La Franca Statistical Modeling

Marta La Franca (Matr. 866590)

29/4/2021

- Analisi dei dati:
- Statistiche descrittive
- Modello lineare
    - Multicollinearità
    - Omoschedasticità
    - Normalità dei residui
    - Outlier
- Autocorrelazione
    - Risoluzione

Vengono caricate le librerie utili ai fini dello svolgimento del problema in esame

```
library(car)
library(olsrr)
library(skedastic)
library(psych)
library(lmtest)
library(systemfit)
library(sandwich)
library(describedata)
library(klaR)
library(DataCombine)
library(pander)
```

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

```
library(lmtest)
```

Si carica la funzione white.test che verrà utilizzata durante lo studio dell'eteroschedasticità:

```
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)
}
```

Analisi dei dati:

Si carica il dataset d'interesse:

```
data<-read.csv("C:/Users/marta/Downloads/sm_esame290421.csv")
```

Si effettua un print delle prime 6 righe del dataset:

```
pander(head(data),big.mark=",")
```

time	x1	x2	x3	y
192	6.132	-0.7008	50.94	20.68
122	14.26	1.254	331	22.85
28	6.296	0.07431	79.15	23.23
94	18.65	2.415	724.9	28.13
57	7.284	-0.6889	63.89	23.7
185	11.78	-1.922	293.9	21.16

Si considerano le variabili numeriche:

```
var_num<-c("x1","x2","x3","y")
```

Si ordinano i dati secondo la variabile “time”:

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

Statistiche descrittive

```
pander(summary(data[,var_num]))
```

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

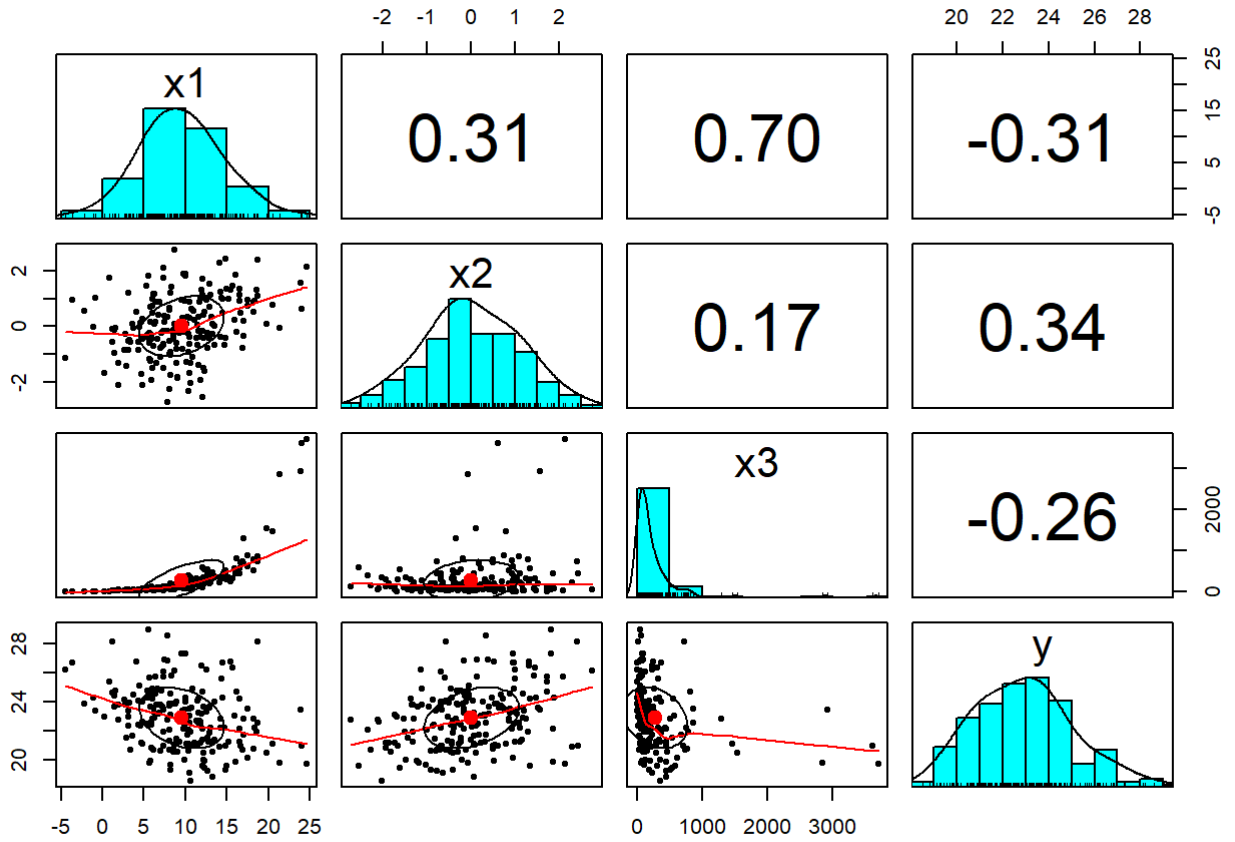
E’ possibile osservare che: le variabili presentano range diversi tra loro; la variabile $x1$ presenta un range da -4.440 a 24.603, mentre la variabile $x3$ presenta un range da 3.281 a 3699.064.

Si può dedurre che nessuna variabile presenta dei valori nulli e/o anomali, e.g. un valore −999 che identifica missing value.

La variabile $x2$ presenta una media di 0.01307; la variabile $x3$ di 276.579.

Si descrivono le correlazioni delle variabili numeriche: si considerano in particolare istogrammi, scatter plot e il coefficiente di correlazio delle coppie di variabili

```
pairs.panels(data[,-1])
```



E’ possibile osservare che: la maggior correlazione è presente tra le variabili $x1$ e $x3$.

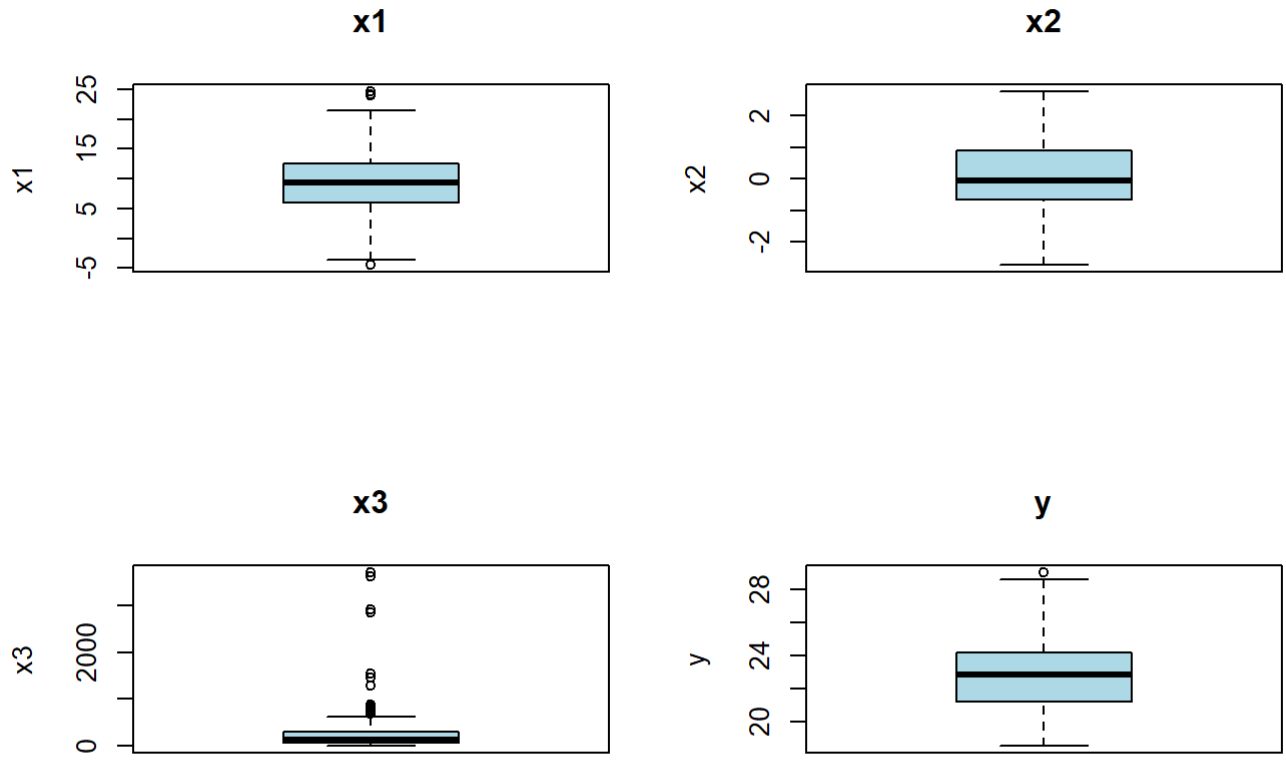
La variabile $x3$ presenta una coda a destra, mentre per le altre variabili non si distinguono particolari andamenti.

Tra le variabili $x3$ e y e le variabili $x1$ e y è presente una correlazione negativa, sebbene essa sia di lieve entità.

Non sembrano esserci dunque particolari correlazioni tra nessuna delle variabili considerate, solitamente infatti da questa matrice di correlazione le variabili che rappresentano delle problematiche sono considerate quelle con indici di correlazione superiori a 0,90.

Si effettuano i boxplot delle variabili numeriche:

```
par(mfrow=c(2,2))
for(i in var_num){
  boxplot(data[,i],main=i,col="light blue",ylab=i)
}
```



Dall'analisi del boxplot si osserva che la variabile $x3$ presenta svariati outliers, e non presenta una distribuzione normale.

Per quanto riguarda le altre variabili numeriche invece si identifica una distribuzione quasinormale (a occhio) che risulta essere un buon punto di partenza, in quanto permettono di assumere le ipotesi del Teorema del Limite Centrale senza troppi problemi.

Modello lineare

Viene adesso svolto il modello lineare di y (variabile dipendente) su x1, x2 e log(x3) (variabili indipendenti): in particolare il test di ipotesi ed interpretazione dei coefficienti.

```
mod1<-lm(y~x1+x2+I(log(x3)),data)
```

```
pander(summary(mod1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.07	1.511	17.91	3.893e-43
x1	0.02248	0.1366	0.1646	0.8694
x2	0.7037	0.1946	3.617	0.0003798
I(log(x3))	-0.8944	0.5645	-1.584	0.1147

Fitting linear model: y ~ x1 + x2 + I(log(x3)) Effettuando la summary del modello si osserva che il modello risulta essere statisticamente significativo (presenta un p-value pari a 3.162e-16).

Observations	Residual Std. Error	R^2	Adjusted R^2
200	1.748	0.3183	0.3078

La variabile $x2$ è l'unica variabile, oltre l'intercetta ad essere statisticamente significativa ed ha un impatto positivo sulla y.

La variabile log(x3) invece ha un impatto negativo rispetto la y.

Il coefficiente R^2 risulta discretamente basso (0.3183); il modello ha dunque una capacità di adattamento bassa, ma non risulta anomalo in quanto solamente una variabile risulta statisticamente significativa.

Multicollinearità

Si effettua lo studio della multicollinearità utilizzando l'indice VIF e il condition index:

```
pander(vif(mod1))
```

x1	x2	I(log(x3))
31.75	2.874	28.81

```
pander(ols_eigen_cindex(mod1))
```

Eigenvalue	Condition Index	intercept	x1	x2	I(log(x3))
2.885	1	0.000754	0.0007927	0.0004438	0.0002257
1.007	1.692	6.178e-05	3.105e-05	0.3387	1.019e-05
0.1061	5.214	0.0307	0.03294	0.04373	9.466e-08
0.001229	48.45	0.9685	0.9662	0.6171	0.9998

Analizzando i risultati si osserva che la variabile $x1$ e la variabile log($x3$) presentano dei VIF elevati; sono infatti considerati valori anomali dei valori maggiori di 10.

Inoltre considerando il condition index, l'autovalore caratterizzato da un condition index più elevato spiega il 96% della varianza di $x1$ e il 99% della varianza di log($x3$).

Dunque è possibile dire che vi sia collinearità tra le due variabili.

Si decide dunque di proseguire nel modello eliminando una delle due variabili, ed in particolare si decide di eliminare la variabile log($x3$)

Si definisce un secondo modello:

```
mod2<-lm(y~x1+x2,data)
```

```
pander(summary(mod2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.71	0.2735	90.37	2.574e-162
x1	-0.1901	0.02562	-7.422	3.376e-12

	Estimate	Std. Error	t value	Pr(>|t|)
x2	0.9454	0.1213	7.794	3.653e-13

Fitting linear model: y ~ x1 + x2

Observations	Residual Std. Error	R^2	Adjusted R^2
200	1.754	0.3095	0.3025

Il modello così composto risulta, come prima, statisticamente significativo.

In particolare adesso tutte le variabili sono statisticamente significative, come ci si poteva aspettare avendo eliminato la collinearità.

Il coefficiente R^2 invece risulta discretamente basso come in precedenza.

Si controlla la collinearità per verificare che si siano risolti i problemi:

```
pander(vif(mod2))
```

x1	x2
1.108	1.108

```
pander(ols_eigen_cindex(mod2))
```

Eigenvalue	Condition Index	intercept	x1	x2
1.898	1	0.0524	0.05272	0.008492
0.9958	1.381	0.006423	3.255e-05	0.8778
0.1061	4.229	0.9412	0.9472	0.1137

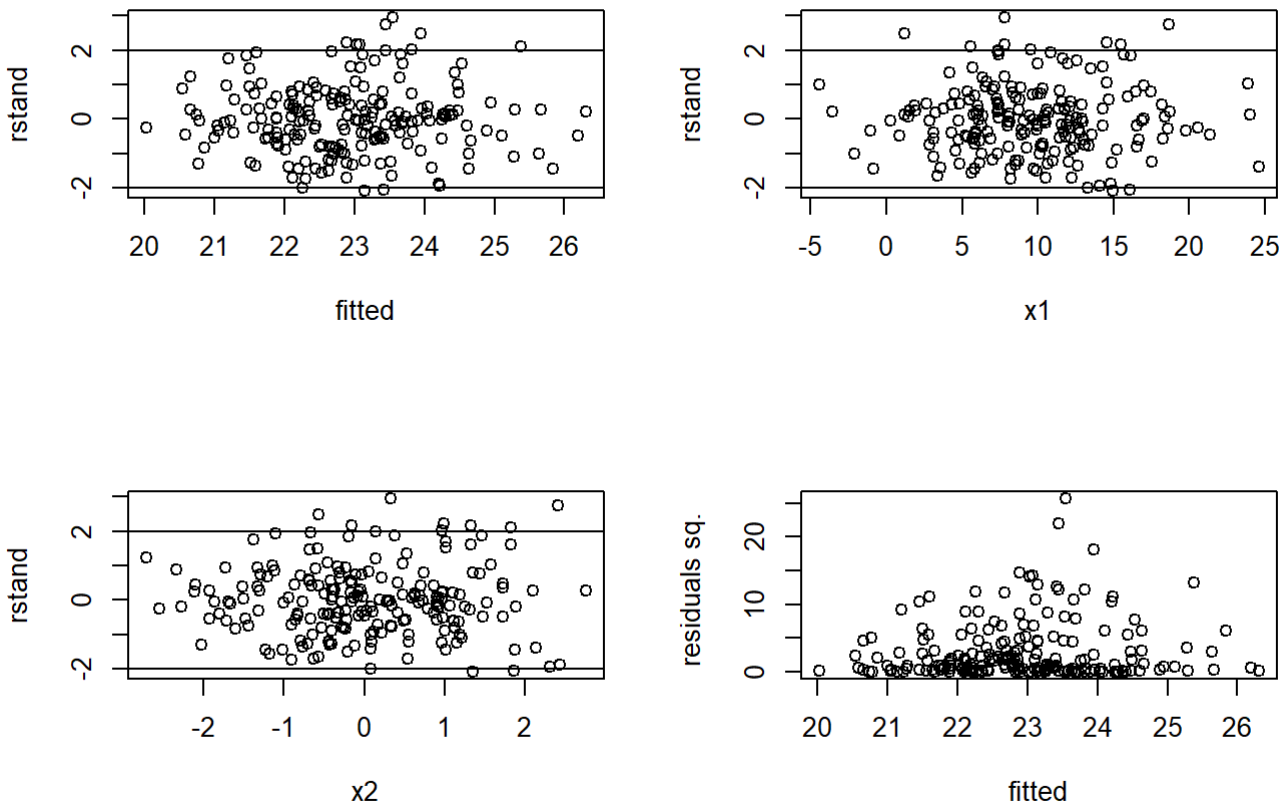
Come è possibile osservare è stato risolto il problema della multicollinearità.

Omoschedasticità

Si effettua ora lo studio dell'omoschedasticità

Si effettua l'**analisi grafica**:

```
par(mfrow=c(2,2))

plot(fitted(mod2),rstudent(mod2),ylab="rstand", xlab='fitted')
abline(h=2)
abline(h=-2)

plot(data$x1,rstudent(mod2), ylab='rstand', xlab='x1')
abline(h=2)
abline(h=-2)

plot(data$x2,rstudent(mod2), ylab='rstand', xlab='x2')
abline(h=2)
abline(h=-2)

plot(mod2$fitted, (mod2$residuals)^2, xlab='fitted',ylab='residuals sq. ')
```



Analizzando i grafici dei residui standardizzati vs i fitted, e dei residui vs le variabili esplicative, così come considerando i residui al quadrato non si rilevano particolari pattern che possano indicare la presenza di eteroschedasticità nel modello.

Tuttavia è fondamentale effettuare anche il **white test** per verificare analiticamente tale osservazione:

```
pander(white.test(mod1),big.mark=',')
```

Test.Statistic	P
3.763	0.1524

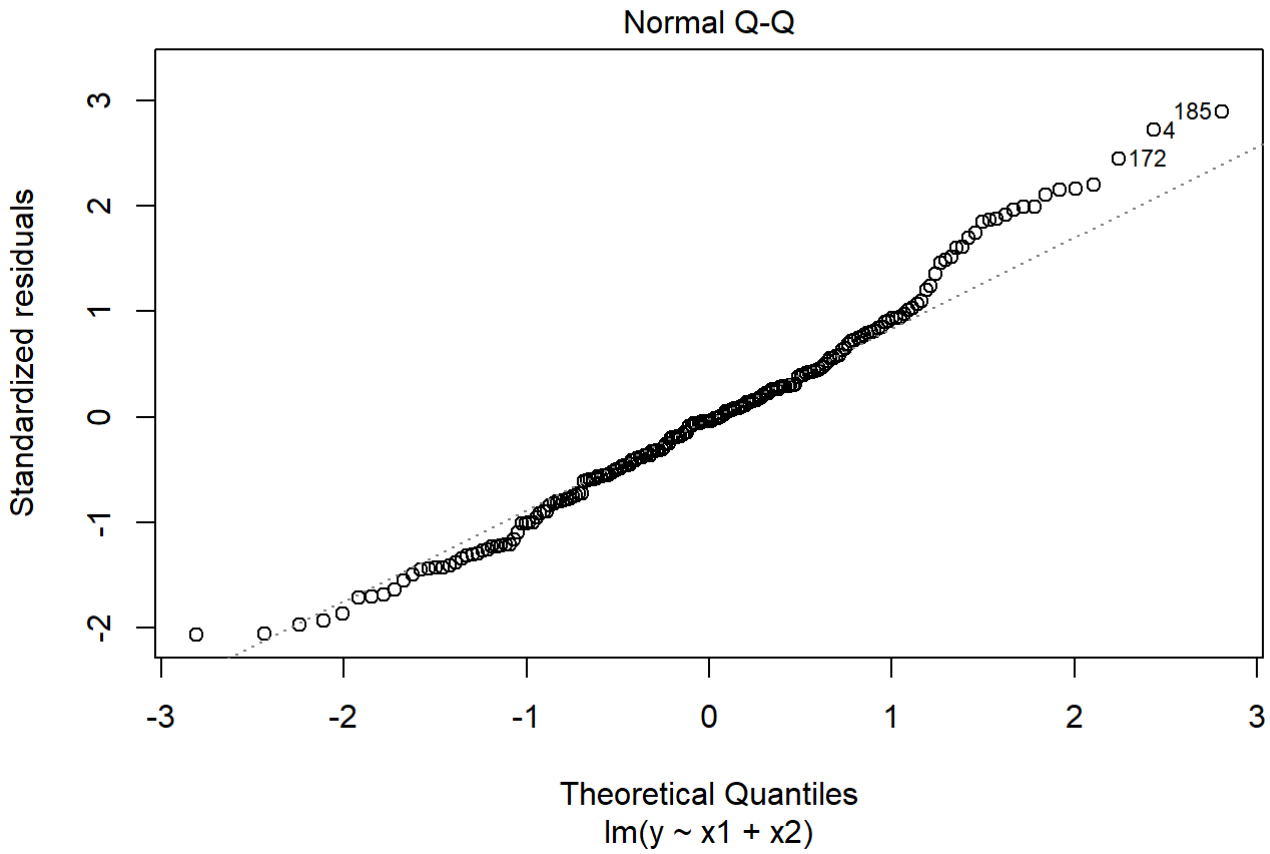
Il test di White per l'omoschedasticità risulta nella regione di accettazione: non si rifiuta l'ipotesi di omoschedasticità dei residui.

Non si ritiene dunque necessario apportare alcun tipo di modifica, come ci si aspettava dall'analisi grafica.

Normalità dei residui

Si consideri il **QQ plot**

```
plot(mod2,which=2)
```



Il QQPlot mostra un andamento irregolare solo sulle code: all'inizio e alla fine i punti si discostano dalla distribuzione teorica, tale andamento potrebbe essere dovuto alla presenza di outliers.

Si effettuano i **test per la normalità**

```
ols_test_normality(mod2)
```

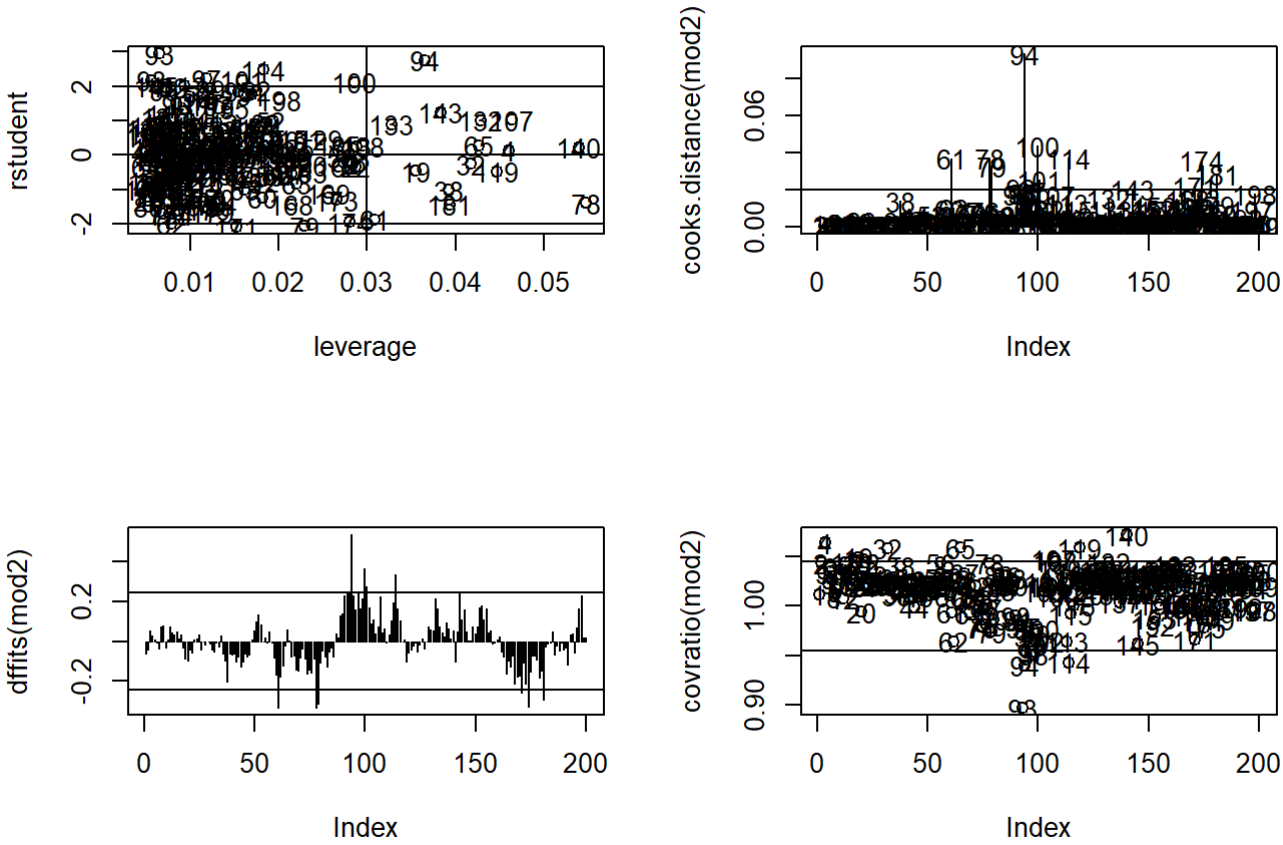
```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9832         0.0175
## Kolmogorov-Smirnov        0.0622         0.4221
## Cramer-von Mises         14.1514       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

Il test di Shapiro-Wilks risulta nella regione di rifiuto dell'ipotesi nulla di normalità dei residui, tuttavia si decide di proseguire nell'analisi delle ipotesi, per verificare la presenza di eventuali valori anomali, considerando anche che n>25 dunque è possibile ricondursi al teorema del limite centrale per cui gli errori standard sono calcolati asintoticamente, dunque si decide di non realizzare correzioni in merito.

Outlier

Si considerino eventuali outliers:

```
n=length(fitted(mod2))
k=length(coef(mod2))

par(mfrow=c(2,2))

plot(hatvalues(mod2),rstudent(mod2),ylab='rstudent',xlab='leverage')
abline(h=2)
abline(h=-2)
abline(h=0)
abline(v=2*k/n)
text(hatvalues(mod2),rstudent(mod2))

plot(cooks.distance(mod2),type='h')
abline(h=4/n)
text(cooks.distance(mod2))

plot(dffits(mod2),type='h')
abline(h=2*sqrt(k/n))
abline(h=-2*sqrt(k/n))

plot(covratio(mod2))
abline(h=1+3*(k/n))
abline(h=1-3*(k/n))
text(covratio(mod2))
```



Si rileva la presenza di diversi outliers, in particolare si notano i valori con indice 94, 79, 100.

Si decide dunque di realizzare un dataset *pulito* eliminando i valori anomali:

```
data2<-data[hatvalues(mod2)<2*k/n & abs(rstudent(mod2))<2 & cooks.distance(mod2)<4/n,]
mod_out<-lm(y~x1+x2,data2)
pander(summary(mod_out))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.49	0.2798	87.52	2.565e-143
x1	-0.1812	0.02695	-6.725	2.554e-10
x2	0.9707	0.1238	7.843	4.651e-13

Fitting linear model: y ~ x1 + x2 Si osserva che tutte le variabili risultano significative.

Observations	Residual Std. Error	R^2	Adjusted R^2
173	1.472	0.3242	0.3163

```
ols_test_normality(mod_out)
```

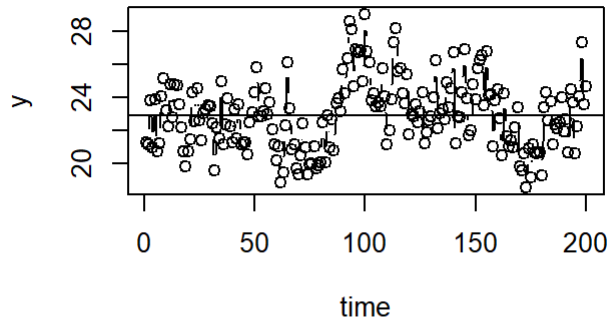
```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9892         0.2133
## Kolmogorov-Smirnov        0.0356         0.9810
## Cramer-von Mises          10.987         0.0000
## Anderson-Darling          0.3198         0.5303
## -----------------------------------------------
```

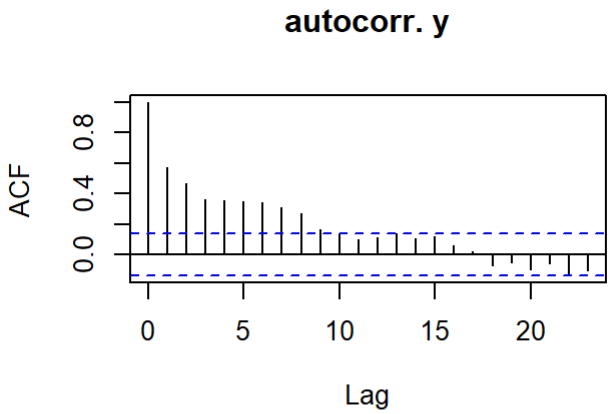
E’ stato risolto anche il problema della normalità.

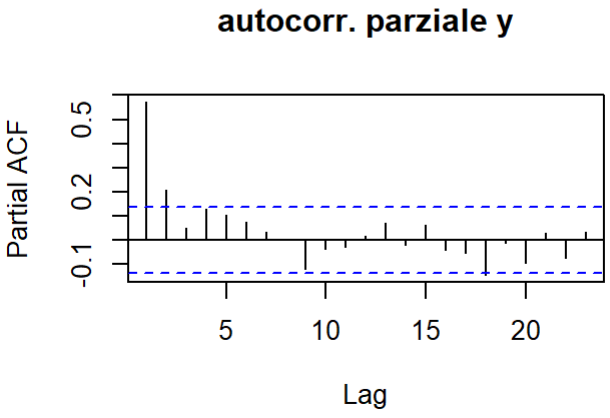
Autocorrelazione

Si effettua l’**analisi grafica**:

```
par(mfrow=c(2,2))

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

acf(data$y, main="autocorr. y")

pacf(data$y, main="autocorr. parziale y")
```







Analzzando i grafici si rileva la presenza di autocorrelazione, in primo luogo viene considerata autocorrelazione di primo ordine.

Si effettua così il *Test di Durbin-Watson* per verificare analiticamente l’autocorrelazione:

```
durbinWatsonTest(mod2,max.lag=8)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.8424635     0.3117179       0
##    2       0.7172077     0.5616804       0
##    3       0.6259871     0.7305112       0
##    4       0.5559689     0.8598509       0
##    5       0.5092048     0.9518115       0
##    6       0.4775257     1.0143303       0
##    7       0.4359449     1.0957253       0
##    8       0.3788820     1.2052920       0
##  Alternative hypothesis: rho[lag] != 0
```

Si rileva la presenza di autocorrelazione, infatti viene respinta l'ipotesi nulla di incorrelazione, e si verifica che la stastitica presenta valori < 1 che indicano la presenza di autocorrelazione positiva.

Si procede prima di tutto considerando autocorrelazione di primo ordine.

Risoluzione

Si decide di utilizzare la **procedura di Cochrane Orcutt** in cui si effettua una stima di ρ, e successivamente si effettuano le stime GLS.

Vengono creati i residui ritardati:

```
data$u_hat<-mod2$residuals
data<-slide(data=data, Var='u_hat',TimeVar= 'time', NewVar='u_hat_lag')
```

Si procede effettuando la regressione di u_hat su u_hat_lag

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.008121	0.06661	0.1219	0.9031
u_hat_lag	0.8426	0.03816	22.08	3.587e-55

Fitting linear model: u_hat ~ u_hat_lag La stima di u_hat_lag è 0.8426

Observations	Residual Std. Error	R^2	Adjusted R^2
199	0.9397	0.7122	0.7107

Si memorizza il coefficiente di autocorrelazione rho:

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

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

Vengono costruite le variabili laggate:

```
data<-slide(data=data, Var='y',TimeVar= 'time', NewVar='y_lag')
data<-slide(data=data, Var='x1',TimeVar= 'time', NewVar='x1_lag')
data<-slide(data=data, Var='x2',TimeVar= 'time', NewVar='x2_lag')
```

Si creano le variabili trasformate:

```
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
```

Si stima il modello con le variabili trasformate:

```
mod3<-lm(y_t~0+interc_t+x1_t+x2_t,data)
pander(summary(mod3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
interc_t	24.8	0.4342	57.11	4.086e-124
x1_t	-0.194	0.01022	-18.97	2.987e-46
x2_t	0.9784	0.05186	18.87	6.061e-46

Fitting linear model: y_t ~ 0 + interc_t + x1_t + x2_t

Observations	Residual Std. Error	R^2	Adjusted R^2
199	0.941	0.9468	0.946

```
pander(summary(mod2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.71	0.2735	90.37	2.574e-162
x1	-0.1901	0.02562	-7.422	3.376e-12
x2	0.9454	0.1213	7.794	3.653e-13

Fitting linear model: y ~ x1 + x2 Si osserva che la bontà di adattamento del modello risulta ampiamente migliorata, il modello adesso presenta un'ottima bontà di adattamento (R^2 = 0.9468)

Observations	Residual Std. Error	R^2	Adjusted R^2
200	1.754	0.3095	0.3025

Il modello come prima è significativo.

Le variabili rimangono stasticamente significative.

Si effettua il test di Durbin Watson per controllo:

```
durbinWatsonTest(mod3, max.lag = 5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.015731459      2.028123   0.838
##    2    -0.032258886      2.033594   0.706
##    3    -0.007593326      1.981864   0.970
##    4    -0.020029177      1.989661   0.862
##    5     0.007448101      1.920812   0.760
##  Alternative hypothesis: rho[lag] != 0
```

Si osserva che l’autocorrelazione è stata risolta.