GLS 1 - Data set: COST FUNCTION

INTRODUZIONE

I dati utilizzati per questa analisi contengono le caratteristiche di una serie storica di carattere economico nel range di anni 1947-1971. Le variabili contenute sono:

1. YEAR: anno
2. COST: indice di costo
3. K: quota del costo capitale
4. L: quota del costo del lavoro
5. E: quota del costo dell'energia
6. M: quota del costo dei materiali
7. PK: costo del capitale
8. PL: costo del lavoro
9. PE: costo dell'energia
10. PM: costo dei materiali

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
library(AER)
data("ManufactCosts")
d <- data.frame(ManufactCosts)
```

```r
names(d) <- c("cost","k","l","e","m","pk","pl","pe","pm")

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("cost","k","l","e","m","pk","pl","pe","pm")

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

cost	k	l	e	m	pk	pl	pe	pm
182.4	0.05107	0.2473	0.04253	0.6591	1	1	1	1
183.2	0.05817	0.2772	0.05127	0.6134	1.003	1.155	1.303	1.055
186.5	0.04602	0.2591	0.05075	0.6441	0.7437	1.156	1.197	1.066
221.7	0.04991	0.2479	0.04606	0.6561	0.925	1.235	1.124	1.124
255.9	0.05039	0.2549	0.04482	0.6499	1.049	1.338	1.252	1.217
264.7	0.04916	0.2666	0.0446	0.6397	0.9974	1.379	1.279	1.2

STATISTICHE DESCRITTIVE

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

Table 2: Table continues below

cost	k	l	e
Min. :182.4	Min. :0.04602	Min. :0.2473	Min. :0.03963
1st Qu.:274.5	1st Qu.:0.05033	1st Qu.:0.2683	1st Qu.:0.04348
Median :358.4	Median :0.05443	Median :0.2772	Median :0.04482
Mean :380.9	Mean :0.05349	Mean :0.2745	Mean :0.04482
3rd Qu.:475.0	3rd Qu.:0.05635	3rd Qu.:0.2834	3rd Qu.:0.04606
Max. :658.2	Max. :0.06185	Max. :0.2975	Max. :0.05127

m	pk	pl	pe	pm
Min. :0.6064	Min. :0.7437	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:0.6181	1st Qu.:1.0065	1st Qu.:1.435	1st Qu.:1.303	1st Qu.:1.206
Median :0.6196	Median :1.2018	Median :1.734	Median :1.376	Median :1.327
Mean :0.6272	Mean :1.1836	Mean :1.772	Mean :1.346	Mean :1.301
3rd Qu.:0.6394	3rd Qu.:1.3246	3rd Qu.:2.055	3rd Qu.:1.392	3rd Qu.:1.375
Max. :0.6591	Max. :1.4990	Max. :2.760	Max. :1.647	Max. :1.550

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

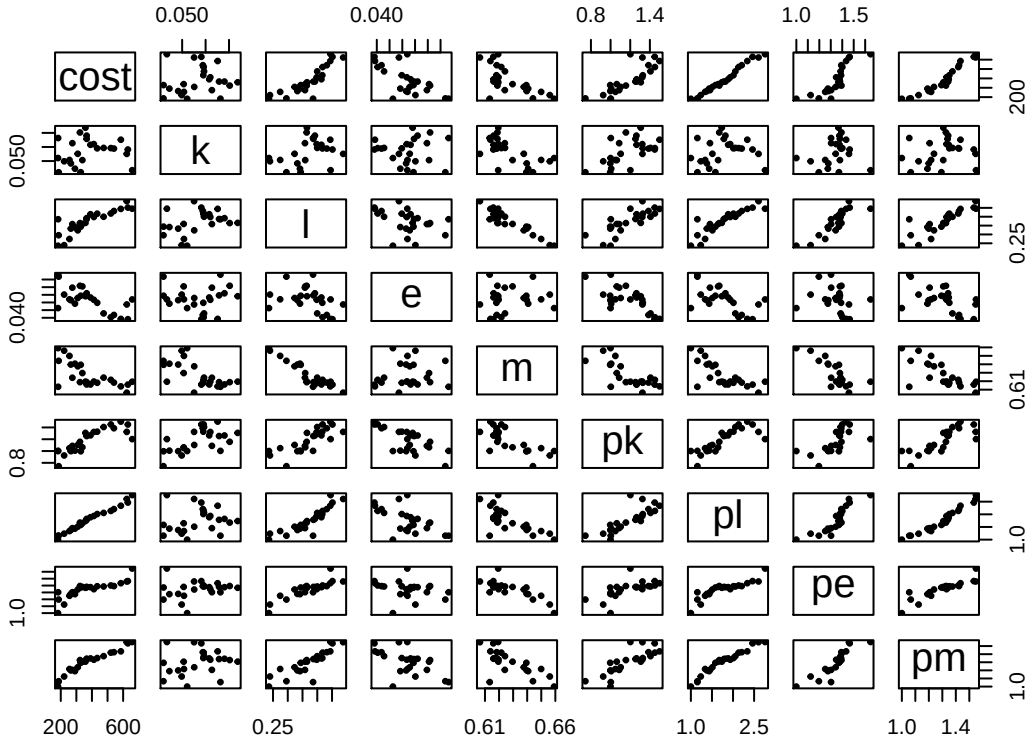
Table 4: Table continues below

	cost	k	l	e	m	pk
cost	1	0.1324	0.8592	-0.6702	-0.6668	0.8236

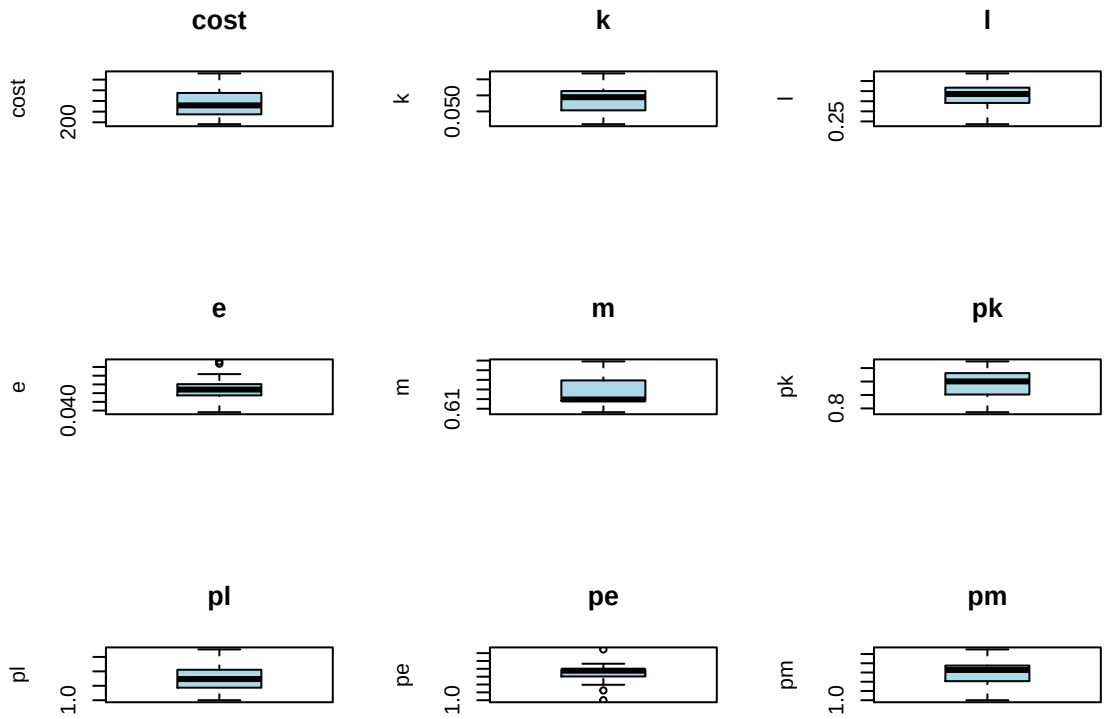
	cost	k	l	e	m	pk
k	0.1324	1	0.3627	-0.02038	-0.6316	0.5728
l	0.8592	0.3627	1	-0.3786	-0.9277	0.7874
e	-0.6702	-0.02038	-0.3786	1	0.1301	-0.7141
m	-0.6668	-0.6316	-0.9277	0.1301	1	-0.7301
pk	0.8236	0.5728	0.7874	-0.7141	-0.7301	1
pl	0.9897	0.1678	0.8863	-0.5809	-0.7213	0.8096
pe	0.8184	0.1937	0.8618	-0.2341	-0.7821	0.6299
pm	0.9558	0.1997	0.8576	-0.5511	-0.7118	0.8053

	pl	pe	pm
cost	0.9897	0.8184	0.9558
k	0.1678	0.1937	0.1997
l	0.8863	0.8618	0.8576
e	-0.5809	-0.2341	-0.5511
m	-0.7213	-0.7821	-0.7118
pk	0.8096	0.6299	0.8053
pl	1	0.8653	0.9717
pe	0.8653	1	0.8877
pm	0.9717	0.8877	1

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



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



REGRESSIONE

Si effettua ora la regressione della variabili “Cost” su “L”, “PK”, “PL”, “PM”.

```
#-- R CODE
mod1 <- lm(cost ~ l + pk + pl + pm, d) #-- stima modello lineare semplice
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	213.6	184.7	1.157	0.2611
l	-1,321	695.5	-1.9	0.07198
pk	71.97	36.73	1.959	0.06418
pl	356	40.08	8.883	2.228e-08
pm	-142.9	116.4	-1.228	0.2337

Table 7: Fitting linear model: cost ~ l + pk + pl + pm

Observations	Residual Std. Error	R^2	Adjusted R^2
25	19.59	0.9847	0.9817

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

Table 8: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
l	1	370,650	370,650	965.8	2.102e-18
pk	1	28,568	28,568	74.44	3.559e-08
pl	1	94,560	94,560	246.4	1.039e-12
pm	1	578.7	578.7	1.508	0.2337
Residuals	20	7,676	383.8	NA	NA

Il modello interpreta bene la variabile dipendente. Tuttavia solo il parametro associato alla variabile "pl" risulta chiaramente significativo.

Verifichiamo ora la sfericità dei residui; il test di White mostra con chiarezza che i residui sono omoschedastici.

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

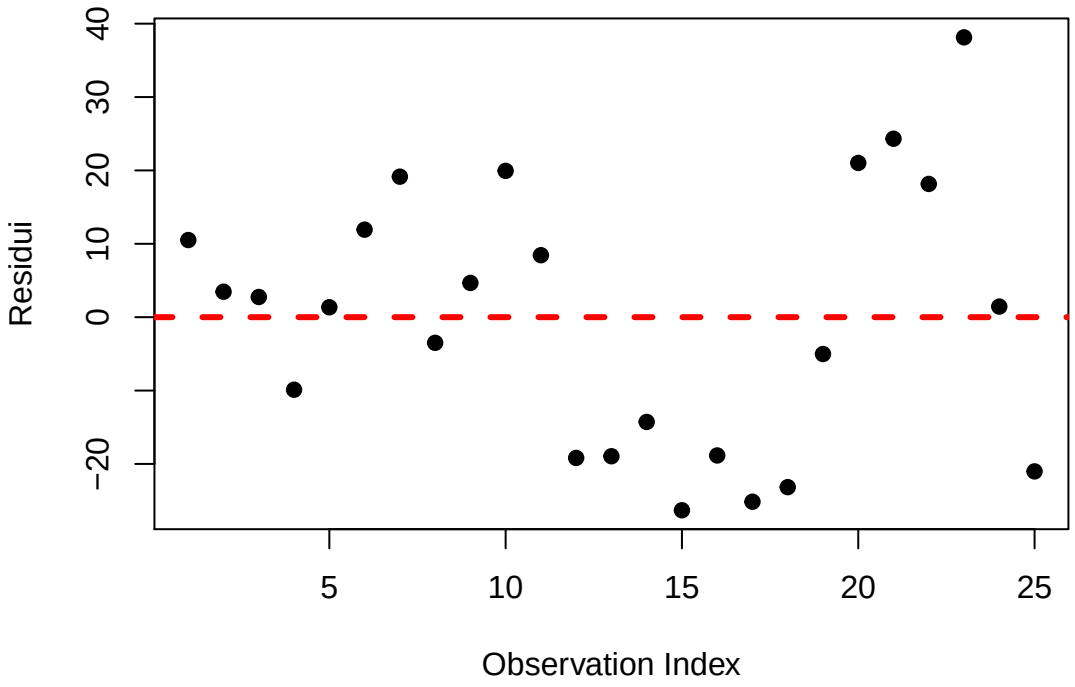
Test.statistic	P.value
7.361	0.0252

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

Table 10: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
0.7531	1.033e-05 * * *	true autocorrelation is greater than 0

```
#-- R CODE
plot(1:nrow(d),resid(mod1),xlab="Observation Index",ylab="Residui",pch=19)
abline(h=0,col=2,lwd=3,lty=2)
```



Il grafico dei residui mostra un andamento "non rettangolare" a segnalare l'esistenza di correlazione. Si calcola perciò il coefficiente di autocorrelazione di primo grado fra i residui regredendo i residui rispetto ai residui ritardati.

```
#-- R CODE

library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.4.3
```

```
d1  <- d
d1$resid <- resid(mod1)
d1$resid_l1 <- Lag(d1$resid,1)

pander(cor(data.frame(d1$resid,d1$resid_l1),use="pairwise.complete.obs"))
```

	d1.resid	d1.resid_l1
d1.resid	1	0.6117
d1.resid_l1	0.6117	1

```
#-- R CODE
mod2 <- arima(d1$cost, order=c(1,0,0), xreg = d1[,c("l","pk","pl","pm")],method="ML")
mod2
```

```
##
## Call:
## arima(x = d1$cost, order = c(1, 0, 0), xreg = d1[, c("l", "pk", "pl", "pm")],
##     method = "ML")
##
## Coefficients:
##          ar1  intercept           l       pk        pl        pm
##       0.7160    77.8908  -1216.9719  68.8727  290.4698   32.2447
## s.e.  0.1468   114.0014    369.0325  29.1789   37.5612  105.9266
##
## sigma^2 estimated as 169.2:  log likelihood = -99.97,  aic = 213.95
```

```
coeftest(mod2)
```

```
##
## z test of coefficients:
##
##              Estimate  Std. Error z value  Pr(>|z|)
## ar1           0.71601     0.14680  4.8774 1.075e-06 ***
## intercept    77.89082   114.00144  0.6832 0.4944526
## l         -1216.97193   369.03245 -3.2977 0.0009747 ***
## pk           68.87275    29.17888  2.3604 0.0182571 *
## pl          290.46983    37.56124  7.7332 1.048e-14 ***
## pm           32.24472   105.92663  0.3044 0.7608185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
durbinWatsonTest(as.numeric(mod2$residuals))
```

```
## [1] 1.620563
```

Come era prevedibile i modelli danno risultati simili: i valori dei D e i p-value per il Durbin Watson mostra che è accettata l'ipotesi di non autocorrelazione dei residui.

Dal punto di vista interpretativo si evince che il fattore determinante il costo della manifattura negli anni considerati è stato il costo del lavoro.

Si osserva dal p-value associato al parametro AR1 che corregge i residui correlati che tale parametro risulta significativo, vale a dire che prima della correzione i residui erano correlati come si era visto nelle precedenti analisi mostrate sopra.