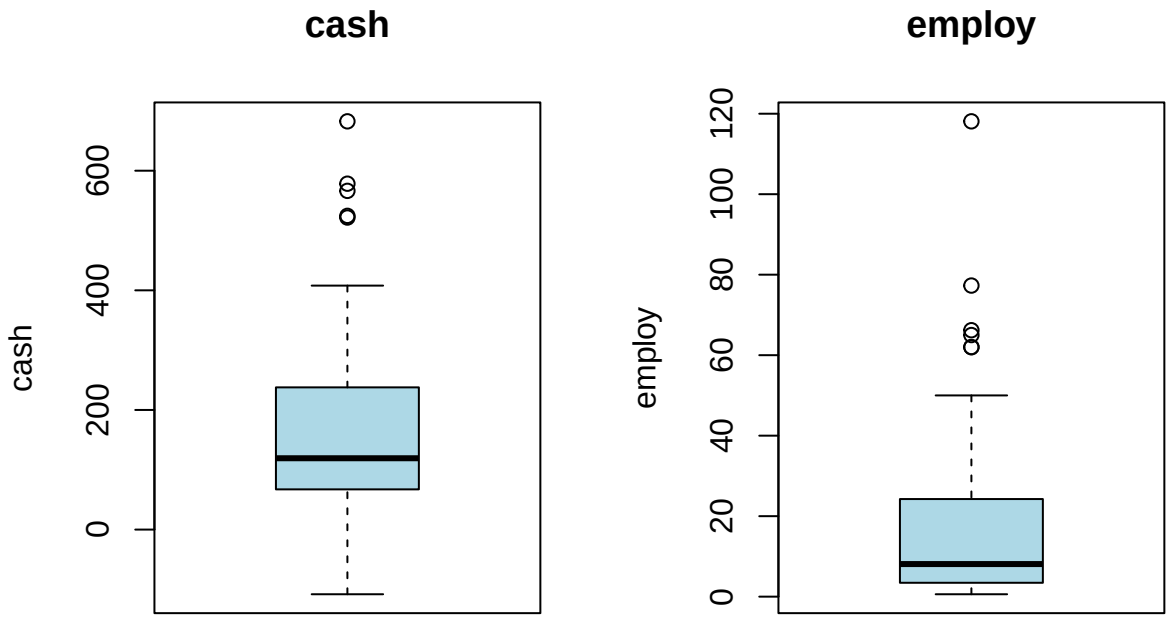
```
# Box-plot
par(mfrow=c(1,2))
for(i in var_num){
  boxplot(data[,i],main=i,col="lightblue",ylab=i)
}
```


assets
assets
0
2000
6000
10000


mark_val
mark_val
0
1000
2000
3000
4000


sales
sales
0
1000
3000
5000

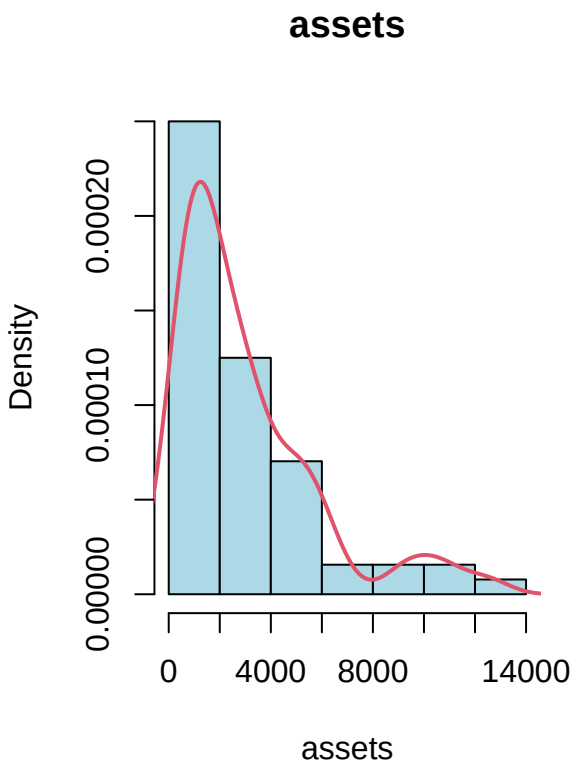

profits
profits
−200
0
100
200
300

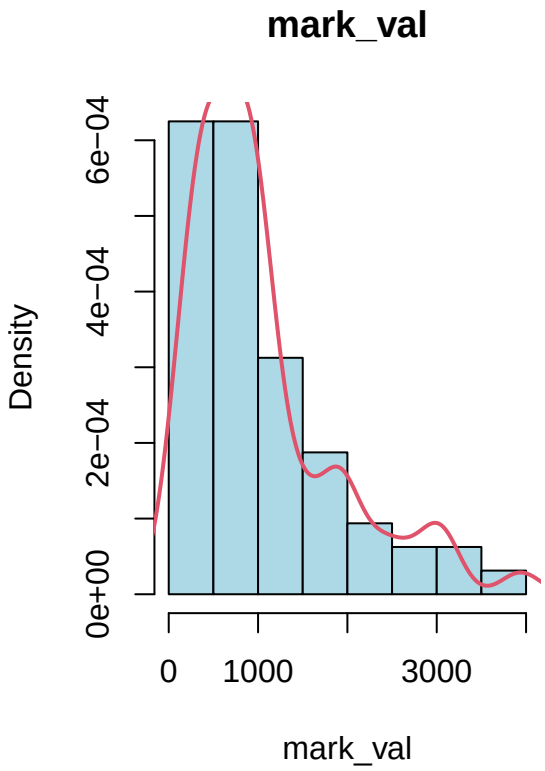


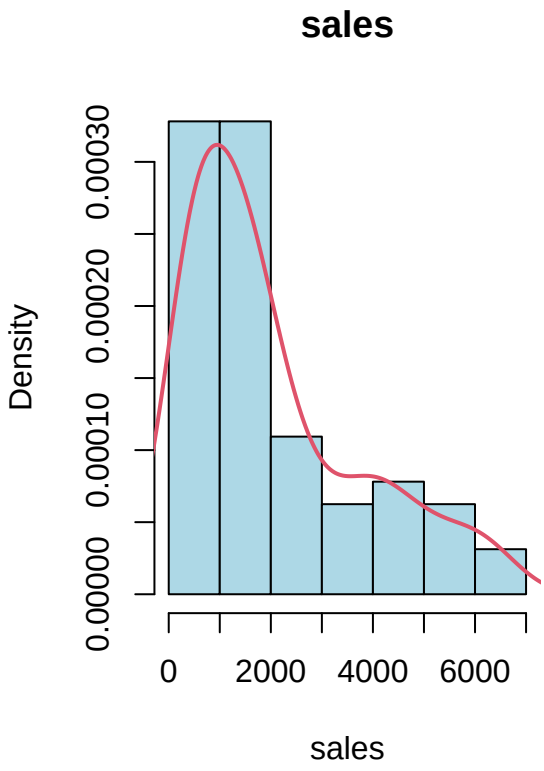
Dai box-plot notiamo che per quasi tutte le variabili la media si discosta dalla mediana, indicando una possibile non-normalità. Inoltre possiamo già vedere i primi outlier, presenti in tutte le variabili.

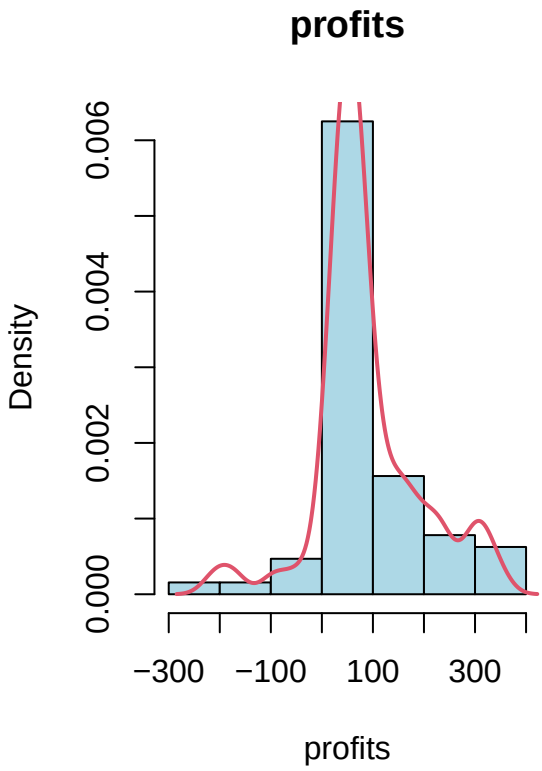
Andiamo a vedere simmetria e possibile non-normalità anche con degli istogrammi.

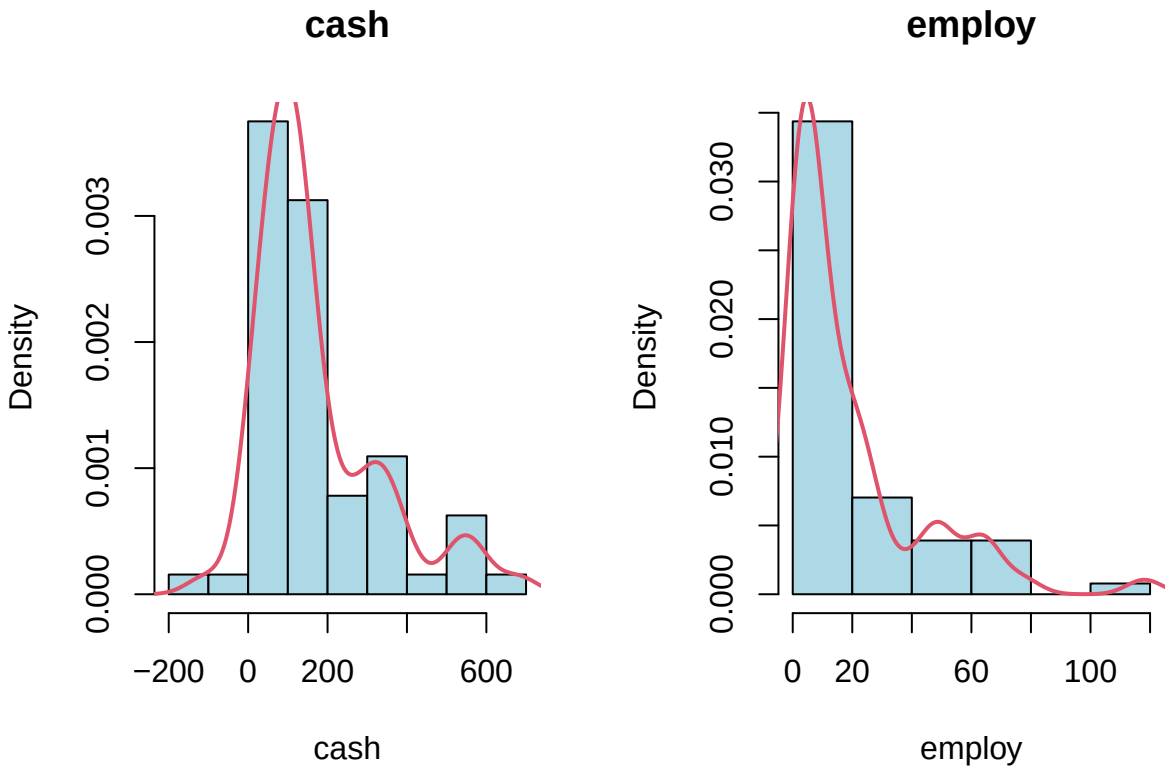
```
# Istogrammi
par(mfrow=c(1,2))
for(i in var_num){
  hist(data[,i],main=i,col="lightblue",xlab=i,freq=F,prob=TRUE)
  lines(density(data[,i]), col=2, lwd=2)
}
```


assets
Density
0.00000
0.00010
0.00020
0
4000
8000
14000
assets


mark_val
Density
0e+00
2e−04
4e−04
6e−04
0
1000
3000
mark_val


sales
Density
0.00000
0.00010
0.00020
0.00030
0
2000
4000
6000
sales

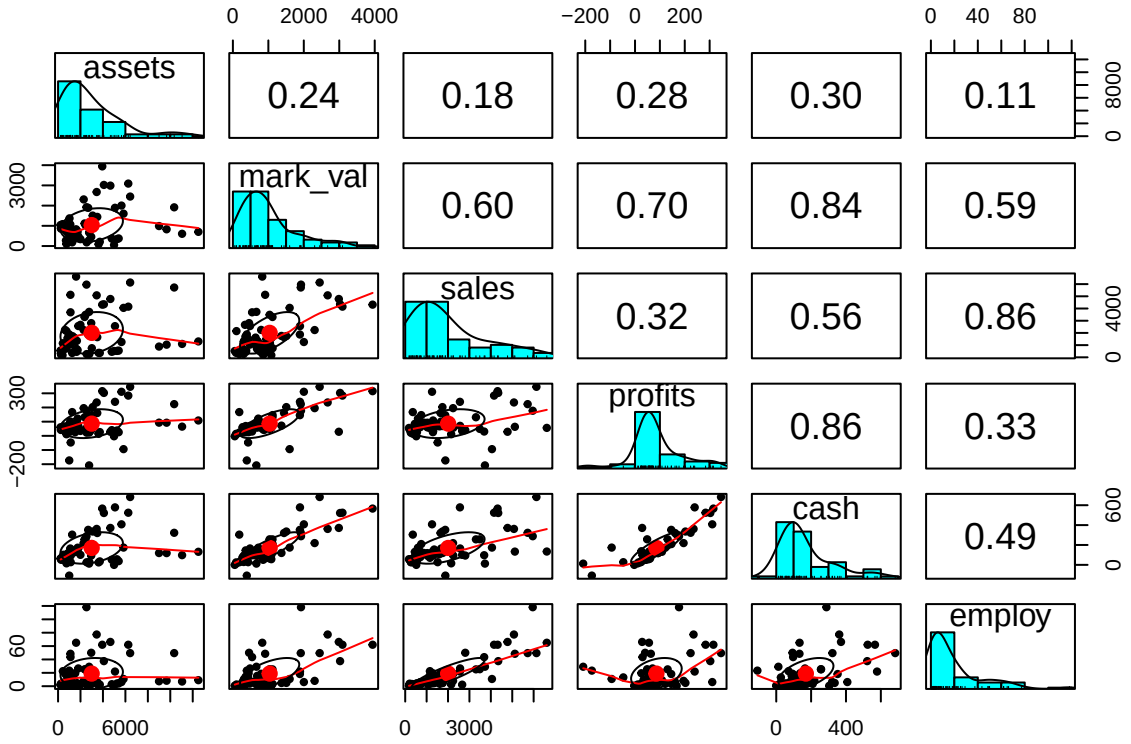

profits
Density
0.000
0.002
0.004
0.006
−300
−100
100
300
profits



Quasi tutte le variabili presentano un'asimmetria positiva (in particolare assets, mark_val, employ, sales), indicando una possibile non-normalità, da verificare con gli appositi grafici e test.

Andiamo a studiare le correlazioni.

```
# Studio delle correlazioni
pairs.panels(data[,var_num])
```




```
cor(data[,var_num])
```

```
##              assets  mark_val     sales   profits      cash    employ
## assets    1.0000000 0.2441519 0.1772699 0.2830843 0.3036877 0.1105166
## mark_val  0.2441519 1.0000000 0.5974219 0.6986348 0.8354449 0.5873249
## sales     0.1772699 0.5974219 1.0000000 0.3172530 0.5630487 0.8634541
## profits   0.2830843 0.6986348 0.3172530 1.0000000 0.8556017 0.3252593
## cash      0.3036877 0.8354449 0.5630487 0.8556017 1.0000000 0.4920105
## employ    0.1105166 0.5873249 0.8634541 0.3252593 0.4920105 1.0000000
```

Le variabili maggiormente correlate sono mark_val con cash (0.84), profits con cash (0.86), sales con employ (0.86) e mark_val con profits (0.7).

Consideriamo da ora in avanti esclusivamente le variabili di nostro interesse (i.e. mark_val, assets, sales, profits, cash ed employ).

Regressione di mark_val su assets, sales, profits, cash ed employ

Effetttuiamo la regressione e commentiamone i risultati.

```
m1<-lm(mark_val~assets+sales+profits+cash+employ,data)
summary(m1)
```

```
##
```

```
## Call:
## lm(formula = mark_val ~ assets + sales + profits + cash + employ,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1228.23  -222.80   -50.87   251.70  1180.29
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.438e+02  1.030e+02   2.367   0.0213 *
## assets       2.066e-04  2.127e-02   0.010   0.9923
## sales       -1.802e-02  7.445e-02  -0.242   0.8096
## profits      8.828e-02  1.176e+00   0.075   0.9404
## cash         3.787e+00  8.620e-01   4.393 4.82e-05 ***
## employ       9.399e+00  4.875e+00   1.928   0.0587 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 444.8 on 58 degrees of freedom
## Multiple R-squared:  0.7394, Adjusted R-squared:  0.7169
## F-statistic: 32.91 on 5 and 58 DF,  p-value: 9.393e-16
```

Le uniche variabili che risultano significative sono cash ed employ (anche se in minor modo). Il modello risulta significativo, poiché il p-value associato alla statistica F è molto basso. Infine l'R quadro risulta buono e spiega circa il 74% della variabilità totale, mentre l'R quadro aggiustato è minore di circa lo 0.02.

Andiamo a verificare eventuali episodi di multicollinearità.

Multicollinearità

Consideriamo gli appositi indici

```
vif(m1)
```

```
##   assets    sales  profits     cash   employ
## 1.117129 5.121912 4.566396 5.939145 4.121352
```

```
ols_vif_tol(m1)
```

```
##   Variables Tolerance      VIF
## 1    assets 0.8951514 1.117129
## 2     sales 0.1952396 5.121912
## 3   profits 0.2189911 4.566396
## 4      cash 0.1683744 5.939145
## 5    employ 0.2426388 4.121352
```

```
ols_eigen_cindex(m1)
```

```
##   Eigenvalue Condition Index  intercept      assets        sales      profits
## 1 4.54322435        1.000000 0.01054237 0.012715279 0.0033125789 0.0044760914
```

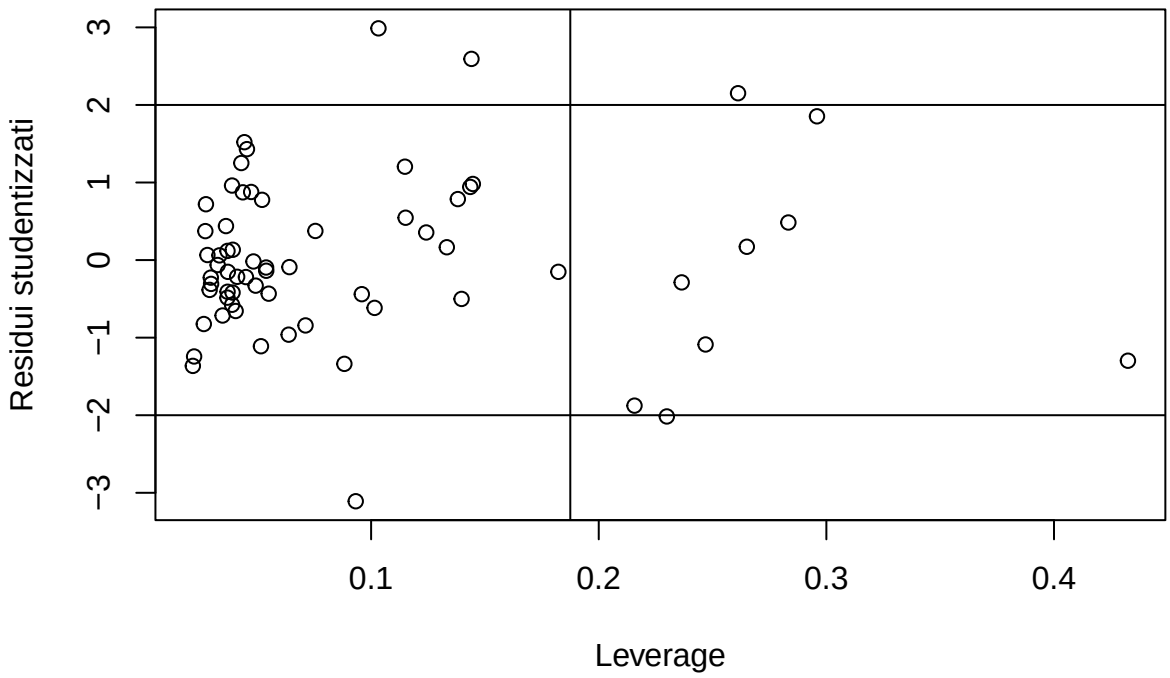
```
## 2 0.59747869        2.757534 0.01587922 0.178927361 0.0249846568 0.0205616751
## 3 0.49313275        3.035290 0.10384716 0.201041217 0.0035788982 0.1016401384
## 4 0.24526276        4.303941 0.69713864 0.598596685 0.0001225628 0.0001331748
## 5 0.08183714        7.450868 0.11718427 0.007204285 0.2675864899 0.2367365382
## 6 0.03906432       10.784300 0.05540834 0.001515173 0.7004148133 0.6364523820
##          cash       employ
## 1 0.003287019 0.0050374176
## 2 0.001747291 0.1008507392
## 3 0.026095211 0.0002876457
## 4 0.000110627 0.0327995688
## 5 0.254895575 0.4885831460
## 6 0.713864277 0.3724414827
```

Sia l'indice di tolleranza, sia il VIF non mostrano episodi di evidente multicollinearità. Il VIF, in particolare, è sotto 10 per ogni variabile esplicativa. Per quanto riguarda il condition index invece, questo risulta maggiore di 10 per l'autovalore più piccolo. Ciò nonostante, la quota di varianza spiegata associata all'indice è sempre minore dell'80% (al massimo vale 70% per sales e 71% per cash). Decidiamo dunque di non cambiare il modello e di proseguire l'analisi.

Outlier

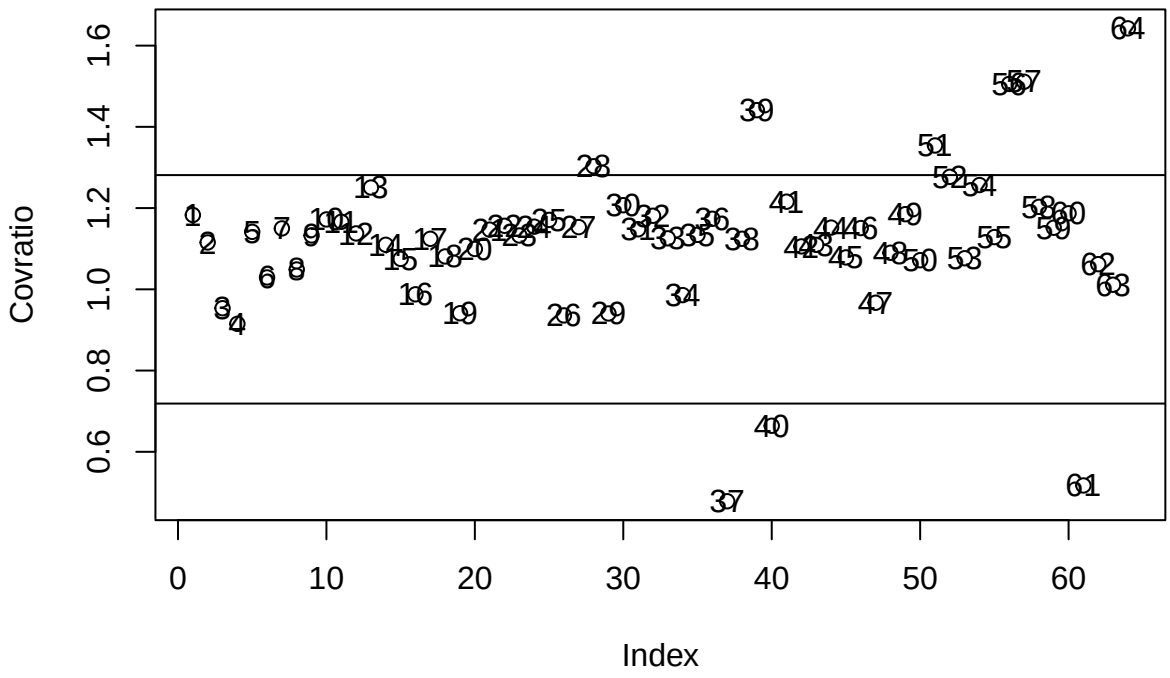
Verifichiamo la presenza o meno di outlier tramite il grafico residui studentizzati vs. leverage.

```
plot(hatvalues(m1), rstudent(m1), ylab="Residui studentizzati", xlab="Leverage")
# Soglie: 2, -2, 2k/n, oltre la quale si considerano outlier
# k=#coefficienti
# n=#osservazioni
abline(h=2)
abline(h=-2)
k=length(coef(m1))
n=nrow(data)
abline(v=2*k/n)
```

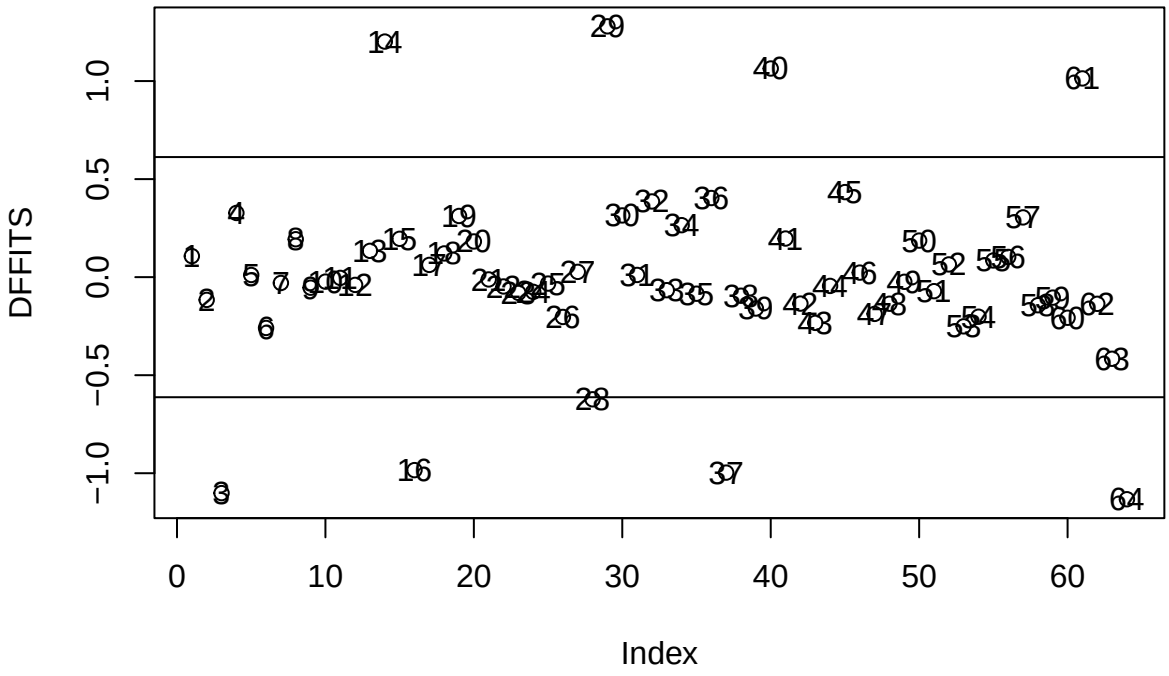


Notiamo la presenza di diversi outlier, i.e. i punti fuori dalla zona delimitata da v, 2 e -2. Utilizziamo i test.

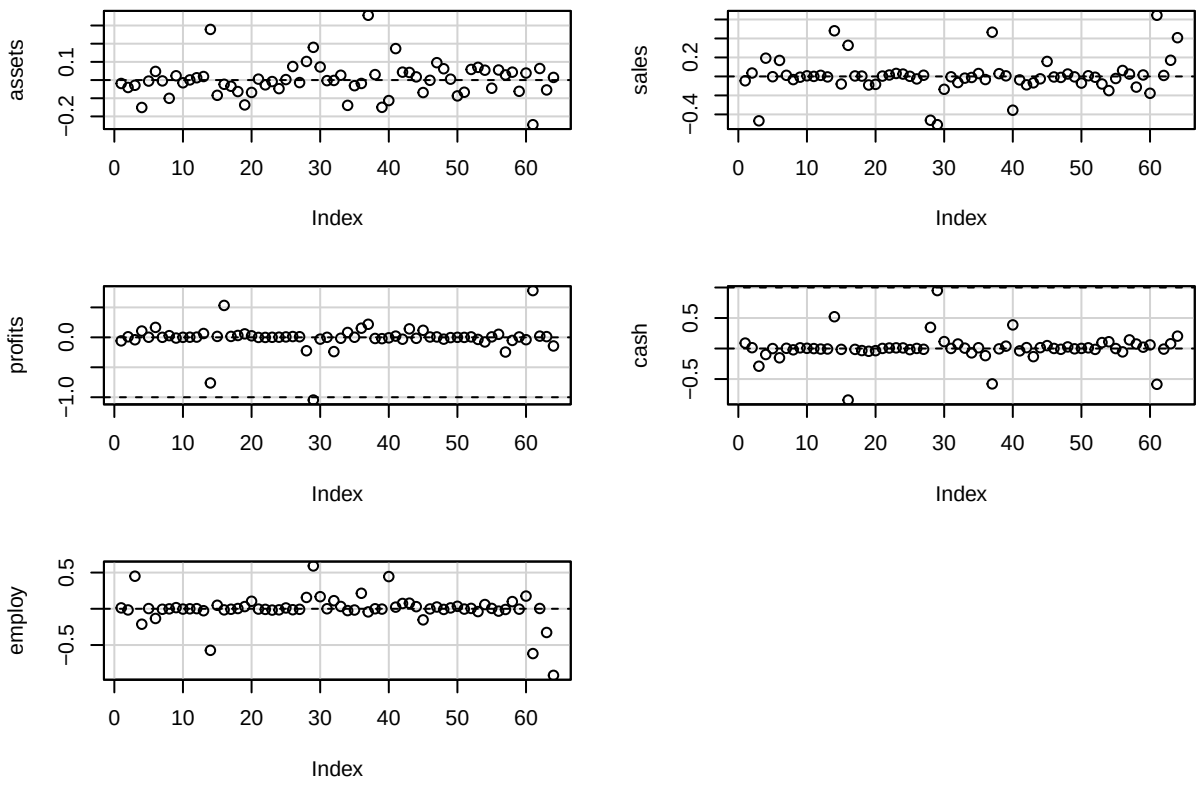
```
# COVRATIO
plot(covratio(m1), ylab="Covratio")
abline(h=1+3*k/n)
abline(h=1-3*k/n)
text(covratio(m1))
```



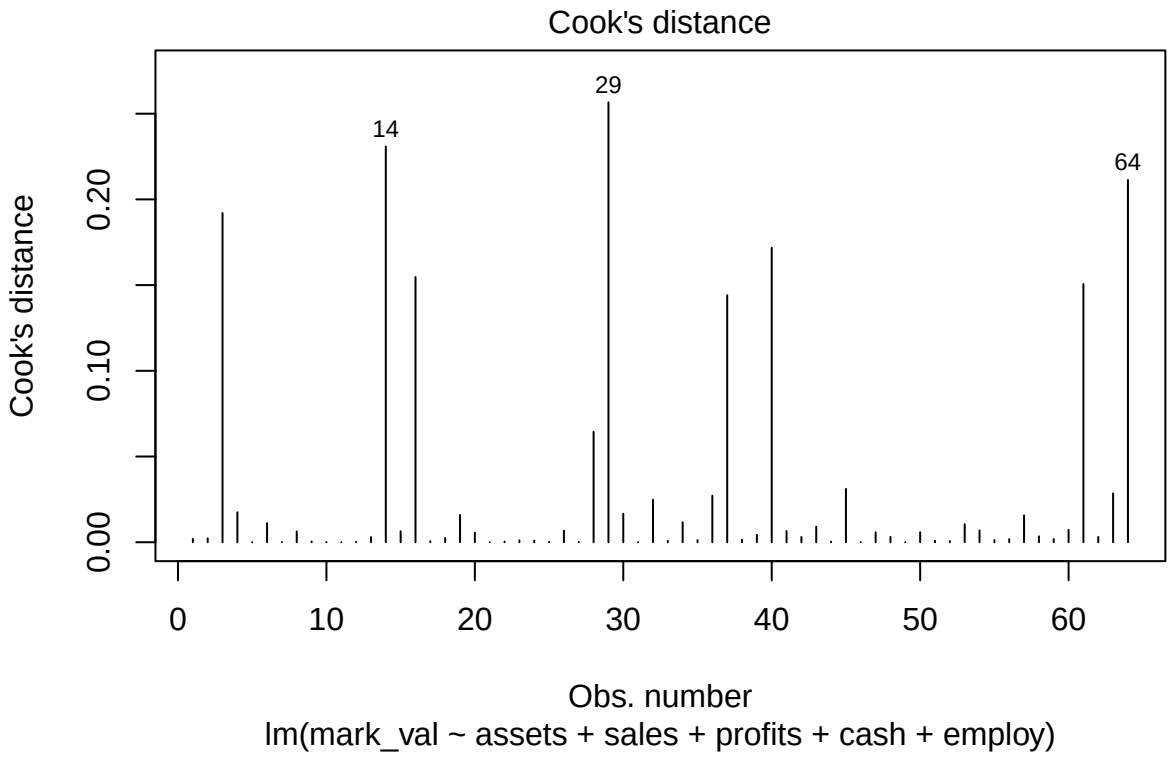

```
# DFITTS
plot(dffits(m1), ylab="DFFITS")
text(dffits(m1))
abline(h=2*sqrt(k/n))
abline(h=-2*sqrt(k/n))
```




```
# DFBETAS
dfbetasPlots(m1)
```




```
# COOK
plot(m1, which=4)
```



I test ci identificano diversi outlier. Facendo una sintesi di tutti i test decidiamo di eliminare le osservazioni 3, 14, 16, 29, 37, 40, 61 e 64.

```
data2 = data[-c(3, 14, 16, 29, 37, 40, 61, 64), ]
```

Andiamo a ristimare il modello.

```
m1<-lm(mark_val~assets+sales+profits+cash+employ,data2)
summary(m1)
```

```
##
## Call:
## lm(formula = mark_val ~ assets + sales + profits + cash + employ,
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -577.71 -178.00  -45.03  190.91  689.38
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 268.782130  77.223277   3.481  0.00105 **
## assets       -0.008898   0.015222  -0.585  0.56146
## sales        -0.096933   0.064520  -1.502  0.13929
## profits       0.421340   1.146367   0.368  0.71476
```

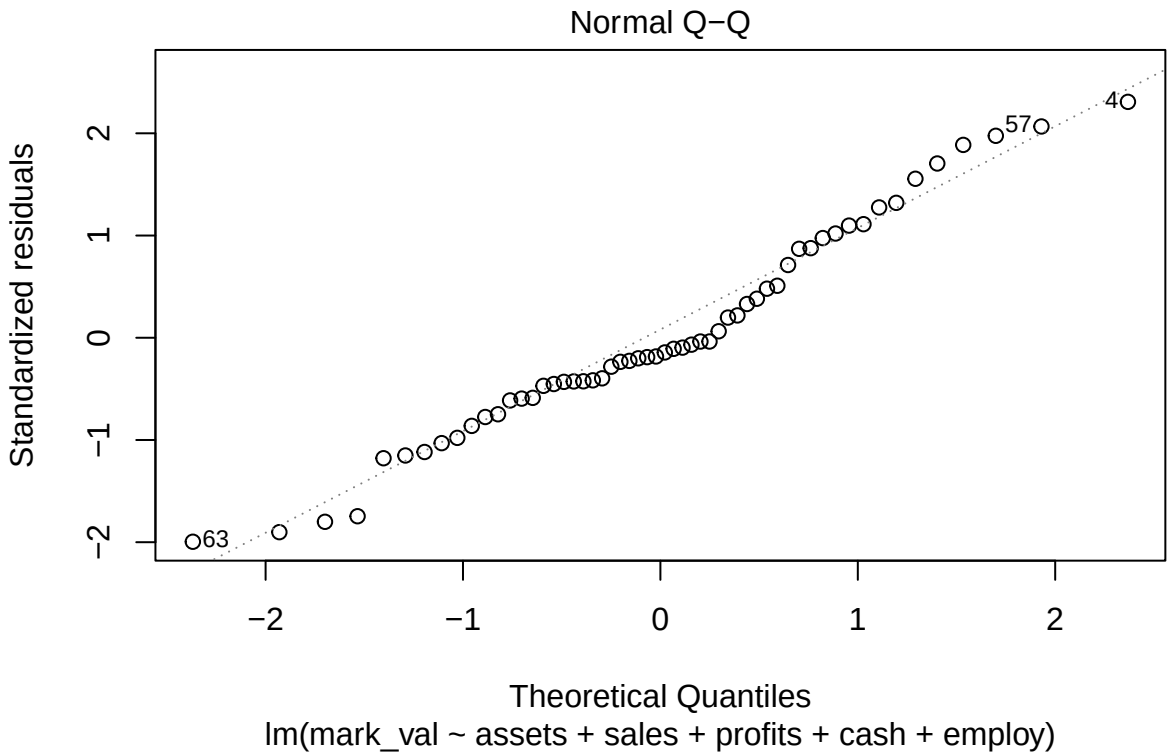
```
## cash          4.182202   0.889733   4.701 2.07e-05 ***
## employ       12.675564   5.032814   2.519  0.01503 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 308.3 on 50 degrees of freedom
## Multiple R-squared:  0.8035, Adjusted R-squared:  0.7839
## F-statistic: 40.89 on 5 and 50 DF,  p-value: < 2.2e-16
```

Come prima, risultano significative cash, employ e viene rifiutata l'ipotesi nulla di non significatività del modello. Tuttavia, l'R quadro e l'R quadro aggiustato sono sensibilmente aumentati (circa 0.07 in più). Il modello senza outlier riesce a spiegare una porzione maggiore della variabilità.

Normalità

Adesso visualizziamo il Normal Q-Q Plot per verificare la normalità.

```
plot(m1, which=2)
```



La distribuzione si discosta leggermente dall'andamento della normale nella parte centrale. Potrebbe esserci non normalità dei residui. Andiamo a vedere coi relativi test.

```
ols_test_normality(m1)
```

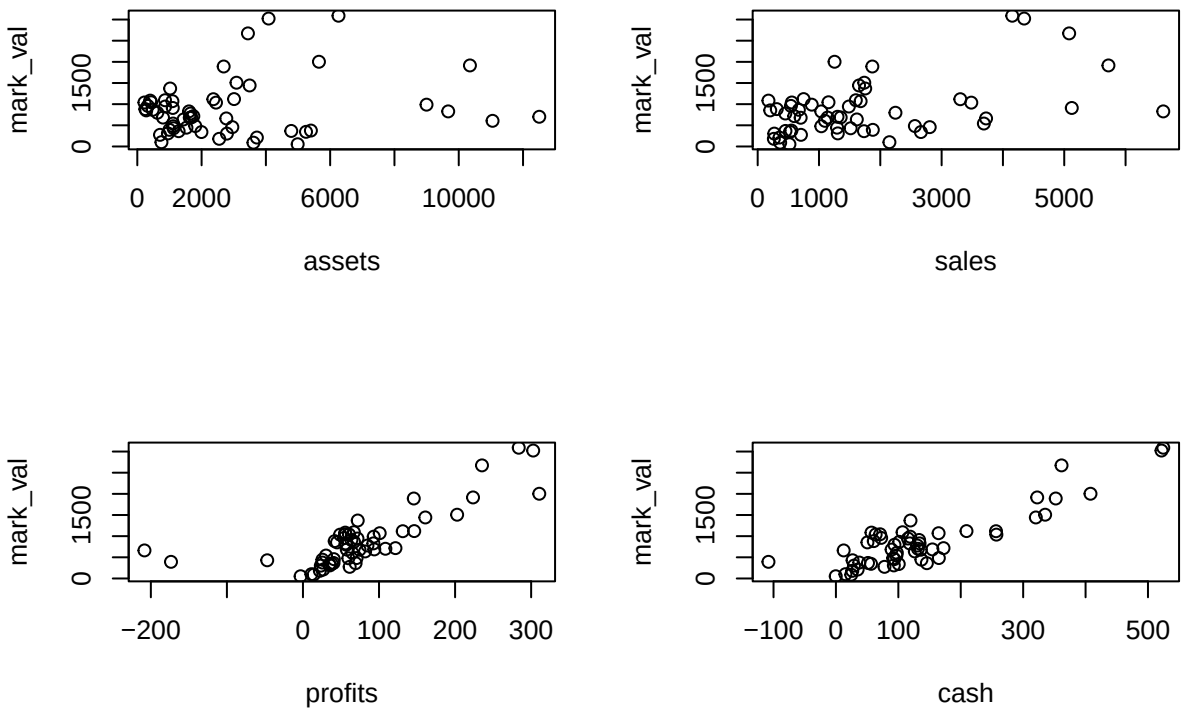
```
## ------------------------------------------------
##        Test              Statistic       pvalue
## ------------------------------------------------
## Shapiro-Wilk               0.9767         0.3477
## Kolmogorov-Smirnov         0.1212         0.3544
## Cramer-von Mises           5.3095         0.0000
## Anderson-Darling           0.5099         0.1894
## ------------------------------------------------
```

Dai test risulta la normalità dei residui e l'ipotesi è pertanto verificata.

Eteroschedasticità

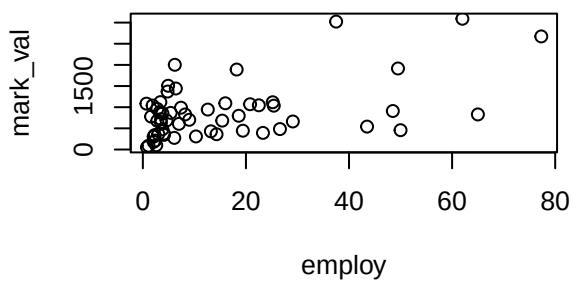
Verifichiamo ora se c'è omoschedasticità o meno dei residui, tramite i seguenti grafici.

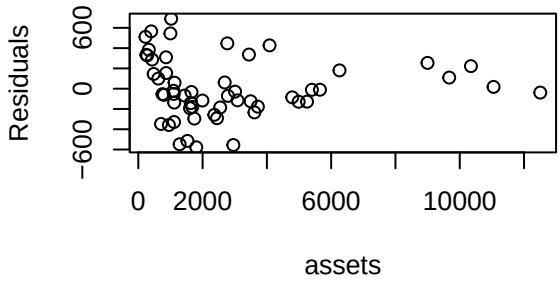
```
# Valori osservati regressore x vs. variabile dipendente y
var_espl = c("assets", "sales", "profits", "cash", "employ")
par(mfrow=c(2,2))
for (i in var_espl){
  plot(data2[,i], data2$mark_val, xlab=i, ylab="mark_val")
}
```

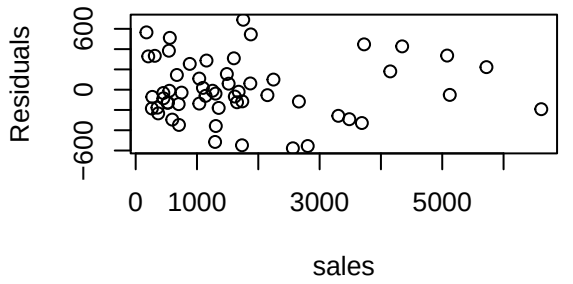


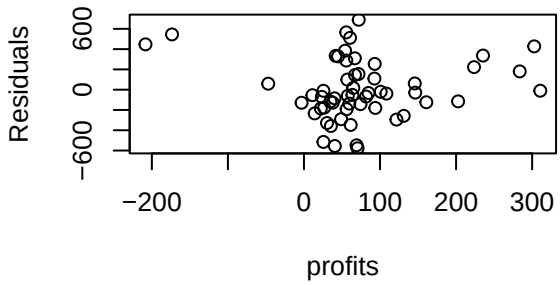

```
# Residui regressore x vs. variabile dipendente y
for (i in var_espl){
```

```
  plot(data2[,i],  m1$residuals, xlab=i, ylab="Residuals")
}
```

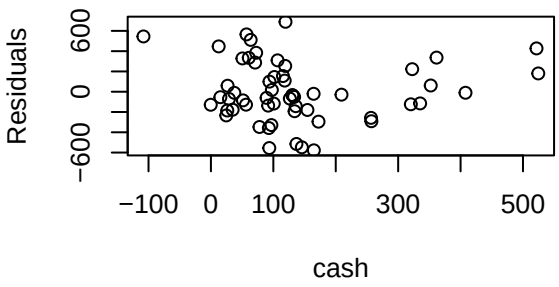


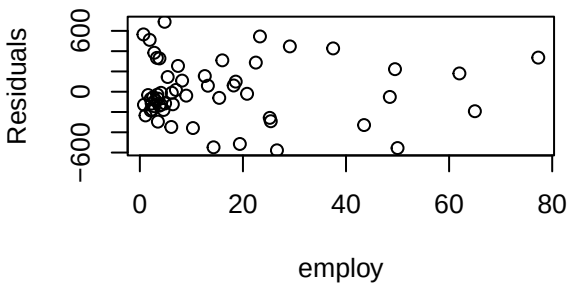




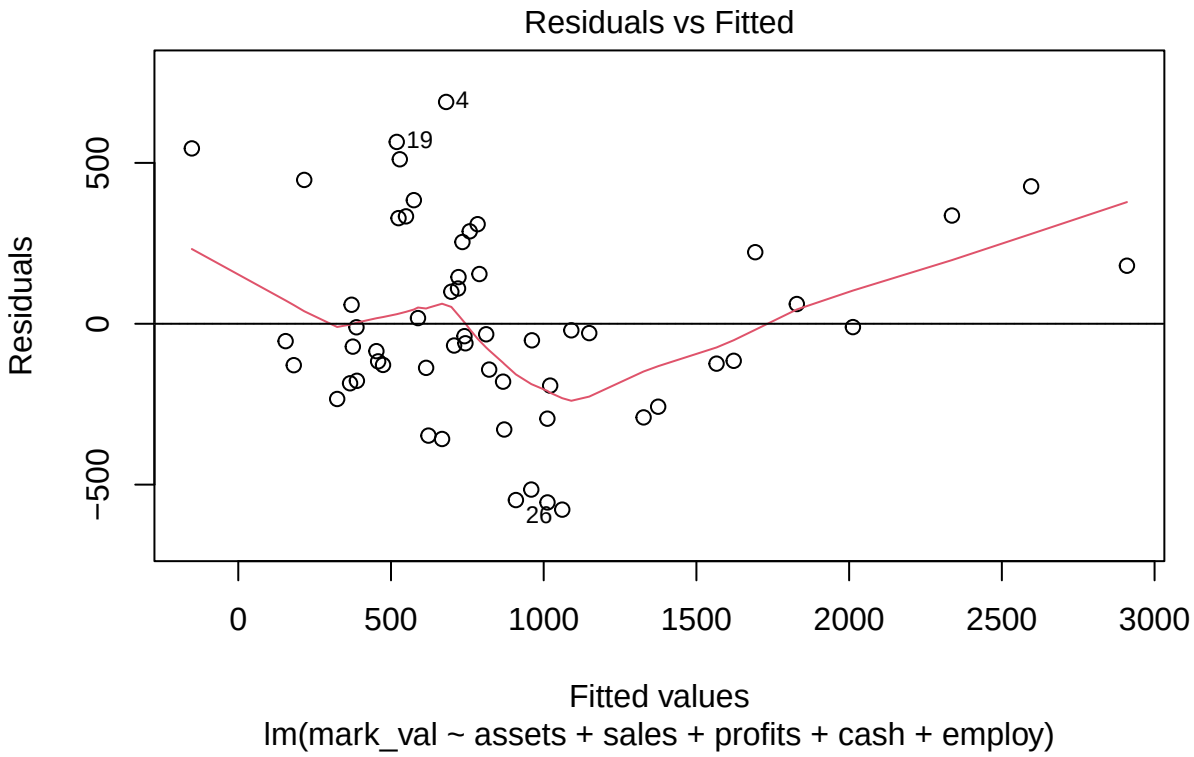



```
par(mfrow=c(1,1))
```

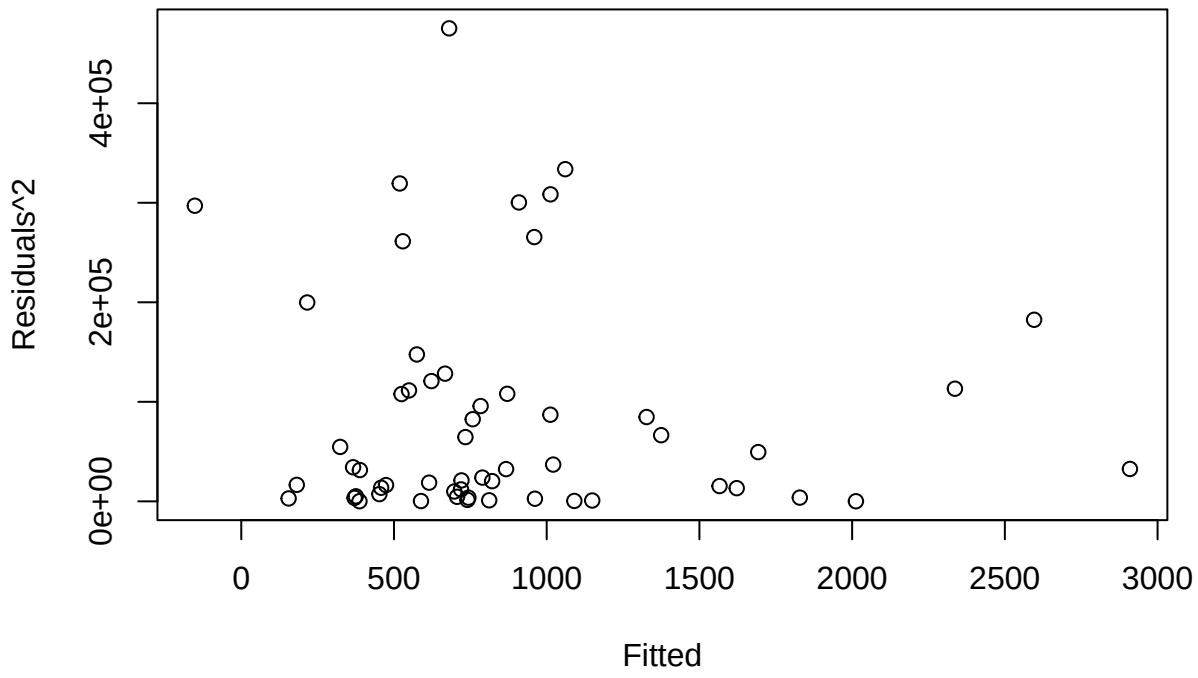





```r
# Residui stimati vs. valori predetti
plot(m1, which=1)
abline(h=0)
```




```
# Residui stimati al quadrato vs. valori predetti
plot(m1$fitted, (m1$residuals)^2, xlab="Fitted", ylab="Residuals^2")
```



Dai grafici osserviamo un'elevata eteroschedasticità. Verifichiamo la nostra ipotesi tramite gli appositi test. di White e Breusch-Pagan.

```
white.test(m1)
```

```
##   Test.Statistic         P
## 1      0.3631767 0.8339445
```

```
ols_test_breusch_pagan(m1)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                 Data
##  -------------------------------------
##  Response : mark_val
##  Variables: fitted values of mark_val
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    0.171846
##  Prob > Chi2   =    0.6784764
```

Accettiamo l'ipotesi di omoschedasticità. Tuttavia, proviamo a migliorare tramite metodo FGLS.

FGLS

Effettuo anzitutto la regressione sui residui al quadrato e calcolo la deviazione standard. Poiché potrei avere varianze negative, applico il logaritmo e poi esponenzio.

```
aux<-lm(I(log(m1$residuals^2))~assets+sales+profits+cash+employ,data2)
summary(aux)
```

```
##
## Call:
## lm(formula = I(log(m1$residuals^2)) ~ assets + sales + profits +
##     cash + employ, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5920 -1.0546  0.3133  1.2375  3.2024
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.055e+01  4.827e-01  21.868   <2e-16 ***
## assets      -2.230e-04  9.514e-05  -2.344   0.0231 *
## sales       -3.150e-04  4.033e-04  -0.781   0.4385
## profits     -8.739e-03  7.165e-03  -1.220   0.2283
## cash         3.994e-03  5.561e-03   0.718   0.4760
## employ       5.144e-02  3.146e-02   1.635   0.1082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.927 on 50 degrees of freedom
## Multiple R-squared:  0.2283, Adjusted R-squared:  0.1512
## F-statistic: 2.959 on 5 and 50 DF,  p-value: 0.02042
```

```
data2$var = exp(aux$fitted.values)
data2$sd<-sqrt(data2$var)
```

Ora trasformo le variabili, dividendo per la deviazione standard.

```
fgls<-lm(I(mark_val/sd)~0+I(1/sd)+I(assets/sd)+I(sales/sd)+I(profits/sd)+I(cash/sd)+I(employ/sd), data2
summary(fgls)
```

```
##
## Call:
## lm(formula = I(mark_val/sd) ~ 0 + I(1/sd) + I(assets/sd) + I(sales/sd) +
##     I(profits/sd) + I(cash/sd) + I(employ/sd), data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8792 -1.2131 -0.3998  0.8379  5.0565
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## I(1/sd)       240.111263  64.335763   3.732 0.000486 ***
## I(assets/sd)   -0.010403   0.007428  -1.400 0.167548
## I(sales/sd)    -0.088542   0.054668  -1.620 0.111606
## I(profits/sd)   2.293804   1.309267   1.752 0.085909 .
## I(cash/sd)      2.920341   0.956240   3.054 0.003614 **
## I(employ/sd)   16.402613   5.260451   3.118 0.003017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.762 on 50 degrees of freedom
## Multiple R-squared:   0.97,  Adjusted R-squared:  0.9664
## F-statistic: 269.3 on 6 and 50 DF,  p-value: < 2.2e-16
```

L'R quadro e l'R quadro aggiustato sono sensibilmente aumentati. Inoltre, ora sono significative cash, employ e profits. Ritentiamo col test di White.

```
white.test(fgls)
```

```
##   Test.Statistic         P
## 1      0.2726916 0.8725409
```

Naturalmente otteniamo errori omoschedastici.

Errori HC

Utilizziamo un approccio alternativo che garantisce errori standard affidabili e in cui la matrice di varianze e covarianze viene costruita (diagonale) e non assumiamo di conoscerla a priori. In pratica si correggono i residui coi gradi di libertà del modello, ottenendo standard error diversi dai precedenti.

```
coeftest(m1, vcov= vcovHC(m1, type="HC2"))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) 268.7821296  84.3327582  3.1872 0.002478 **
## assets       -0.0088982   0.0114993 -0.7738 0.442689
## sales        -0.0969333   0.0635867 -1.5244 0.133703
## profits       0.4213405   1.7689553  0.2382 0.812710
## cash          4.1822017   1.0812125  3.8681 0.000318 ***
## employ       12.6755643   6.0832506  2.0837 0.042321 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(m1, vcov= vcovHC(m1, type="HC3"))
```

```
##
## t test of coefficients:
##
```

```
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) 268.7821296  89.8910782  2.9901 0.004319 **
## assets       -0.0088982   0.0130592 -0.6814 0.498778
## sales        -0.0969333   0.0707268 -1.3705 0.176643
## profits       0.4213405   2.3561529  0.1788 0.858797
## cash          4.1822017   1.4034309  2.9800 0.004441 **
## employ       12.6755643   6.8282980  1.8563 0.069304 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In questo caso abbiano usato errori HC2 e HC3.

Modello finale

Testiamo la normalità sul modello finale fgls.

```
ols_test_normality(fgls)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9517         0.0253
## Kolmogorov-Smirnov        0.1195         0.3707
## Cramer-von Mises          4.8889         0.0000
## Anderson-Darling          0.8291         0.0304
## -----------------------------------------------
```

Poiché Shapiro-Wilk è molto sensibile agli outlier, consideriamo significativo il test di Kolmogorov-Smirnov che ci porta ad accettare l'ipotesi di normalità.