GLS 2 - Data set: COMPANIES

INTRODUZIONE

I dati contengono alcune informazioni riguardanti 64 compagnie. Le variabili presenti nel dataset sono:

1. ASSETS: attivo in bilancio (milioni di dollari)
2. SALES: fatturato relativo alle vendite
3. MARK_VAL: valore di mercato della compagnia (milioni di dollari)
4. PROFITS: profitto (milioni di dollari)
5. CASH: flusso di cassa
6. EMPLOY: numero complessivo di dipendenti
7. SECTOR: settore di mercato in cui opera la compagnia (comunicazioni, energia, finanza, hitech, manufatturiero, medico, retail, trasporti, altro)

Analisi proposte:

1. Statistiche descrittive
2. Regressione
3. Studio dell'autocorrelazione

```r
#-- R CODE

library(pander)
library(car)
library(olsrr)
library(systemfit)
library(het.test)
panderOptions('knitr.auto.asis', FALSE)

#-- White test function
white.test <- function(lmod,data=d){
  u2 <- lmod$residuals^2
  y <- fitted(lmod)
  Ru2 <- summary(lm(u2 ~ y + I(y^2)))$r.squared
  LM <- nrow(data)*Ru2
  p.value <- 1-pchisq(LM, 2)
  data.frame("Test statistic"=LM,"P value"=p.value)
}

#-- funzione per ottenere osservazioni outlier univariate
FIND_EXTREME_OBSERVARION <- function(x,sd_factor=2){
  which(x>mean(x)+sd_factor*sd(x) | x<mean(x)-sd_factor*sd(x))
}

#-- import dei dati
ABSOLUTE_PATH <- "C:\\Users\\sbarberis\\Dropbox\\MODELLI STATISTICI"
d <- read.csv(paste0(ABSOLUTE_PATH,"\\F. Esercizi(22) copia\\1.Error-GLS copy(8)\\2.Error-GLS\\companies

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("assets","sales","mark_val","profits","cash","employ")
```

```r
#-- print delle prime 6 righe del dataset
pander(head(d),big.mark=",")
```

Table 1: Table continues below

company	assets	sales	mark_val	profits	cash
Air Products	2,687	1,870	1,890	145.7	352.2
American Savings Bank FSB	3,614	367	90	14.1	24.6
AMR	6,425	6,131	2,448	345.8	682.5
Apple Computer	1,022	1,754	1,370	72	119.5
Armstrong World Industries	1,093	1,679	1,070	100.9	164.5
Bally Manufacturing	1,529	1,295	444	25.6	137

employ	sector
18.2	Other
1.1	Finance
49.5	Transportation
4.8	HiTech
20.8	Manufacturing
19.4	Other

STATISTICHE DESCRITTIVE

```r
#-- R CODE
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

Table 3: Table continues below

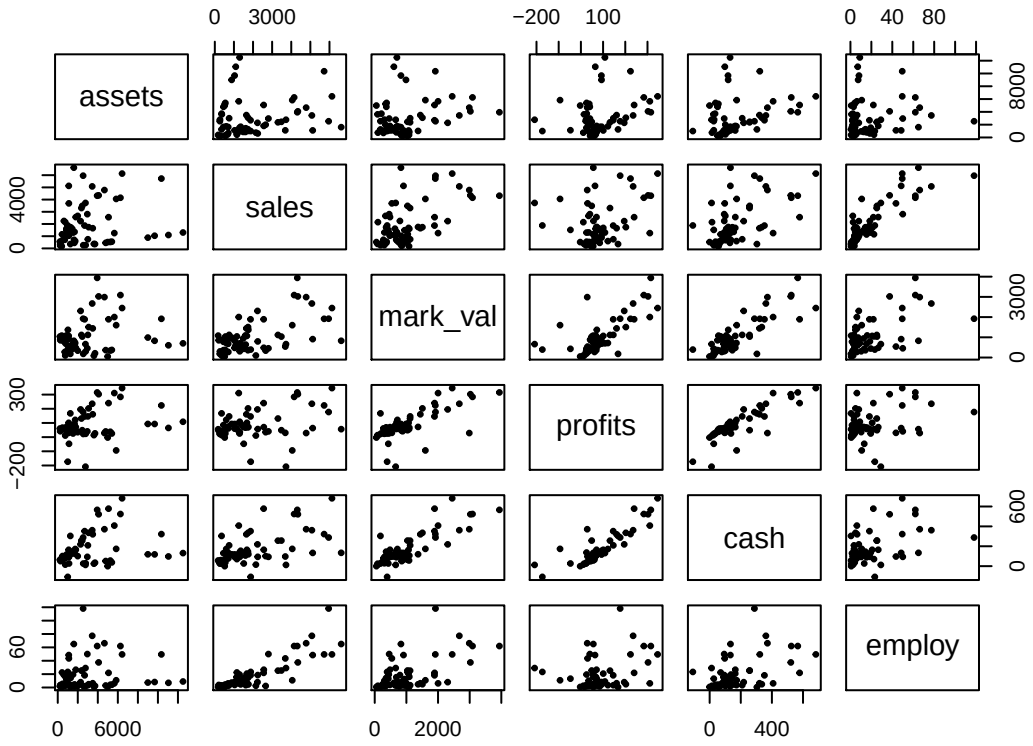
assets	sales	mark_val	profits
Min. : 223	Min. : 176.0	Min. : 53.0	Min. :-208.40
1st Qu.: 1075	1st Qu.: 653.2	1st Qu.: 440.5	1st Qu.: 37.20
Median : 2140	Median :1501.5	Median : 829.0	Median : 64.30
Mean : 2997	Mean :2006.5	Mean :1036.7	Mean : 86.01
3rd Qu.: 3976	3rd Qu.:2698.0	3rd Qu.:1182.5	3rd Qu.: 124.00
Max. :12505	Max. :6615.0	Max. :3940.0	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

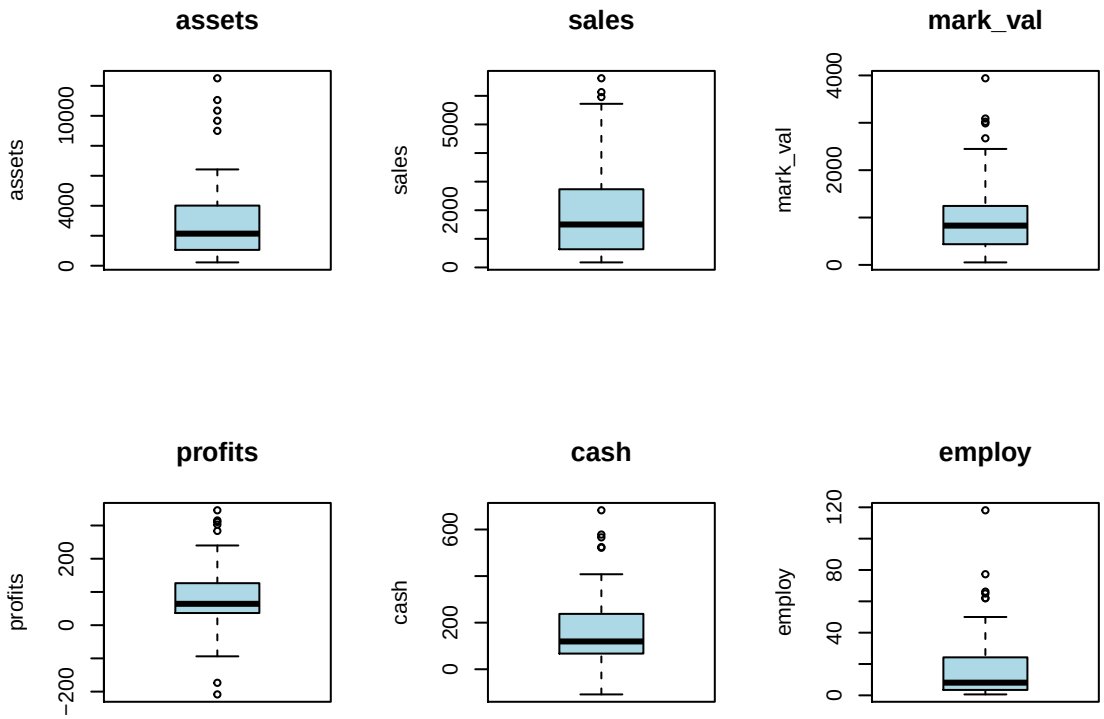
```r
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

	assets	sales	mark_val	profits	cash	employ
assets	1	0.1773	0.2442	0.2831	0.3037	0.1105
sales	0.1773	1	0.5974	0.3173	0.563	0.8635
mark_val	0.2442	0.5974	1	0.6986	0.8354	0.5873
profits	0.2831	0.3173	0.6986	1	0.8556	0.3253
cash	0.3037	0.563	0.8354	0.8556	1	0.492
employ	0.1105	0.8635	0.5873	0.3253	0.492	1

```r
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```r
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```



Non esistono correlazioni elevate tra le variabili.

REGRESSIONE

Si effettua la regressione con variabile dipendente con “mark_val” e variabile esplicativa “assets”, “sales”, “profits”, “cash”, “employ”.

```
#-- R CODE
mod1 <- lm(mark_val ~ assets + sales + profits + cash + employ, d) #-- stima modello lineare semplice
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	243.8	103	2.367	0.02132
assets	0.0002066	0.02127	0.009716	0.9923
sales	-0.01802	0.07445	-0.242	0.8096
profits	0.08828	1.176	0.0751	0.9404
cash	3.787	0.862	4.393	4.819e-05
employ	9.399	4.875	1.928	0.05875

Table 7: Fitting linear model: mark_val ~ assets + sales + profits + cash + employ

Observations	Residual Std. Error	R^2	Adjusted R^2
64	444.8	0.7394	0.7169

```
pander(anova(mod1),big.mark=",")
```

Table 8: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
assets	1	2,624,580	2,624,580	13.27	0.000578
sales	1	13,958,764	13,958,764	70.55	1.315e-11
profits	1	11,824,616	11,824,616	59.77	1.732e-10
cash	1	3,410,595	3,410,595	17.24	0.0001095
employ	1	735,472	735,472	3.717	0.05875
Residuals	58	11,475,057	197,846	NA	NA

Il modello interpreta bene la variabile dipendente ($R^2 = 0.7394$) ma solo "cash" e in parte "employ" hanno associato parametri significativi. Si verifica ora omoschedasticità e incorrelazione.

```
#-- R CODE
pander(white.test(mod1),big.mark=",") #-- white test
```

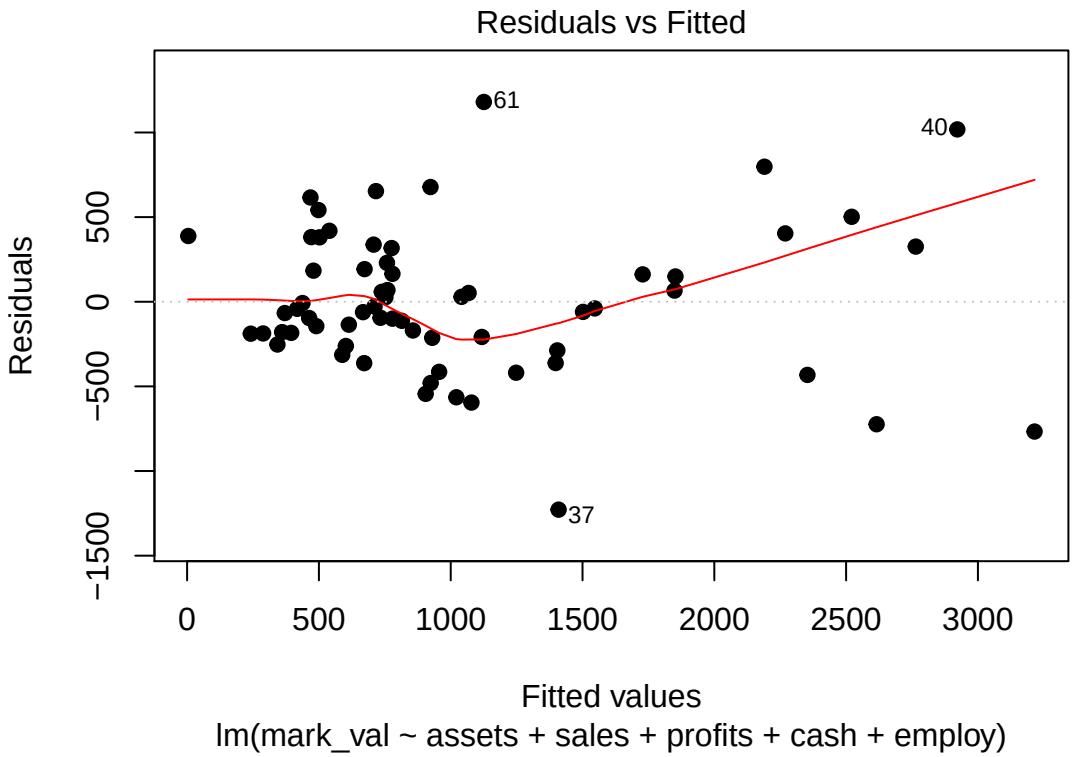
Test.statistic	P.value
9.406	0.009068

```
pander(dwtest(mod1),big.mark=",") #-- Durbin-Whatson test
```

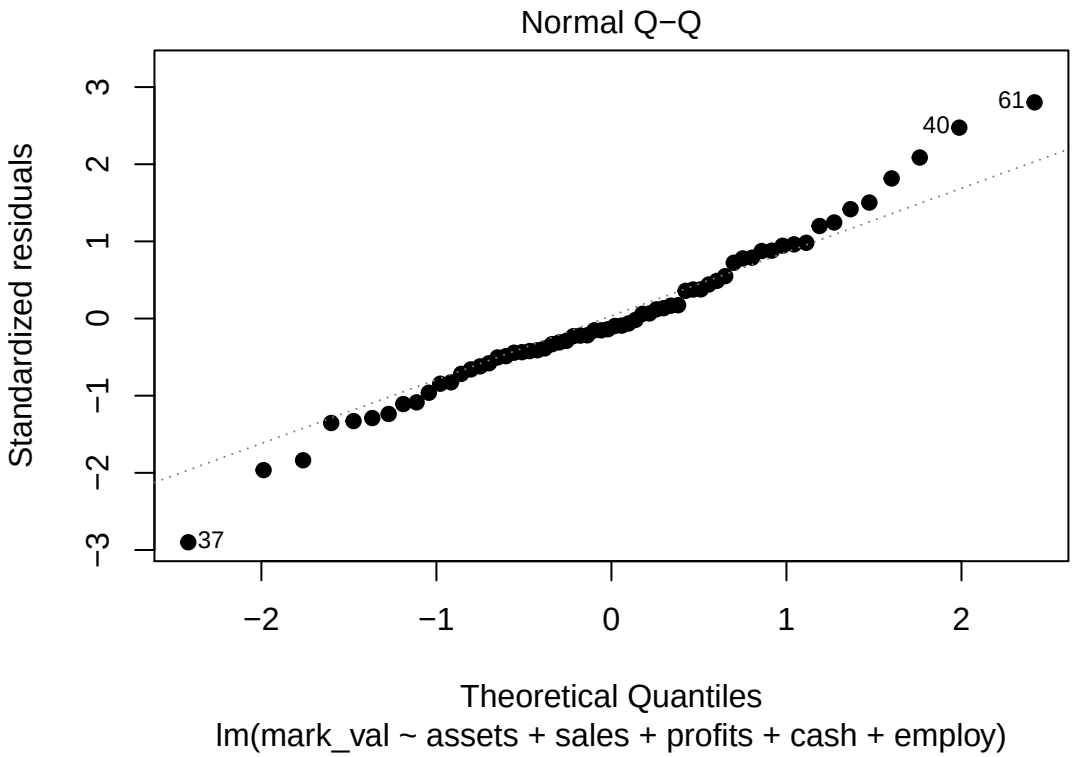
Table 10: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
2.064	0.5685	true autocorrelation is greater than 0

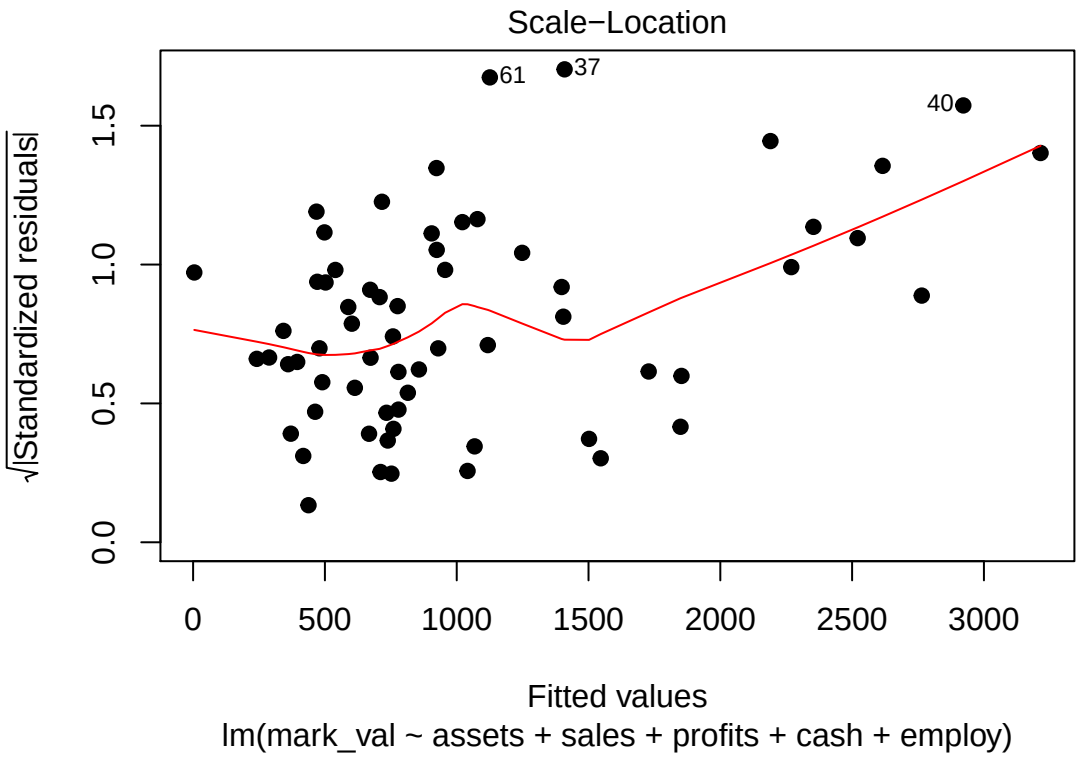
```
#-- R CODE
plot(mod1,which=1,pch=19)
```



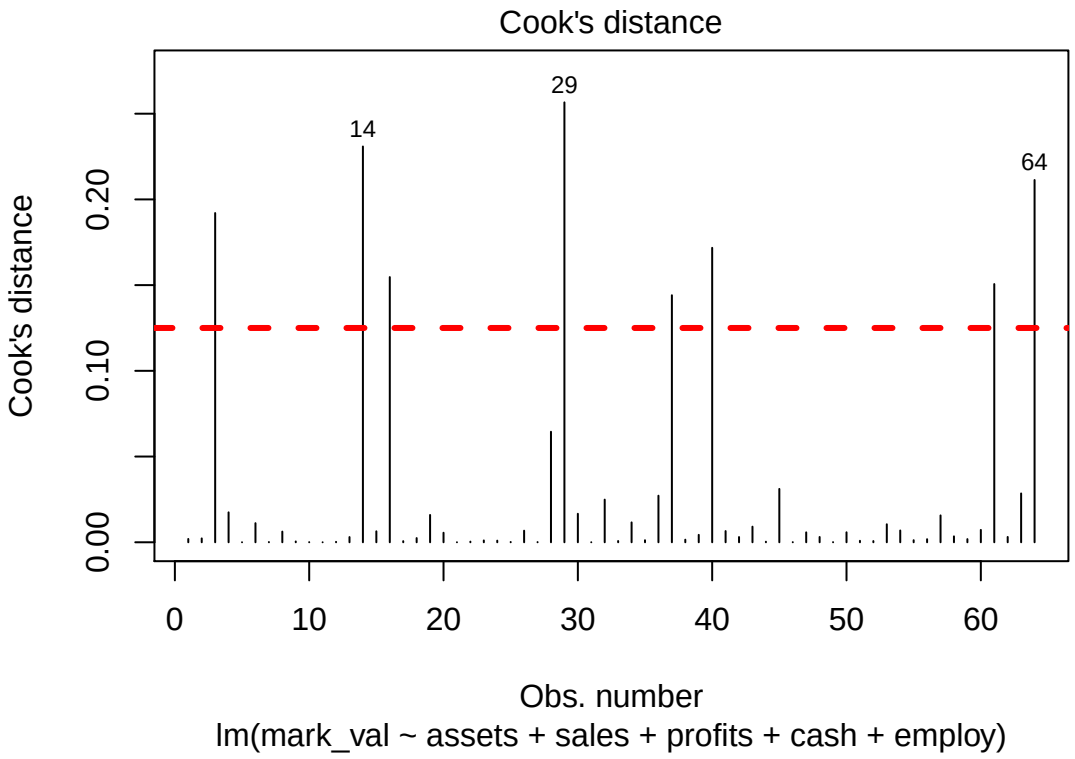

```
plot(mod1,which=2,pch=19)
```



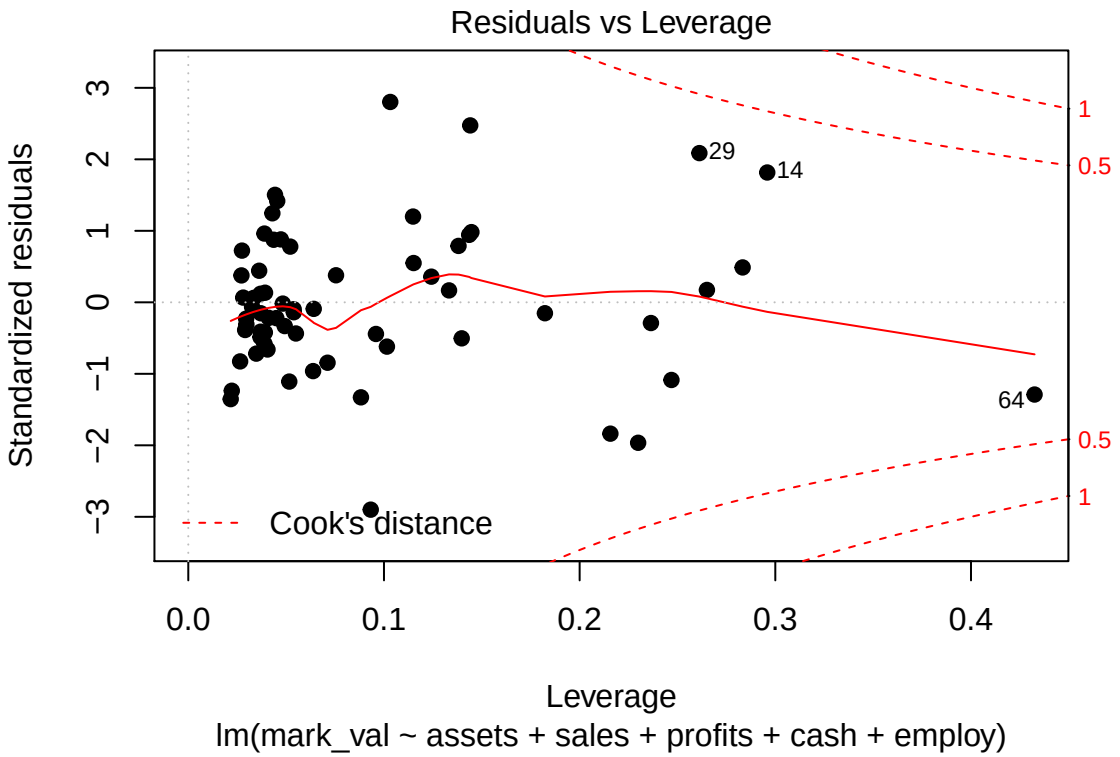

```
plot(mod1,which=3,pch=19)
```




```
plot(mod1,which=4,pch=19)
abline(h=2*4/nrow(d),col=2,lwd=3,lty=2)
```




```
plot(mod1,which=5,pch=19)
```



I grafici dei residui hanno una configurazione non regolare che suggerisce eteroschedasticità. Proviamo ora a regredire i residui al quadrato del modello sviluppato precedentemente sui regressori. Si nota che il modello risulta significativo, cosa che non accadrebbe se i residui stessi fossero omoschedastici.

```
#-- R CODE
mod2 <- lm(resid(mod1)^2 ~ assets + sales + profits + cash + employ, d) #-- stima modello lineare sempl
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	103,853	64,252	1.616	0.1114
assets	-21.87	13.26	-1.649	0.1046
sales	-10.26	46.44	-0.221	0.8259
profits	-571.7	733.3	-0.7797	0.4387
cash	1,267	537.7	2.356	0.02185
employ	-243.6	3,041	-0.08012	0.9364

Table 12: Fitting linear model: resid(mod1)^2 ~ assets + sales + profits + cash + employ

Observations	Residual Std. Error	R^2	Adjusted R^2
64	277442	0.209	0.1408

```
pander(anova(mod2),big.mark=",")
```

Table 13: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
assets	1	2.561e+10	2.561e+10	0.3327	0.5663
sales	1	2.742e+11	2.742e+11	3.562	0.06411
profits	1	4.377e+11	4.377e+11	5.686	0.02039
cash	1	4.415e+11	4.415e+11	5.736	0.01987
employ	1	494,071,534	494,071,534	0.006419	0.9364
Residuals	58	4.464e+12	7.697e+10	NA	NA

Da tale procedimento si calcola il valore della varianza dei residui che si ottiene dai valori previsti della regressione dei residui al quadrato rispetto ai regressori. La varianza dei residui sarà la varianza di tali valori previsti.

```
#-- R CODE
var(fitted(mod2))
```

[1] 18722296190

```
sd(fitted(mod2))
```

[1] 136829.4

Il test Durbin Watson non respinge l'ipotesi di non correlazione fra gli errori.

```
#-- R CODE
pander(white.test(mod1),big.mark=",") #-- white test
```

Test.statistic	P.value
9.406	0.009068

```
pander(dwtest(mod2),big.mark=",") #-- Durbin-Whatson test
```

Table 15: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
2.176	0.74	true autocorrelation is greater than 0

Per eliminare l'eteroschedasticità si propone un modello lineare basato su FGLS-WLS avendo costruito errori omoschedastici in cui tutte le variabili e gli errori sono divisi per il reciproco dello scarto quadratico medio della varianza dei valori previsti.

Approfondimento: implementazione del metodo FGLS - Feasible Generalized Least Squares

Si tratta di considerare l'eteroschedasticita' nel processo di stima; i passi principali sono i seguenti:

1. Si stima il modello di regressione utilizzando il metodo OLS (Ordinary Least Squares). In questo modo

si ottengono i residui OLS

2. Si assume un modello che descriva la varianza degli errori in funzione dei regressori (ad esempio, una forma lineare, quadratica, logaritmica, ecc.). Da un punto di vista operativo, cio' equivale ad assumere una relazione fra i residui al quadrato e i regressori. Quindi, si stimano i parametri del modello cosi' specificato
3. Si utilizza il modello stimato al punto 2 per ottenere i valori previsti dei residui al quadrato, che rappresentano i valori previsti della varianza.
4. In luogo delle variabili di origine $Y, X_1, X_2, ..., X_k$ si considerano le nuove variabili $\hat{Y}, \hat{X}_1, \hat{X}_2, ..., \hat{X}_k$, ottenute rapportando le variabili di origine ai valori previsti della deviazione standard (ossia, la radice quadrata della varianza)
5. Infine, da un punto di vista operativo si hanno due possibilità. (1) Con il metodo OLS si stima il modello di regressione (senza intercetta) costruito sulle variabili che derivano dal punto 4. (2) Si applica il metodo dei minimi quadrati pesati (Weighted Least Squares) al modello di origine usando come peso il reciproco dei valori stimati della varianza.

Proviamo con la procedura con metodo OLS calcolando quindi le nuove variabili:

```
#-- R CODE

sd_error <- sqrt(fitted(mod2))
```

```
## Warning in sqrt(fitted(mod2)): NaNs produced
```

```
mod3 <- lm(I(mark_val/sd_error) ~ 0 + I(1/sd_error) + I(assets/sd_error) + I(sales/sd_error) + I(profits
pander(summary(mod3),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
I(1/sd_error)	296.5	85.74	3.458	0.001081
I(assets/sd_error)	-0.01998	0.02139	-0.934	0.3546
I(sales/sd_error)	-0.009926	0.06646	-0.1493	0.8818
I(profits/sd_error)	-1.348	1.085	-1.243	0.2194
I(cash/sd_error)	4.63	0.9661	4.792	1.373e-05
I(employ/sd_error)	7.264	4.745	1.531	0.1317

Table 17: Fitting linear model: I(mark_val/sd_error) ~ 0 + I(1/sd_error) + I(assets/sd_error) + I(sales/sd_error) + I(profits/sd_error) + I(cash/sd_error) + I(employ/sd_error)

Observations	Residual Std. Error	R^2	Adjusted R^2
59	1.029	0.8727	0.8583

```
pander(anova(mod3),big.mark=",")
```

Table 18: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(1/sd_error)	1	242.8	242.8	229.2	6.861e-21
I(assets/sd_error)	1	2.385	2.385	2.25	0.1395
I(sales/sd_error)	1	65.22	65.22	61.54	1.98e-10

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(profits/sd_error)	1	48.79	48.79	46.05	9.98e-09
I(cash/sd_error)	1	23.2	23.2	21.89	2.032e-05
I(employ/sd_error)	1	2.484	2.484	2.344	0.1317
Residuals	53	56.16	1.06	NA	NA

Proviamo con la procedura con metodo WLS mantenendo quindi le varibili di origine e applicando un vettore di pesi. Si noti che i pesi negativi non hanno senso, pertanto è necessatio eliminare le relative osservazioni.

```
#-- R CODE
weight <- 1/fitted(mod2)

mod4 <- lm(mark_val ~ assets + sales + profits + cash + employ, d[-which(weight<0),],weights = weight[-w
pander(summary(mod4),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	296.5	85.74	3.458	0.001081
assets	-0.01998	0.02139	-0.934	0.3546
sales	-0.009926	0.06646	-0.1493	0.8818
profits	-1.348	1.085	-1.243	0.2194
cash	4.63	0.9661	4.792	1.373e-05
employ	7.264	4.745	1.531	0.1317

Table 20: Fitting linear model: mark_val ~ assets + sales + profits + cash + employ

Observations	Residual Std. Error	R^2	Adjusted R^2
59	1.029	0.7167	0.69

```
pander(anova(mod4),big.mark=",")
```

Table 21: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
assets	1	2.385	2.385	2.25	0.1395
sales	1	65.22	65.22	61.54	1.98e-10
profits	1	48.79	48.79	46.05	9.98e-09
cash	1	23.2	23.2	21.89	2.032e-05
employ	1	2.484	2.484	2.344	0.1317
Residuals	53	56.16	1.06	NA	NA

Commentiamo i risultati per modello 3 (mod3). Il modello risulta ancora significativo, anzi il fitting migliora (R2=0.8727) e solo "cash" risulta significativa.

Inoltre si può verificare dal grafico residui-predetti e dai test di White e Durbin Watson che gli errori sono ora omoschedastici e incorrelati.

```r
#-- R CODE
pander(white.test(mod3),big.mark=",")
```

Test.statistic	P.value
0.07808	0.9617

```r
pander(dwtest(mod3),big.mark=",")
```

Table 23: Durbin-Watson test: `mod3`

Test statistic	P value	Alternative hypothesis
2.171	0.719	true autocorrelation is greater than 0

Tuttavia si vede che il modello non usa tutte le osservazioni perché alcune delle stime delle suddette osservazioni hanno varianze negative per limiti computazionali del programma FGLS. Si propone quindi una nuova stima del modello basata su esponenziale FGLS che per proprietà della funzione esponenziale non può avere stime con varianze negative.

```r
#-- R CODE
mod2 <- lm(log(resid(mod1)^2) ~ assets + sales + profits + cash + employ, d)

sd_error <- sqrt(exp(fitted(mod2)))

mod5 <- lm(I(mark_val/sd_error) ~ 0 + I(1/sd_error) + I(assets/sd_error) + I(sales/sd_error) + I(profits
pander(summary(mod5),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
I(1/sd_error)	243.7	77.77	3.134	0.002706
I(assets/sd_error)	-0.01072	0.009896	-1.083	0.2832
I(sales/sd_error)	0.05157	0.07408	0.6961	0.4891
I(profits/sd_error)	1.719	1.339	1.283	0.2044
I(cash/sd_error)	2.838	1.067	2.659	0.01011
I(employ/sd_error)	4.518	5.938	0.7609	0.4498

Table 25: Fitting linear model: I(mark_val/sd_error) ~ 0 + I(1/sd_error) + I(assets/sd_error) + I(sales/sd_error) + I(profits/sd_error) + I(cash/sd_error) + I(employ/sd_error)

Observations	Residual Std. Error	R^2	Adjusted R^2
64	1.834	0.8922	0.8811

```r
pander(anova(mod5),big.mark=",")
```

Table 26: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(1/sd_error)	1	1,195	1,195	355.3	2.071e-26
I(assets/sd_error)	1	1.844	1.844	0.5484	0.462
I(sales/sd_error)	1	177.3	177.3	52.71	1.071e-09
I(profits/sd_error)	1	216.2	216.2	64.28	5.724e-11
I(cash/sd_error)	1	22.88	22.88	6.803	0.01156
I(employ/sd_error)	1	1.948	1.948	0.579	0.4498
Residuals	58	195.1	3.364	NA	NA

```
pander(white.test(mod5),big.mark=",")
```

Test.statistic	P.value
0.175	0.9162

```
pander(dwtest(mod5),big.mark=",")
```

Table 28: Durbin-Watson test: `mod5`

Test statistic	P value	Alternative hypothesis
2.047	0.5524	true autocorrelation is greater than 0

ll modello ora usa tutte le osservazioni, migliora il fitting e si conferma come unica variabile con parametro significativo “cash”.