MULTI 4 - Data set: CIGARETTE

INTRODUZIONE

In questo dataset sono contenute 48 osservazioni e le seguenti variabili:

1. STATE: stato
2. YEAR: anno
3. CPI: consumer price index
4. POP: popolazione
5. PACKPC: numero di pacchetti consumati pro-capite
6. INCOME: state personal income
7. TAX: tassazione
8. AVGPRS: prezzo medio incluse le tasse
9. TAXS: tassazione per esercizio

Analisi proposte:

1. Statistiche descrittive
2. Regressione Multivariata

```r
#-- R CODE
library(car)
library(sjstats)
library(plotrix)
library(sjPlot)
library(sjmisc)
library(lme4)
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"
```

```r
d <- read.csv(paste0(ABSOLUTE_PATH,"\\esercizi (5)  copia\\4.mult\\Cigarette.txt"),sep=" ")

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("cpi","pop","packpc","income","tax")

d_ca <- d[d$state=="CA",]
names(d_ca) <- paste0(names(d_ca),"_CA")

d_ar <- d[d$state=="AR",]
names(d_ar) <- paste0(names(d_ar),"_AR")

d1 <- cbind(d_ar,d_ca)

d_tx <- d[d$state=="TX",]
names(d_tx) <- paste0(names(d_tx),"_TX")

d2 <- cbind(d_tx,d_ca)

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

ID	state	year	cpi	pop	packpc	income	tax	avgprs	taxs
1	AL	1,985	1.076	3,973,000	116.5	46,014,968	32.5	102.2	33.35
2	AR	1,985	1.076	2,327,000	128.5	26,210,736	37	101.5	37
3	AZ	1,985	1.076	3,184,000	104.5	43,956,936	31	108.6	36.17
4	CA	1,985	1.076	26,444,000	100.4	447,102,816	26	107.8	32.1
5	CO	1,985	1.076	3,209,000	113	49,466,672	31	94.27	31
6	CT	1,985	1.076	3,201,000	109.3	60,063,368	42	128	51.48

STATISTICHE DESCRITTIVE

Si propongono la matrice di correlazione tra le variabili e alcune descrittive di base.

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

Table 2: Table continues below

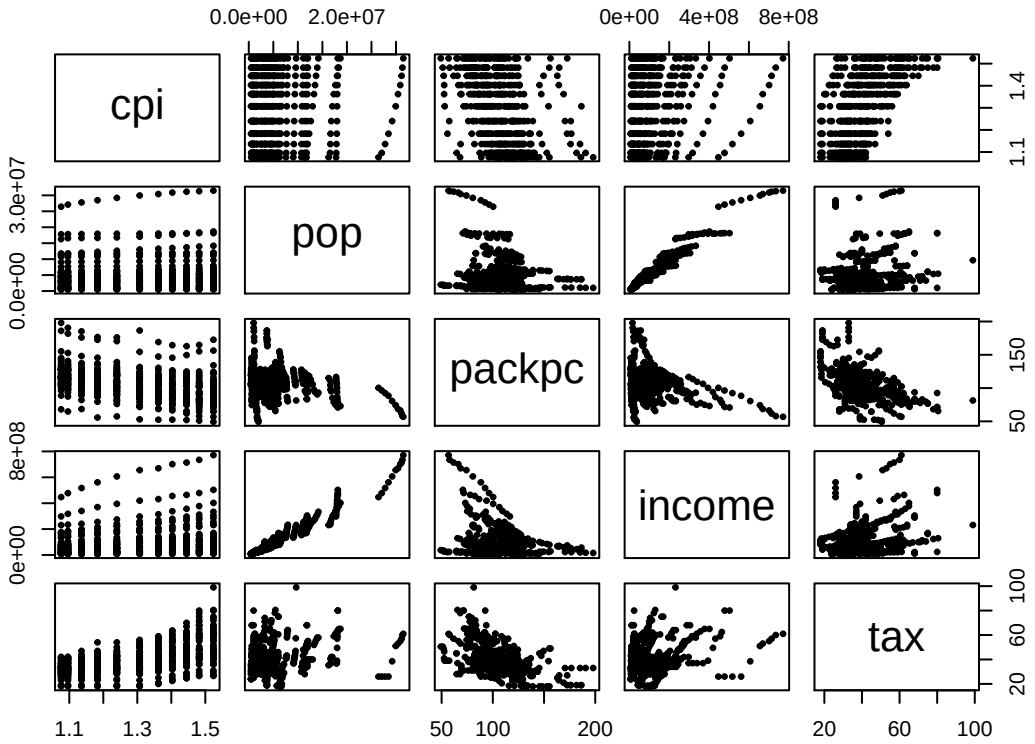
cpi	pop	packpc	income
Min. :1.076	Min. : 453401	Min. : 49.27	Min. : 6769883
1st Qu.:1.136	1st Qu.: 1579248	1st Qu.: 91.41	1st Qu.: 25100460
Median :1.307	Median : 3628254	Median :105.91	Median : 61692010
Mean :1.296	Mean : 5159738	Mean :106.45	Mean : 99245556
3rd Qu.:1.445	3rd Qu.: 6012163	3rd Qu.:119.59	3rd Qu.:118084118
Max. :1.524	Max. :31493524	Max. :197.99	Max. :771470144

tax
Min. :18.00
1st Qu.:32.00
Median :39.00
Mean :40.35
3rd Qu.:46.31
Max. :99.00

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

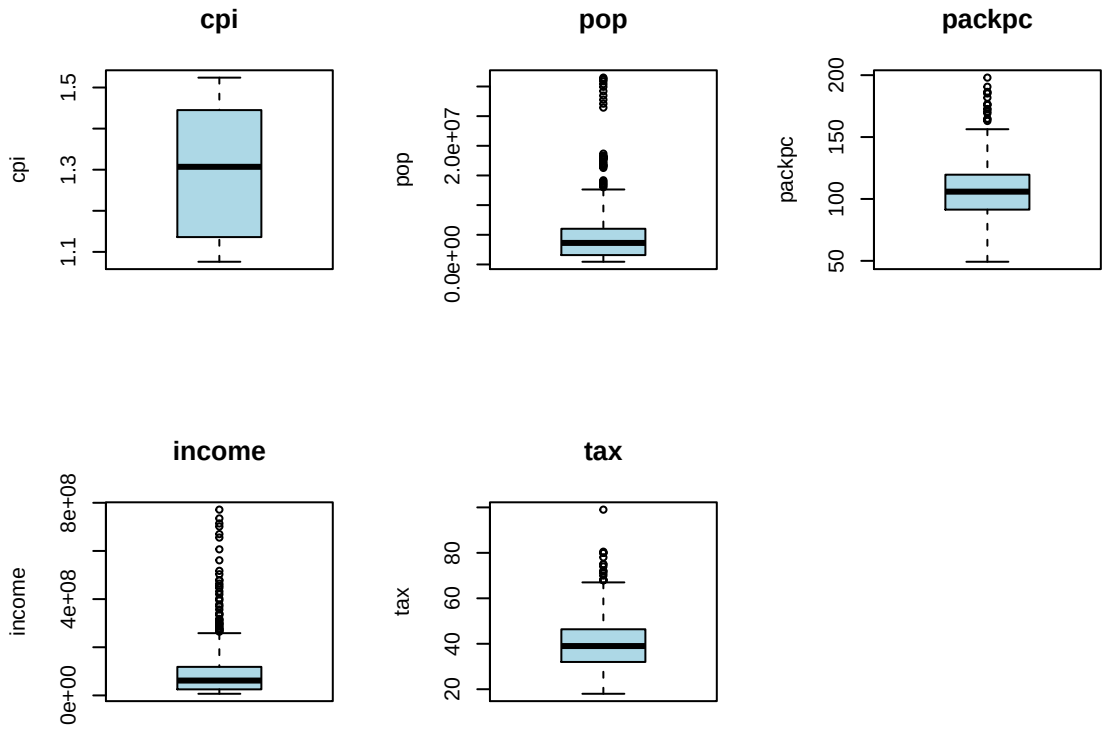
	cpi	pop	packpc	income	tax
cpi	1	0.03033	-0.4036	0.1511	0.5744
pop	0.03033	1	-0.1813	0.977	0.1541
packpc	-0.4036	-0.1813	1	-0.2554	-0.5677
income	0.1511	0.977	-0.2554	1	0.2681
tax	0.5744	0.1541	-0.5677	0.2681	1

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

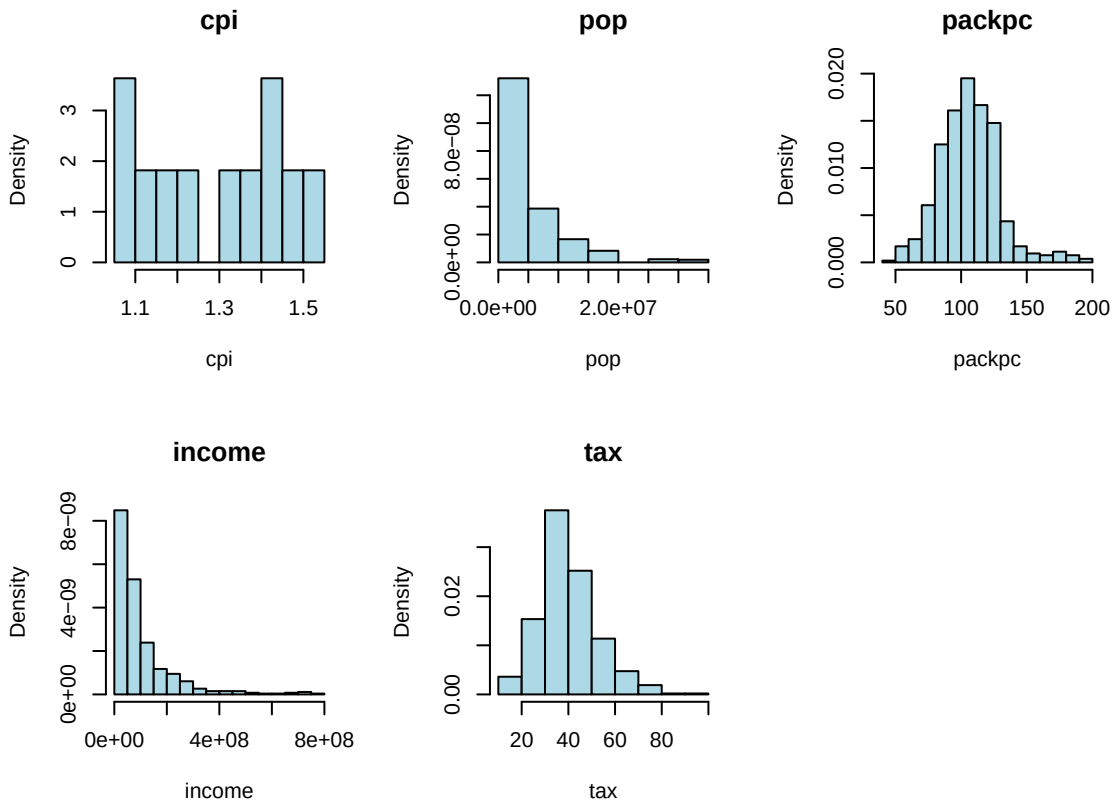



```
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```

```r
par(mfrow=c(2,3))
```




```r
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```



ESERCIZIO 1

L'obiettivo dell'analisi sarà quello di spiegare la variabile "packpc" tramite i regressori "cpi", "pop", "income" e "tax". L'analisi in questione si svolgerà su base regionale considerando due stati: Arkansas e California.

```
#-- R CODE
mod1_AR <- lm(packpc_AR ~ cpi_AR + pop_AR + income_AR + tax_AR, d1)
pander(summary(mod1_AR),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	419.5	84.39	4.971	0.002524
cpi_AR	-257.7	65.89	-3.912	0.007877
pop_AR	-6.304e-05	4.495e-05	-1.403	0.2103
income_AR	6.858e-06	1.869e-06	3.67	0.01046
tax_AR	-1.333	0.5979	-2.229	0.06735

Table 6: Fitting linear model: packpc_AR ~ cpi_AR + pop_AR + income_AR + tax_AR

Observations	Residual Std. Error	R^2	Adjusted R^2
11	2.311	0.9563	0.9272

Observations	Residual Std. Error	R^2	Adjusted R^2

```
pander(anova(mod1_AR),big.mark=",")
```

Table 7: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cpi_AR	1	539.7	539.7	101	5.624e-05
pop_AR	1	83.78	83.78	15.68	0.007448
income_AR	1	51.15	51.15	9.575	0.02127
tax_AR	1	26.54	26.54	4.969	0.06735
Residuals	6	32.05	5.342	NA	NA

Il modello spiega molto bene la variabile dipendente packpc ($R^2 = 0.95$), ma le uniche variabili significative sono "cpi" ad un livello $\alpha = 0.01$, "income" ($\alpha = 0.005$) e "tax" $\alpha = 0.01$. Vediamo ora cosa accade in California.

```
#-- R CODE
mod1_CA <- lm(packpc_CA ~ cpi_CA + pop_CA + income_CA + tax_CA, d1)
pander(summary(mod1_CA),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	278.8	33.36	8.358	0.0001595
cpi_CA	-61.38	12.9	-4.756	0.003138
pop_CA	-4.806e-06	1.593e-06	-3.016	0.02351
income_CA	3.84e-08	3.305e-08	1.162	0.2894
tax_CA	-0.112	0.07017	-1.596	0.1616

Table 9: Fitting linear model: packpc_CA ~ cpi_CA + pop_CA + income_CA + tax_CA

Observations	Residual Std. Error	R^2	Adjusted R^2
11	0.7991	0.9985	0.9974

```
pander(anova(mod1_CA),big.mark=",")
```

Table 10: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cpi_CA	1	2,478	2,478	3,881	1.15e-09
pop_CA	1	12.4	12.4	19.42	0.004533
income_CA	1	1.325	1.325	2.075	0.1998
tax_CA	1	1.626	1.626	2.547	0.1616
Residuals	6	3.831	0.6386	NA	NA

Il modello interpreta quasi perfettamente la variabile dipendente e in questo caso le variabili significative sono “cpi”, “pop”.

Si ricorda che se si sceglie il metodo OLS la regressione multivariata può essere costruita per ciò che riguarda l’ottenimento del fitting e dei parametri sulla base delle regressioni multiple costruite separatamente.

La correlazione tra i residui delle variabili dipendenti nelle due equazioni che risulta essere 0.5662. Si passa ora a un modello in cui le variabili esplicative rimangano identiche in entrambe le regressioni e vi è correlazione fra gli stessi individui nelle diverse equazioni

```
#-- R CODE
pander(cor(data.frame(resid(mod1_CA),resid(mod1_AR))))
```

```
##
## ---------------------------------------------------
##                   resid.mod1_CA.   resid.mod1_AR.
## -------------------- ---------------- ----------------
##  **resid.mod1_CA.**          1               0.5666
##
##  **resid.mod1_AR.**        0.5666             1
## ---------------------------------------------------
```

```
e1 <- packpc_AR ~ cpi_AR + pop_AR + income_AR + tax_AR
e2 <- packpc_CA ~ cpi_CA + pop_CA + income_CA + tax_CA
sistema <- list(e1=e1,e2=e2)

mod1 <- systemfit(sistema,"SUR",data=d1)
summary(mod1)
```

```
##
## systemfit results
## method: SUR
##
##          N DF     SSR detRCov   OLS-R2 McElroy-R2
## system 22 12 38.9895 1.24531 0.987932   0.998651
##
##     N DF      SSR      MSE     RMSE      R2   Adj R2
## e1 11  6 34.79245 5.798741 2.408057 0.95255 0.920916
## e2 11  6  4.19707 0.699512 0.836368 0.99832 0.997199
##
## The covariance matrix of the residuals used for estimation
##         e1       e2
## e1 5.34166 1.046491
## e2 1.04649 0.638571
##
## The covariance matrix of the residuals
##         e1       e2
## e1 5.79874 1.676596
## e2 1.67660 0.699512
##
## The correlations of the residuals
##          e1       e2
## e1 1.000000 0.832461
## e2 0.832461 1.000000
##
```

```
##
## SUR estimates for 'e1' (equation 1)
## Model Formula: packpc_AR ~ cpi_AR + pop_AR + income_AR + tax_AR
##
##                 Estimate   Std. Error  t value  Pr(>|t|)
## (Intercept)  3.63192e+02  7.08173e+01  5.12857 0.0021600 **
## cpi_AR      -2.49217e+02  5.84900e+01 -4.26084 0.0053163 **
## pop_AR      -3.74634e-05  3.75376e-05 -0.99803 0.3567999
## income_AR    6.52862e-06  1.65535e-06  3.94394 0.0075902 **
## tax_AR      -1.43426e+00  5.24627e-01 -2.73386 0.0340115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.408057 on 6 degrees of freedom
## Number of observations: 11 Degrees of Freedom: 6
## SSR: 34.792445 MSE: 5.798741 Root MSE: 2.408057
## Multiple R-Squared: 0.95255 Adjusted R-Squared: 0.920916
##
##
## SUR estimates for 'e2' (equation 2)
## Model Formula: packpc_CA ~ cpi_CA + pop_CA + income_CA + tax_CA
##
##                 Estimate   Std. Error  t value   Pr(>|t|)
## (Intercept)  2.88314e+02  2.80560e+01 10.27636 4.9571e-05 ***
## cpi_CA      -6.77451e+01  1.20047e+01 -5.64320  0.0013272 **
## pop_CA      -5.20450e-06  1.36478e-06 -3.81344  0.0088280 **
## income_CA    5.63196e-08  2.82632e-08  1.99268  0.0933686 .
## tax_CA      -1.27173e-01  6.12950e-02 -2.07477  0.0833353 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.836368 on 6 degrees of freedom
## Number of observations: 11 Degrees of Freedom: 6
## SSR: 4.197071 MSE: 0.699512 Root MSE: 0.836368
## Multiple R-Squared: 0.99832 Adjusted R-Squared: 0.997199
```

Si vede la differenza dei risultati rispetto al caso OLS senza correlazione: "cpl_AR" e "income_AR" sono chiaramente significative molto più di quanto avveniva in precedenza (per $\alpha = 0.01$). Ciò è dovuto al fatto che mentre nel caso precedente esisteva una completa incorrelazione sia fra gli errori appartenenti alla stessa equazione che con gli errori dell' altra equazione in questo caso esiste correlazione fra i medesimi individui considerati nelle due diverse equazioni. Si verifica ora se il parametro relativo a "cpl_AR" si può ritenere in questa equazione uguale a quello relativo a "tax_CA".

```r
#-- R CODE
R1 <- matrix(0,nrow=1,ncol=10)
R1[ 1, 2 ] <- 1
R1[ 1, 10 ] <- -1

pander(linearHypothesis(mod1,R1,test="FT"),big.mark=",")
```

Res.Df	Df	F	Pr(>F)

Table 11: Linear hypothesis test (Theil's F test)

Res.Df	Df	F	Pr(>F)
13	NA	NA	NA
12	1	21.38	0.0005858

Si verifica ora se i parametri relativi alle medesime variabili "clp_AR" e "cpl_CA" e "tax_AR" e "tax_CA" hanno la stessa influenza nelle due equazioni.

```
#-- R CODE
R1 <- matrix(0,nrow=1,ncol=10)
R1[ 1, 2 ] <- 1
R1[ 1, 7 ] <- -1
pander(linearHypothesis(mod1,R1,test="FT")) #-- TEST: clp_AR=cpl_CA
```

Table 12: Linear hypothesis test (Theil's F test)

Res.Df	Df	F	Pr(>F)
13	NA	NA	NA
12	1	10.97	0.006204

```
R2 <- matrix(0,nrow=1,ncol=10)
R2[ 1, 5 ] <- 1
R2[ 1, 10 ] <- -1
pander(linearHypothesis(mod1,R2,test="FT")) #-- TEST: tax_AR=tax_CA
```

Table 13: Linear hypothesis test (Theil's F test)

Res.Df	Df	F	Pr(>F)
13	NA	NA	NA
12	1	7.359	0.01886

La prima è respinta per $\alpha = 0.01$, la seconda solo per $\alpha = 0.05$.

Passiamo ora a modello SURE con variabili esplicative differenti. In un primo caso le variabili esplicative per la prima equazione relativa all'Arkansas sono "cpl", "income", "tax" mentre per la seconda relativa alla California "cpi" e "pop" Si propone dapprima la stima OLS.

```
#-- R CODE
mod1_AR <- lm(packpc_AR ~ cpi_AR + income_AR + tax_AR, d1)
pander(summary(mod1_AR),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	309.7	33.65	9.205	3.684e-05

	Estimate	Std. Error	t value	Pr(>|t|)
cpi_AR	-291.8	65.36	-4.464	0.002921
income_AR	7.74e-06	1.878e-06	4.122	0.004449
tax_AR	-1.985	0.4007	-4.955	0.001646

Table 15: Fitting linear model: packpc_AR ~ cpi_AR + income_AR + tax_AR

Observations	Residual Std. Error	R^2	Adjusted R^2
11	2.466	0.942	0.9171

```
pander(anova(mod1_AR),big.mark=",")
```

Table 16: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cpi_AR	1	539.7	539.7	88.78	3.163e-05
income_AR	1	1.679	1.679	0.2762	0.6154
tax_AR	1	149.3	149.3	24.56	0.001646
Residuals	7	42.56	6.08	NA	NA

```
#-- R CODE
mod1_CA <- lm(packpc_CA ~ cpi_CA + pop_CA, d1)
pander(summary(mod1_CA),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	255.7	13.57	18.83	6.527e-08
cpi_CA	-61.51	9.883	-6.224	0.0002529
pop_CA	-3.368e-06	8.804e-07	-3.825	0.005055

Table 18: Fitting linear model: packpc_CA ~ cpi_CA + pop_CA

Observations	Residual Std. Error	R^2	Adjusted R^2
11	0.9208	0.9973	0.9966

```
pander(anova(mod1_CA),big.mark=",")
```

Table 19: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cpi_CA	1	2,478	2,478	2,923	1.519e-11
pop_CA	1	12.4	12.4	14.63	0.005055
Residuals	8	6.783	0.8478	NA	NA

I risultati cambiano ancora. In entrambe le equazioni il fitting rimane molto alto ma nella prima tutte e 3 le variabili esplicative "cpi", "income", "tax" risultano significative per un p-value inferiore che in precedenza $\alpha = 0.005$ mentre nella seconda entrambe le variabili esplicative "cpl" e "pop" risultano significative per p-value rispettivamente $\alpha = 0.0005$ e $\alpha = 0.01$.

Si nota che la correlazione tra i valori stimati delle variabili dipendenti stimate diminuisce sensibilmente rispetto al primo caso OLS.

```
#-- R CODE
pander(cor(data.frame(resid(mod1_CA),resid(mod1_AR))))
```

	resid.mod1_CA.	resid.mod1_AR.
resid.mod1_CA.	1	0.2067
resid.mod1_AR.	0.2067	1

Si propone il modello con variabili esplicative diverse stimate con il metodo SURE.

```
#-- R CODE
e1 <- packpc_AR ~ cpi_AR + income_AR + tax_AR
e2 <- packpc_CA ~ cpi_CA + pop_CA
sistema <- list(e1=e1,e2=e2)

mod1 <- systemfit(sistema,"SUR",data=d1)
summary(mod1)
```

```
##
## systemfit results
## method: SUR
##
##         N DF     SSR detRCov   OLS-R2 McElroy-R2
## system 22 15 50.5545 4.65182 0.984353   0.995121
##
##     N DF      SSR      MSE     RMSE       R2   Adj R2
## e1 11  7 43.73176 6.247395 2.499479 0.940358 0.914798
## e2 11  8  6.82269 0.852837 0.923492 0.997268 0.996585
##
## The covariance matrix of the residuals used for estimation
##          e1       e2
## e1 6.079631 0.469357
## e2 0.469357 0.847833
##
## The covariance matrix of the residuals
##          e1       e2
## e1 6.247395 0.822304
## e2 0.822304 0.852837
##
## The correlations of the residuals
##          e1       e2
## e1 1.000000 0.356246
## e2 0.356246 1.000000
```

```
##
##
## SUR estimates for 'e1' (equation 1)
## Model Formula: packpc_AR ~ cpi_AR + income_AR + tax_AR
##
##                 Estimate   Std. Error  t value   Pr(>|t|)
## (Intercept)  2.95198e+02  3.29457e+01  8.96016 4.3905e-05 ***
## cpi_AR      -2.63278e+02  6.39618e+01 -4.11618  0.0044810 **
## income_AR    6.91440e-06  1.83716e-06  3.76363  0.0070413 **
## tax_AR      -1.83126e+00  3.93425e-01 -4.65465  0.0023291 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.499479 on 7 degrees of freedom
## Number of observations: 11 Degrees of Freedom: 7
## SSR: 43.731764 MSE: 6.247395 Root MSE: 2.499479
## Multiple R-Squared: 0.940358 Adjusted R-Squared: 0.914798
##
##
## SUR estimates for 'e2' (equation 2)
## Model Formula: packpc_CA ~ cpi_CA + pop_CA
##
##                 Estimate   Std. Error  t value   Pr(>|t|)
## (Intercept)  2.52756e+02  1.34815e+01 18.74839 6.7657e-08 ***
## cpi_CA      -6.36196e+01  9.81623e+00 -6.48106  0.0001919 ***
## pop_CA      -3.17627e-06  8.74255e-07 -3.63311  0.0066550 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.923492 on 8 degrees of freedom
## Number of observations: 11 Degrees of Freedom: 8
## SSR: 6.822694 MSE: 0.852837 Root MSE: 0.923492
## Multiple R-Squared: 0.997268 Adjusted R-Squared: 0.996585
```

La variabili esplicative rimangono tutte significative ma con parametri e significatività diversa: $\alpha = 0.005$ per cpi_AR e tax_AR, $\alpha = 0.01$ per income_AR e pop_CA e $\alpha = 0.0005$ per cpl_CA. Si testano ora le ipotesi che non siano mutati i valori dei parametri per cpl_AR e tax_AR nella prima equazione SURE cambiando le variabili esplicative.

```r
#-- R CODE
pander(linearHypothesis(mod1,"e1_cpi_AR = -249.217",test="FT"),big.mark=",")
```

Table 21: Linear hypothesis test (Theil's F test)

Res.Df	Df	F	Pr(>F)
16	NA	NA	NA
15	1	0.04914	0.8275

```r
pander(linearHypothesis(mod1,"e1_tax_AR = -1.43426",test="FT"),big.mark=",")
```

Table 22: Linear hypothesis test (Theil's F test)

Res.Df	Df	F	Pr(>F)
16	NA	NA	NA
15	1	1.035	0.325

Sono respinte le ipotesi che i parametri siano cambiati. Si testa quindi l'ipotesi che non siano mutati i valori dei parametri per "cpl_CA" nella seconda equazione SURE.

```
#-- R CODE
pander(linearHypothesis(mod1,"e2_cpi_CA = -67.7451",test="FT"),big.mark=",")
```

Table 23: Linear hypothesis test (Theil's F test)

Res.Df	Df	F	Pr(>F)
16	NA	NA	NA
15	1	0.1796	0.6777

Anche in questo caso l'ipotesi è respinta.

ESERCIZIO 2

Si analizza ora per Texas e California la dipendenza lineare della variabile "packpc" da "income" e "avgprs".

```
#-- R CODE
mod1_CA <- lm(packpc_CA ~ income_CA + avgprs_CA, d2)
pander(summary(mod1_CA),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	161.9	5.336	30.34	1.513e-09
income_CA	-1.219e-07	2.077e-08	-5.87	0.0003742
avgprs_CA	-0.05424	0.05023	-1.08	0.3117

Table 25: Fitting linear model: packpc_CA ~ income_CA + avgprs_CA

Observations	Residual Std. Error	R^2	Adjusted R^2
11	1.959	0.9877	0.9846

```
pander(anova(mod1_CA),big.mark=",")
```

Table 26: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income_CA	1	2,462	2,462	641.6	6.322e-09
avgprs_CA	1	4.475	4.475	1.166	0.3117
Residuals	8	30.7	3.838	NA	NA

```
mod1_TX <- lm(packpc_TX ~ income_TX + avgprs_TX, d2)
pander(summary(mod1_TX),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	159.2	4.625	34.43	5.534e-10
income_TX	-4.62e-08	3.827e-08	-1.207	0.2618
avgprs_TX	-0.3409	0.06176	-5.519	0.0005607

Table 28: Fitting linear model: packpc_TX ~ income_TX + avgprs_TX

Observations	Residual Std. Error	R^2	Adjusted R^2
11	2.848	0.9732	0.9665

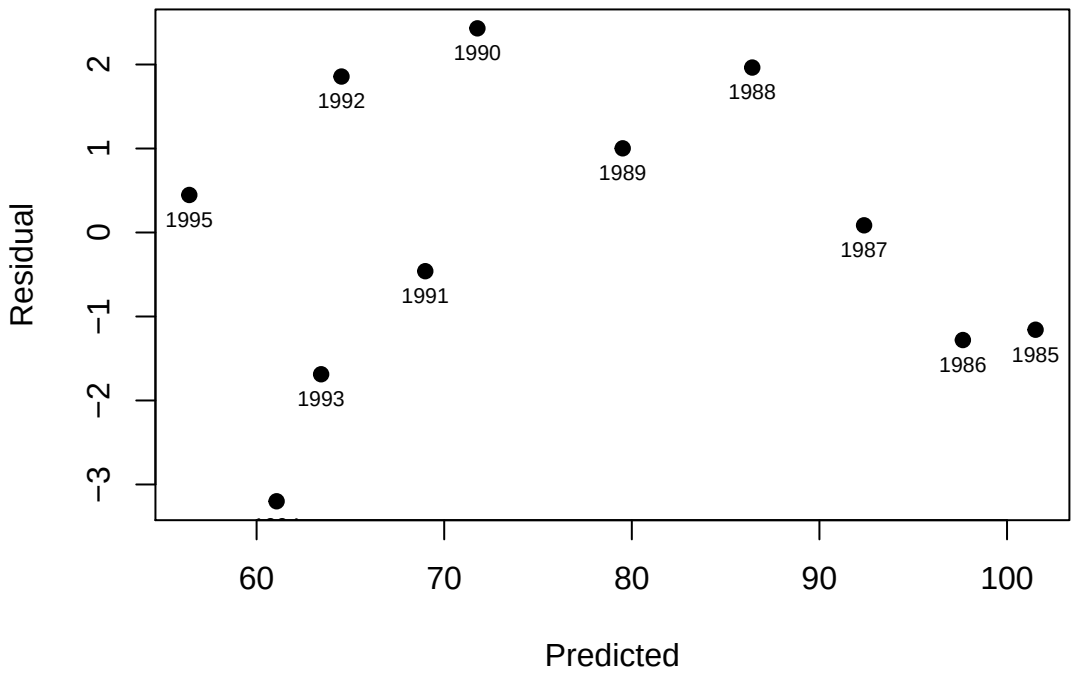
```
pander(anova(mod1_TX),big.mark=",")
```

Table 29: Analysis of Variance Table

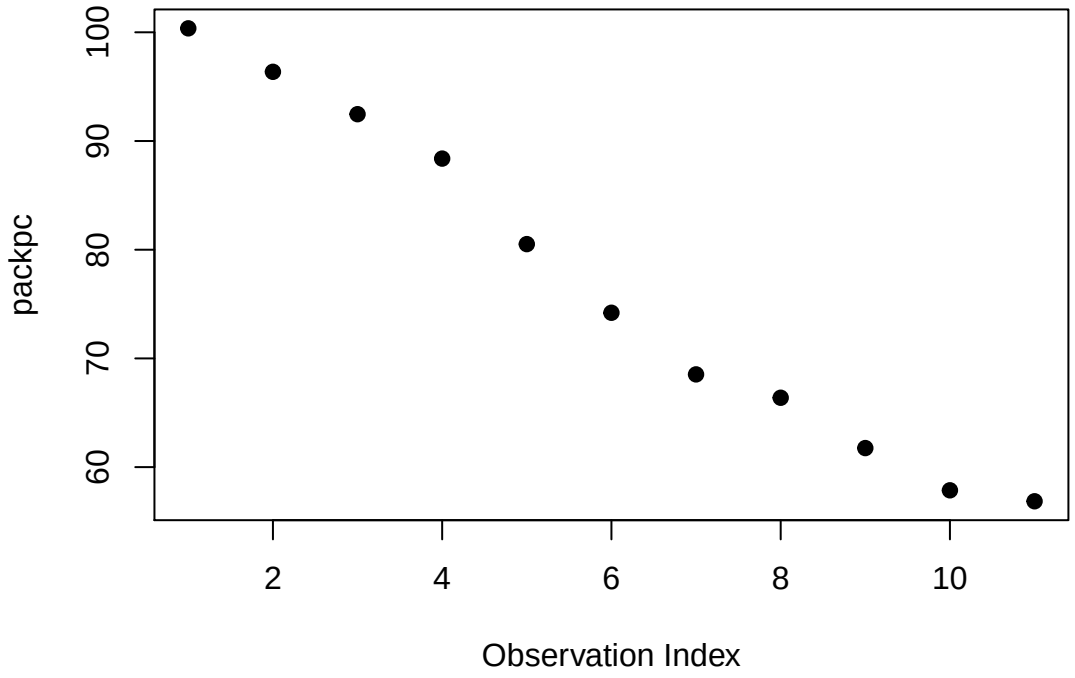
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income_TX	1	2,113	2,113	260.5	2.182e-07
avgprs_TX	1	247.2	247.2	30.46	0.0005607
Residuals	8	64.91	8.114	NA	NA

In entrambe le equazioni il fitting à altissimo ma mentre per la California l'unica variabile con parametro significativo è "income" nel Texas è "avgprs". Si verifica ora omoschedasticità e incorrelazione dei residui.

```
#-- R CODE
plot(fitted(mod1_CA),resid(mod1_CA),pch=19,xlab="Predicted",ylab="Residual")
text(fitted(mod1_CA),resid(mod1_CA),d2$year_TX,pos=1,cex=0.7)
```




```
plot(1:length(d2$packpc_CA),d2$packpc_CA,pch=19,xlab="Observation Index",ylab="packpc")
```




```
#-- R CODE
pander(white.test(mod1_CA),big.mark=",")
```

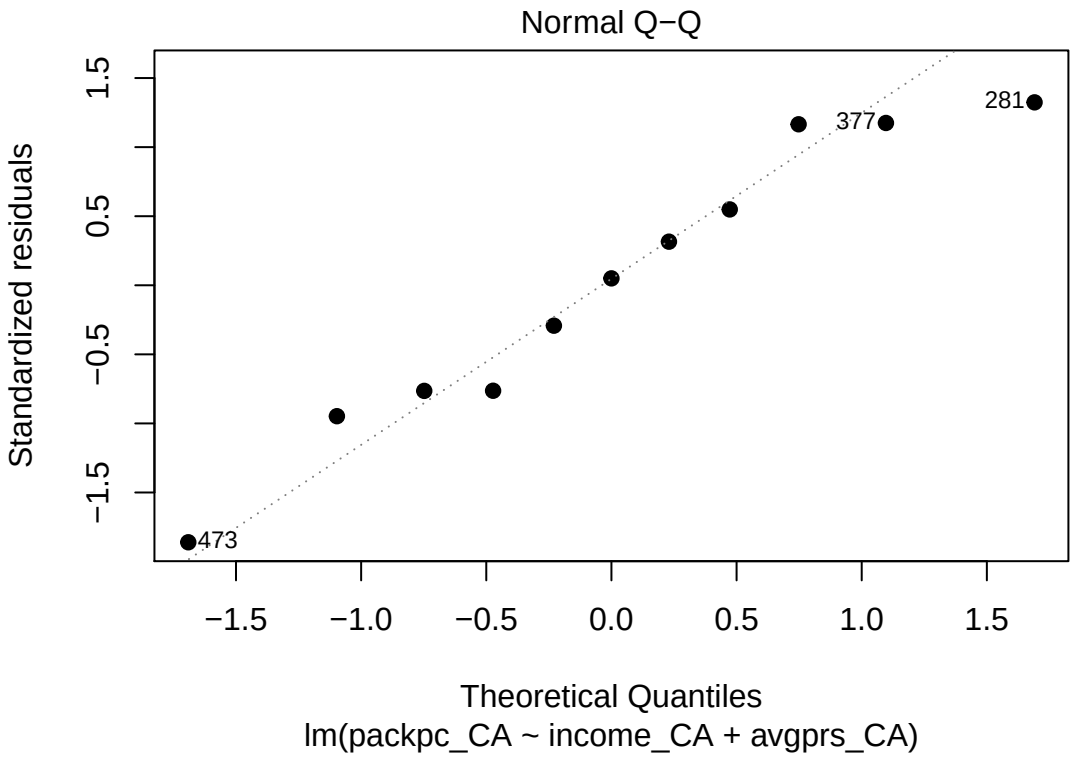
Test.statistic	P.value
65.39	6.328e-15

```
pander(dwtest(mod1_CA),big.mark=",")
```

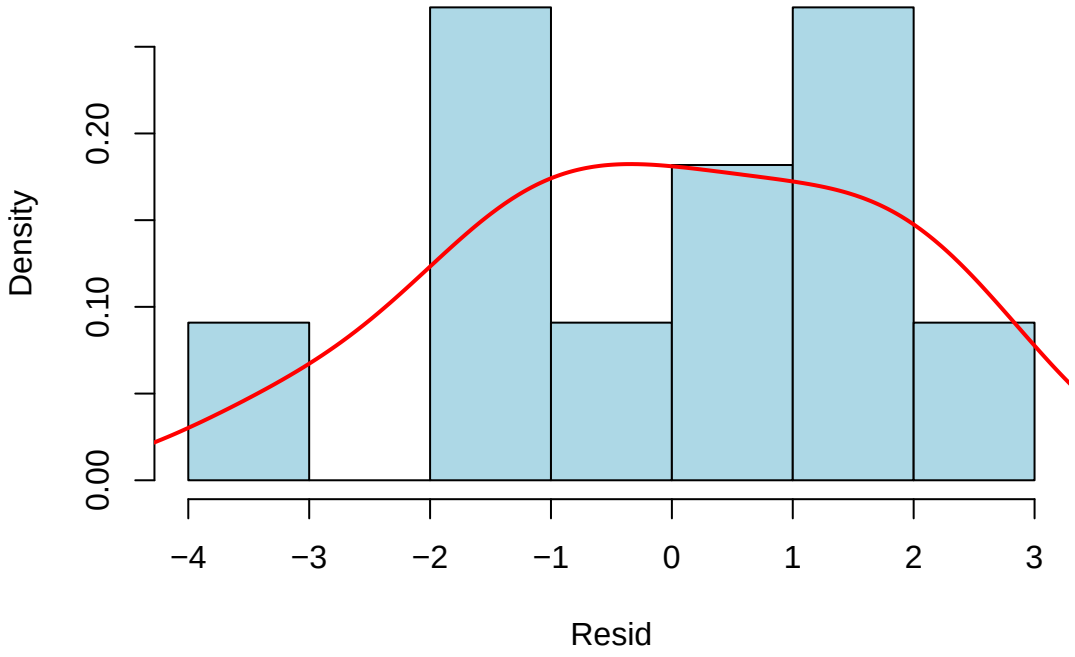
Table 31: Durbin-Watson test: `mod1_CA`

Test statistic	P value	Alternative hypothesis
1.637	0.08236	true autocorrelation is greater than 0

```
#-- R CODE
plot(mod1_CA,which=2,pch=19)
```

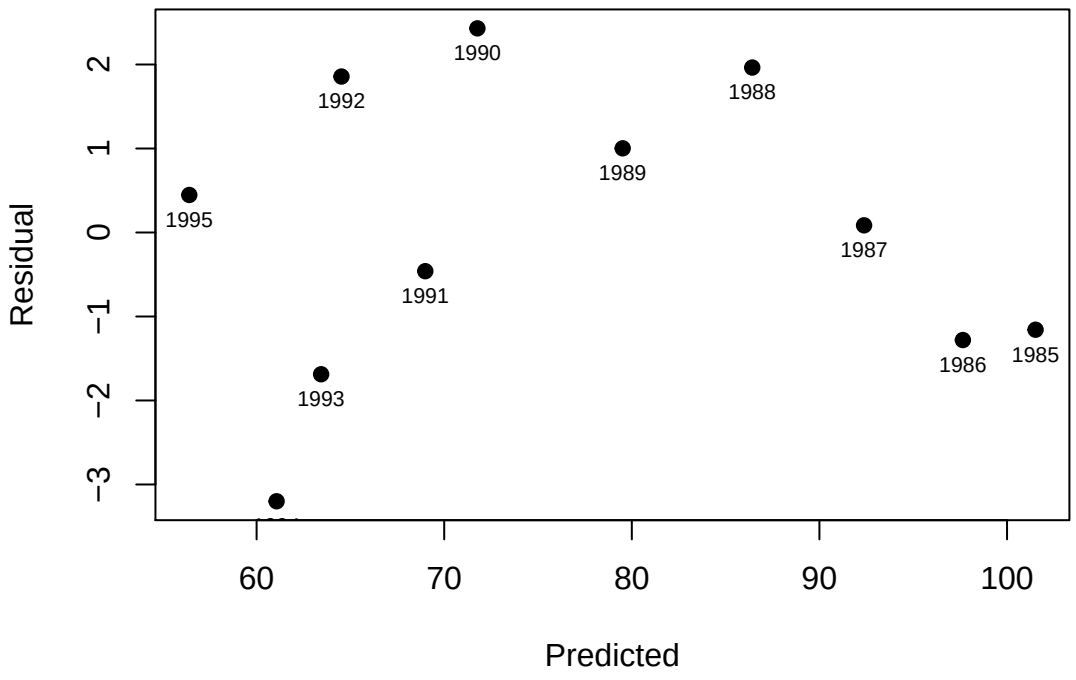



```
hist(resid(mod1_CA),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1_CA)),col=2,lwd=2)
```

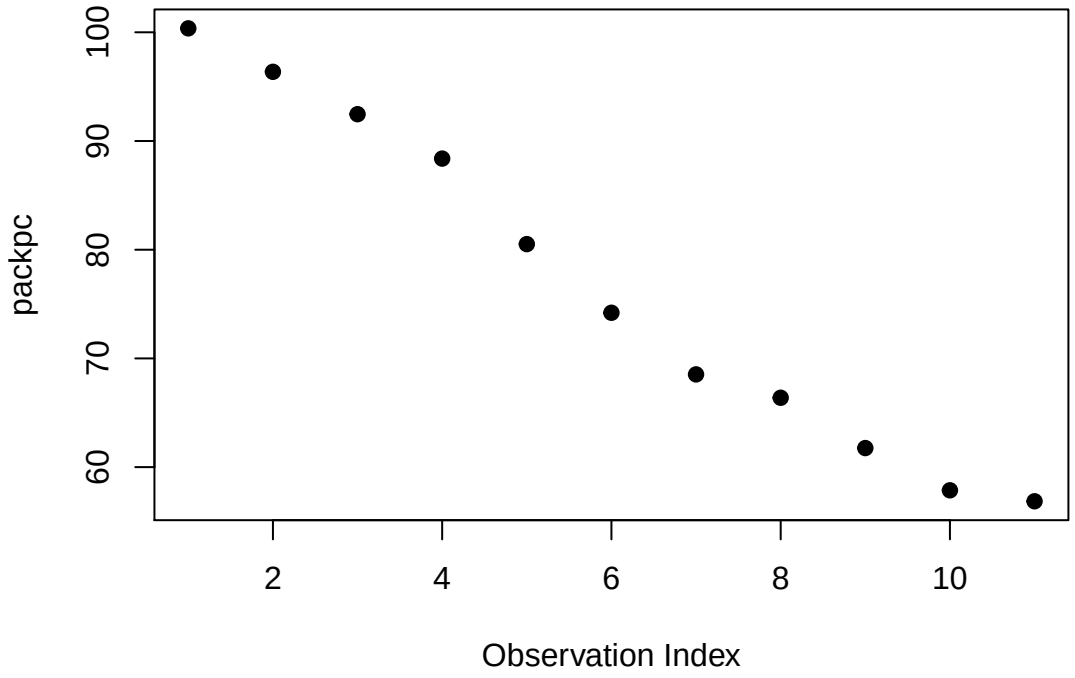


I grafici indicano che i residui sono omoschedastici mentre il valore del test di Durbin-Watson mostra che l'ipotesi di non correlazione è da accettare.
Sulla normalità si osservano problemi: c'è uno scostamento abbastanza netto della distribuzione empirica da quella teorica. Anche in questo caso necessiterebbe una opportuna correzione. Si consideri ora la seconda equazione.

```
#-- R CODE
plot(fitted(mod1_CA),resid(mod1_CA),pch=19,xlab="Predicted",ylab="Residual")
text(fitted(mod1_CA),resid(mod1_CA),d2$year_TX,pos=1,cex=0.7)
```




```
plot(1:length(d2$packpc_CA),d2$packpc_CA,pch=19,xlab="Observation Index",ylab="packpc")
```




```
#-- R CODE
pander(white.test(mod1_TX),big.mark=",")
```

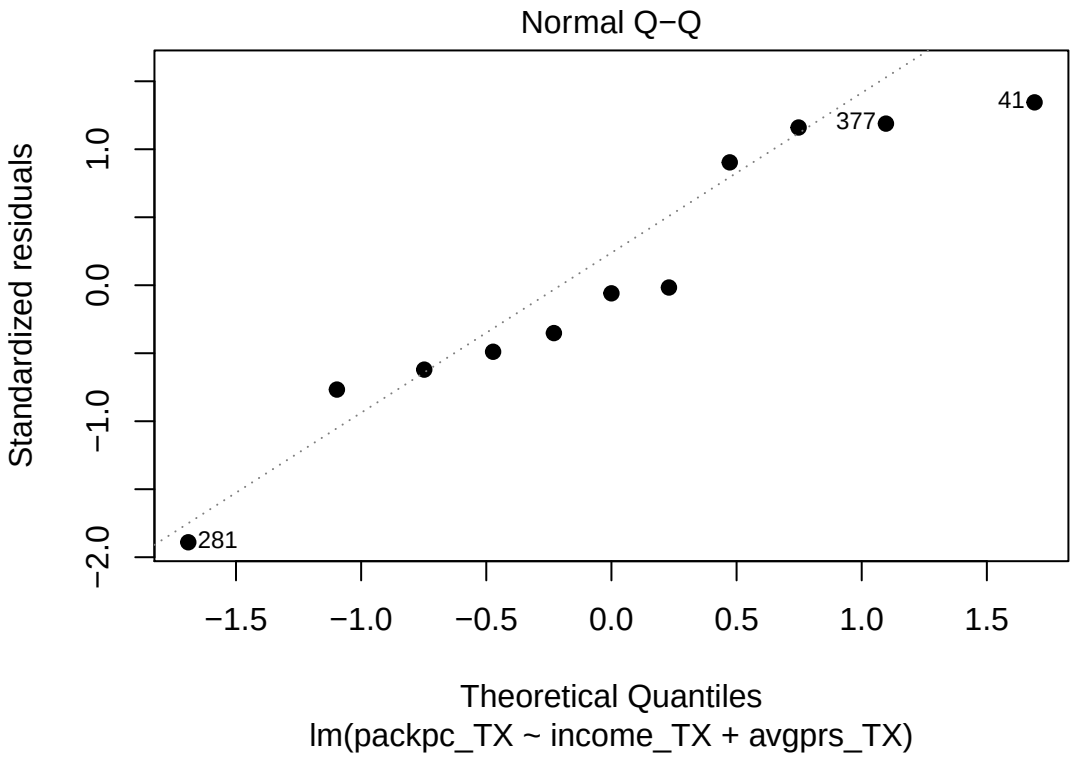
Test.statistic	P.value
29.25	4.458e-07

```
pander(dwtest(mod1_TX),big.mark=",")
```

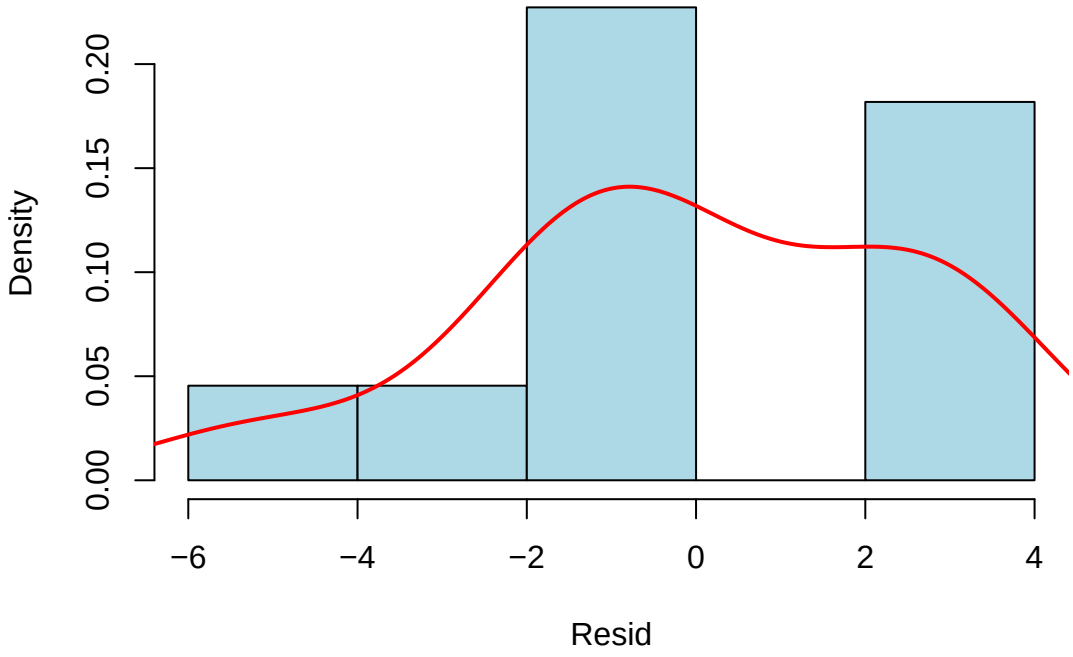
Table 33: Durbin-Watson test: `mod1_TX`

Test statistic	P value	Alternative hypothesis
1.414	0.02881 *	true autocorrelation is greater than 0

```
#-- R CODE
plot(mod1_TX,which=2,pch=19)
```




```
hist(resid(mod1_TX),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1_TX)),col=2,lwd=2)
```



I tre grafici (valori residui- predetti; residui-osservati, predetti-osservati) indicano che i residui sono omoschedastici mentre il valore del test di Durbin-Watson mostra che l'ipotesi di non correlazione è da accettare. Sulla normalità si osservano problemi: c'è uno scostamento abbastanza netto della distribuzione empirica da quella teorica. Anche in questo caso necessiterebbe una opportuna correzione.

Si propone ora il modello multivariato con le stesse variabili esplicative e errori correlati per medesimo individuo.

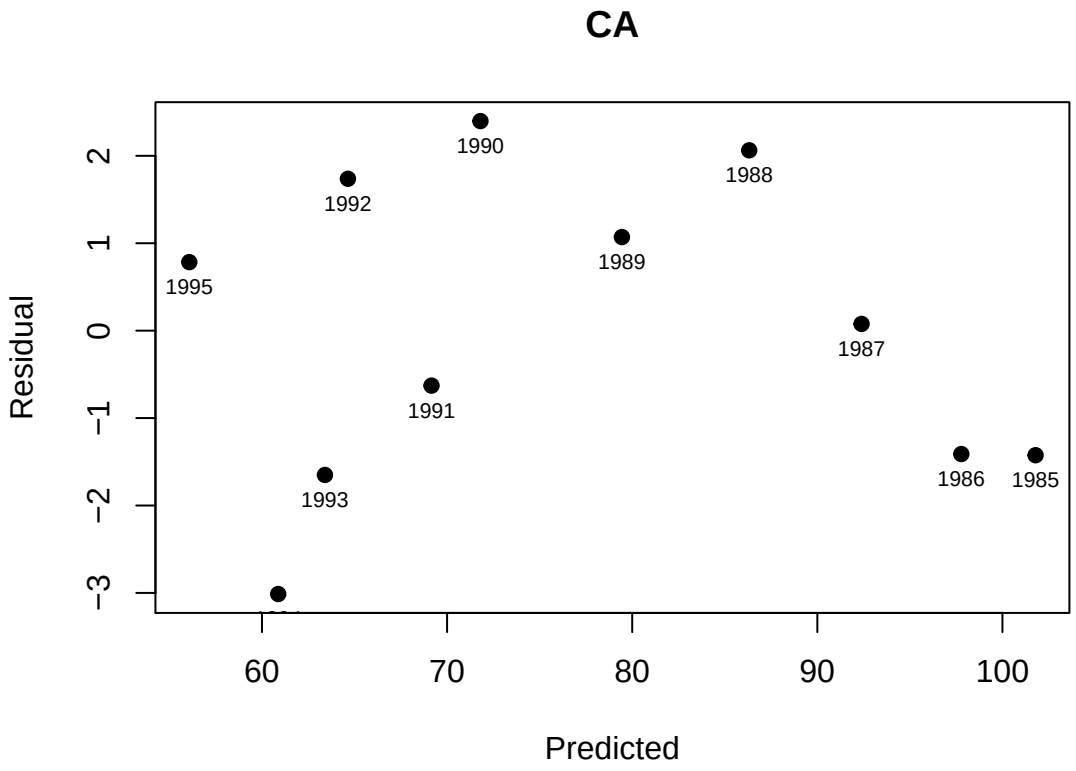
```
#-- R CODE
e1 <- packpc_CA ~ income_CA + avgprs_CA
e2 <- packpc_TX ~ income_TX + avgprs_TX
sistema <- list(e1=e1,e2=e2)

mod3 <- systemfit(sistema,"SUR",data=d2)
summary(mod3)
```

```
##
## systemfit results
## method: SUR
##
##          N DF     SSR detRCov   OLS-R2 McElroy-R2
## system 22 16 97.7515 23.5213 0.980144   0.991559
##
##     N DF     SSR     MSE    RMSE       R2   Adj R2
## e1 11  8 31.0017 3.87522 1.96856 0.987587 0.984484
## e2 11  8 66.7497 8.34371 2.88855 0.972479 0.965599
```

```
##
## The covariance matrix of the residuals used for estimation
##          e1       e2
## e1  3.83806 -2.42561
## e2 -2.42561  8.11369
##
## The covariance matrix of the residuals
##          e1       e2
## e1  3.87522 -2.96857
## e2 -2.96857  8.34372
##
## The correlations of the residuals
##           e1        e2
## e1  1.000000 -0.522059
## e2 -0.522059  1.000000
##
##
## SUR estimates for 'e1' (equation 1)
## Model Formula: packpc_CA ~ income_CA + avgprs_CA
##
##                 Estimate   Std. Error  t value   Pr(>|t|)
## (Intercept)  1.63343e+02  5.11283e+00 31.94758 1.0035e-09 ***
## income_CA   -1.27253e-07  1.95504e-08 -6.50899 0.00018633 ***
## avgprs_CA   -4.32075e-02  4.73160e-02 -0.91317 0.38784439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.968557 on 8 degrees of freedom
## Number of observations: 11 Degrees of Freedom: 8
## SSR: 31.001728 MSE: 3.875216 Root MSE: 1.968557
## Multiple R-Squared: 0.987587 Adjusted R-Squared: 0.984484
##
##
## SUR estimates for 'e2' (equation 2)
## Model Formula: packpc_TX ~ income_TX + avgprs_TX
##
##                 Estimate   Std. Error  t value   Pr(>|t|)
## (Intercept)  1.60196e+02  4.57059e+00 35.04929 4.8044e-10 ***
## income_TX   -6.43707e-08  3.60162e-08 -1.78727 0.11170404
## avgprs_TX   -3.12964e-01  5.84008e-02 -5.35889 0.00067857 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.888549 on 8 degrees of freedom
## Number of observations: 11 Degrees of Freedom: 8
## SSR: 66.749723 MSE: 8.343715 Root MSE: 2.888549
## Multiple R-Squared: 0.972479 Adjusted R-Squared: 0.965599
```

In questo caso vi sono piccoli cambiamenti nei valori dei parametri che non modificano la loro significatività né il livello del p-value. Si verifica ora se gli errori sono omoschedastici e incorrelati all'interno di ogni equazione.

```
#-- R CODE
plot(fitted(mod3)[,1],resid(mod3)[,1],pch=19,xlab="Predicted",ylab="Residual",main="CA")
text(fitted(mod3)[,1],resid(mod3)[,1],d2$year_TX,pos=1,cex=0.7)
```

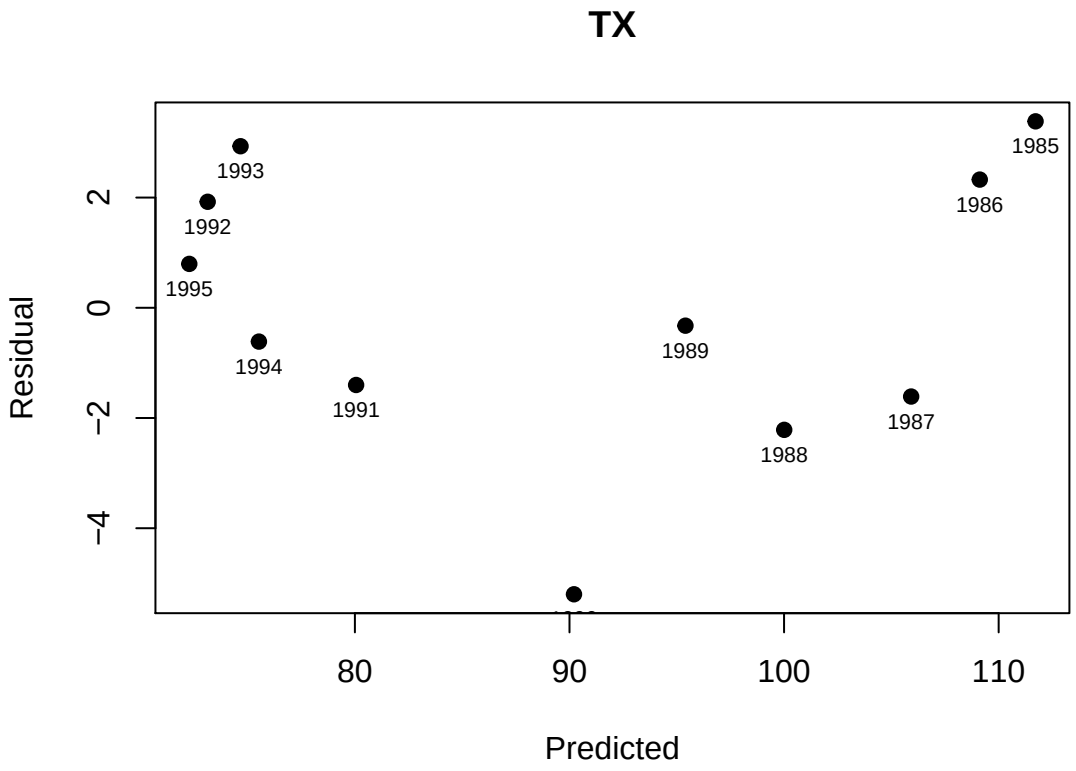



```
#-- R CODE
pander(dwtest(mod3[[1]][[1]]),big.mark=",")
```

Table 34: Durbin-Watson test: `mod3[[1]][[1]]`

Test statistic	P value	Alternative hypothesis
1.637	0.08236	true autocorrelation is greater than 0

```
#-- R CODE
plot(fitted(mod3)[,2],resid(mod3)[,2],pch=19,xlab="Predicted",ylab="Residual",main="TX")
text(fitted(mod3)[,2],resid(mod3)[,2],d2$year_TX,pos=1,cex=0.7)
```




```
#-- R CODE
pander(dwtest(mod3[[1]][[2]]),big.mark=",")
```

Table 35: Durbin-Watson test: `mod3[[1]][[2]]`

Test statistic	P value	Alternative hypothesis
1.414	0.02881 *	true autocorrelation is greater than 0

I grafici indicano che i residui sono omoschedastici all'interno di ogni equazione mentre il valore del test di Durbin-Watson mostra che l'ipotesi di non correlazione è da accettare.

ESERCIZIO 3

Si consideri ora il modello con regressori diversi: "income" per California e "avgprs" per Texas. Iniziamo con le stime OLS

```
#-- R CODE
```

```
mod1_CA <- lm(packpc_CA ~ income_CA, d2)
pander(summary(mod1_CA),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	166.1	3.613	45.98	5.445e-12
income_CA	-1.435e-07	5.716e-09	-25.1	1.215e-09

Table 37: Fitting linear model: packpc_CA ~ income_CA

Observations	Residual Std. Error	R^2	Adjusted R^2
11	1.977	0.9859	0.9843

```
pander(anova(mod1_CA),big.mark=",")
```

Table 38: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income_CA	1	2,462	2,462	630	1.215e-09
Residuals	9	35.18	3.909	NA	NA

```
mod1_TX <- lm(packpc_TX ~ avgprs_TX, d2)
pander(summary(mod1_TX),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	156.5	4.111	38.06	2.96e-11
avgprs_TX	-0.4095	0.02467	-16.6	4.672e-08

Table 40: Fitting linear model: packpc_TX ~ avgprs_TX

Observations	Residual Std. Error	R^2	Adjusted R^2
11	2.92	0.9684	0.9648

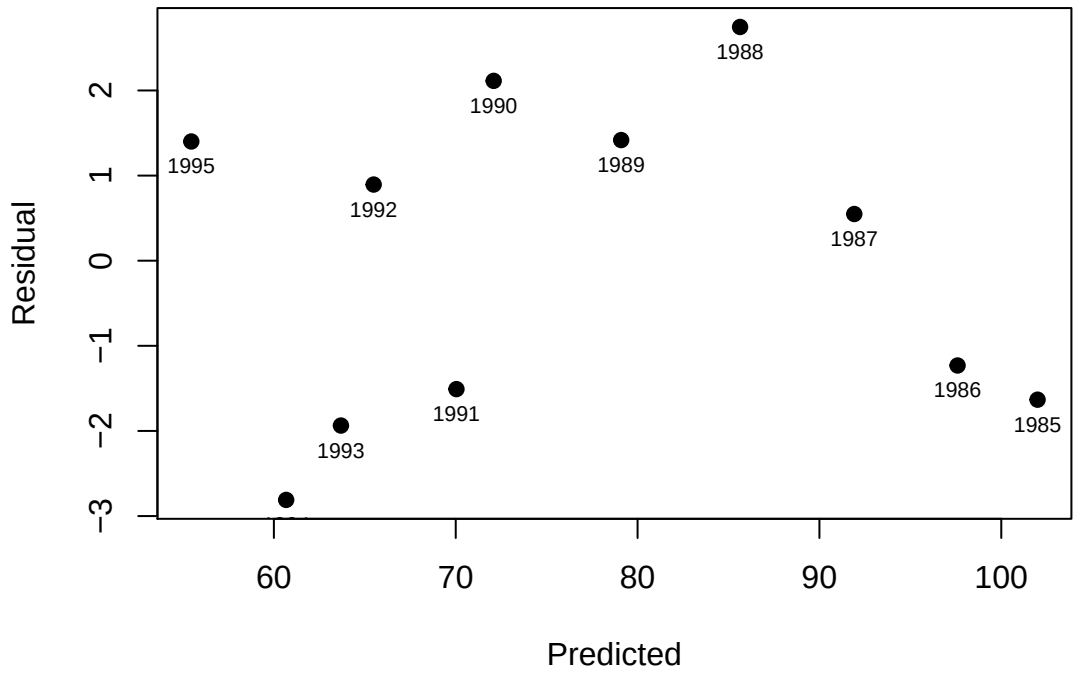
```
pander(anova(mod1_TX),big.mark=",")
```

Table 41: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
avgprs_TX	1	2,349	2,349	275.5	4.672e-08
Residuals	9	76.74	8.526	NA	NA

I modelli offrono entrambi un fitting molto elevato e le variabili esplicative hanno entrambe per la prima volta una significatività legata a un p-value pari a inferiore a 0.0001 e hanno un legame negativo e più rilevante che nel modello precedente sulle rispettiva variabili dipendenti. Verifichiamo ora omoschedasticità e incorrelazione degli errori.

```
#-- R CODE
plot(fitted(mod1_CA),resid(mod1_CA),pch=19,xlab="Predicted",ylab="Residual")
text(fitted(mod1_CA),resid(mod1_CA),d2$year_TX,pos=1,cex=0.7)
```




```
#-- R CODE
pander(white.test(mod1_CA),big.mark=",")
```

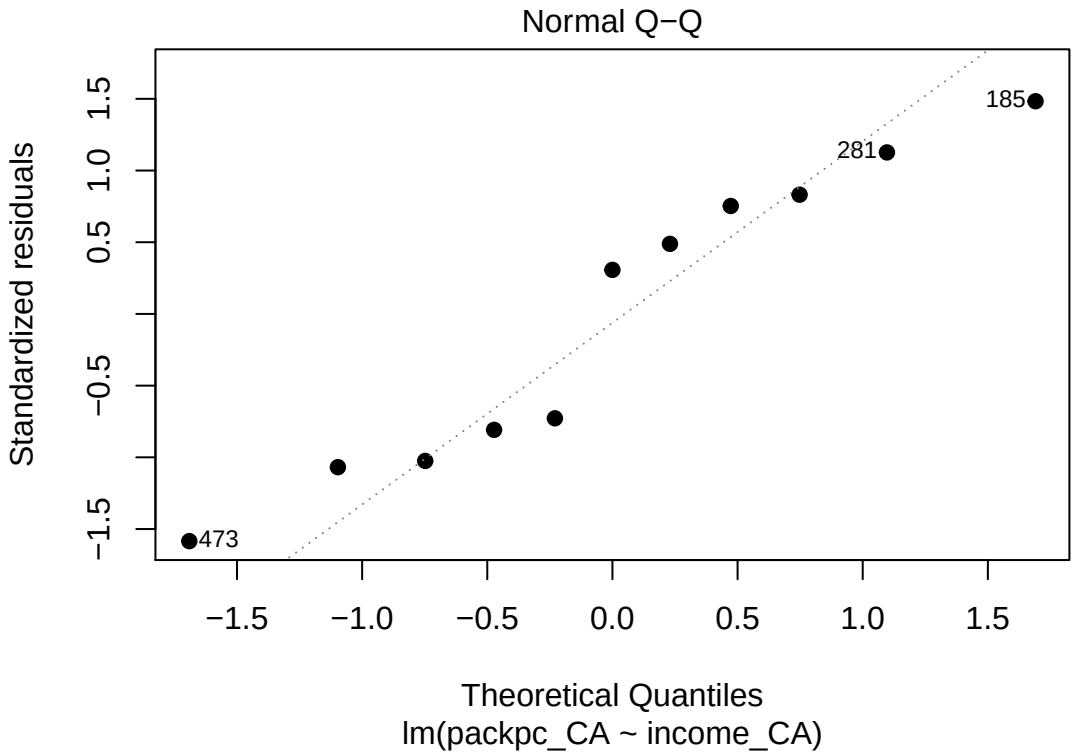
Test.statistic	P.value
31.9	1.186e-07

```
pander(dwtest(mod1_CA),big.mark=",")
```

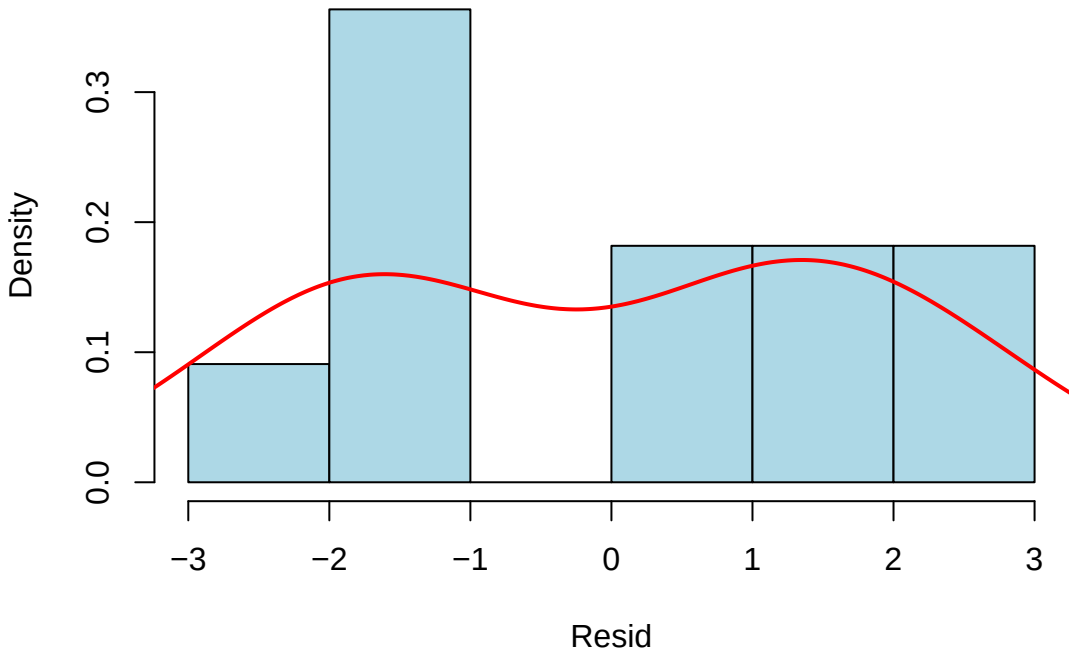
Table 43: Durbin-Watson test: `mod1_CA`

Test statistic	P value	Alternative hypothesis
1.586	0.1285	true autocorrelation is greater than 0

```r
#-- R CODE
plot(mod1_CA,which=2,pch=19)
```




```r
hist(resid(mod1_CA),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1_CA)),col=2,lwd=2)
```



Nell'equazione relativa alla California si conferma anzi migliora ancora la situazione di omoschedasticità ma anche la non correlazione tra gli errori è verificata. Vediamo ora per il Texas.

```
#-- R CODE
pander(white.test(mod1_TX),big.mark=",")
```

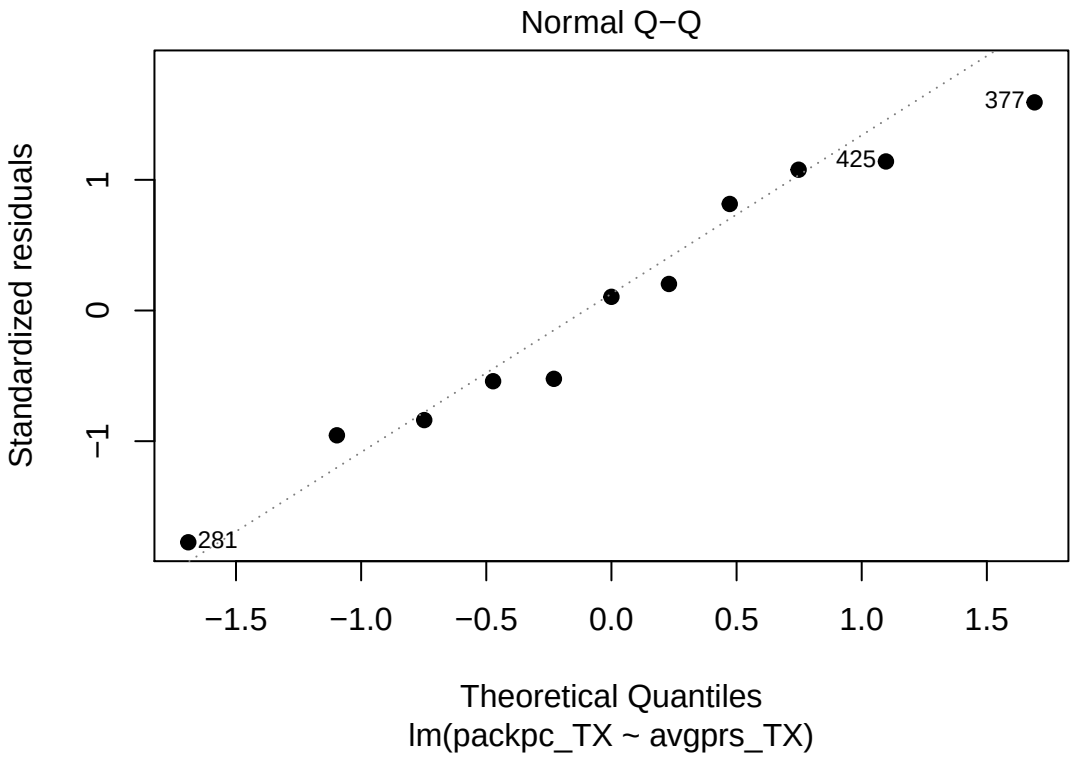
Test.statistic	P.value
39.78	2.301e-09

```
pander(dwtest(mod1_TX),big.mark=",")
```

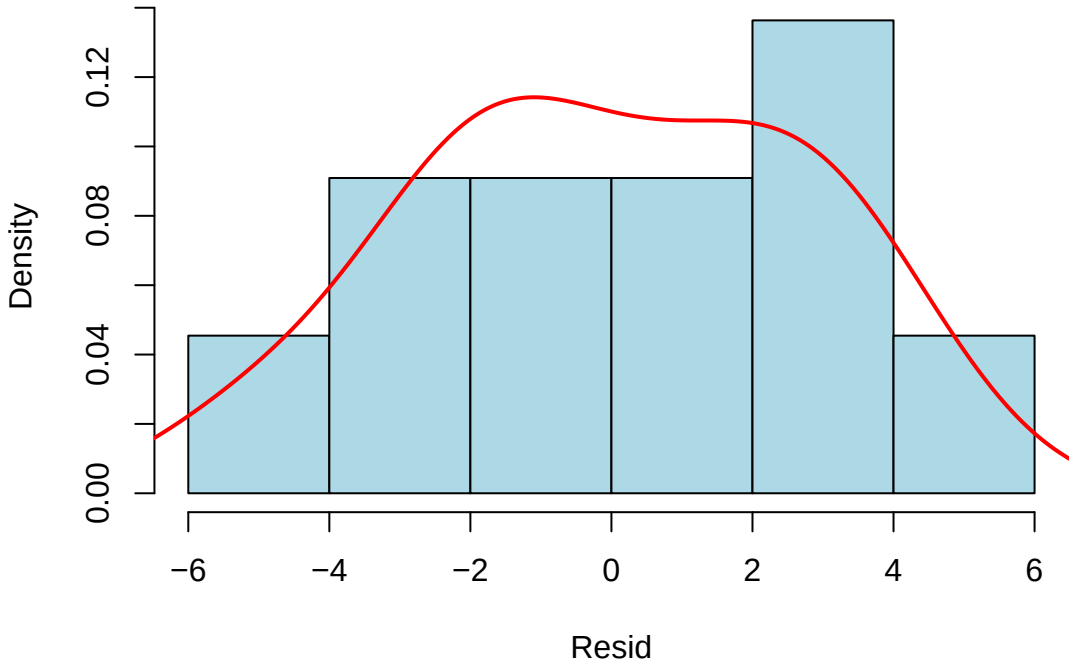
Table 45: Durbin-Watson test: `mod1_TX`

Test statistic	P value	Alternative hypothesis
1.527	0.1083	true autocorrelation is greater than 0

```
#-- R CODE
plot(mod1_TX,which=2,pch=19)
```




```
hist(resid(mod1_TX),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1_TX)),col=2,lwd=2)
```



> >

Anche in questo caso si può concludere in linea di massima che esiste omoschedasticità anche se il grafico valori predetti-osservati si mostra nella parte finale molto distante da una situazione di ottimalità in cui esista una chiara relazione lineare tra detti valori. Si conferma inoltre la non correlazione tra gli errori.

Passiamo ora alla stima Sure.

```
#-- R CODE
e1 <- packpc_CA ~ income_CA
e2 <- packpc_TX ~ avgprs_TX
sistema <- list(e1=e1,e2=e2)

mod4 <- systemfit(sistema,"SUR",data=d2)
summary(mod4)
```

```
##
## systemfit results
## method: SUR
##
##          N DF     SSR detRCov   OLS-R2 McElroy-R2
## system 22 18 112.192 27.9028 0.977211   0.988147
##
##     N DF     SSR     MSE    RMSE       R2   Adj R2
## e1 11  9 35.4547 3.93941 1.98479 0.985804 0.984227
## e2 11  9 76.7378 8.52642 2.92000 0.968361 0.964846
##
## The covariance matrix of the residuals used for estimation
```

```
##          e1       e2
## e1  3.90883 -2.26904
## e2 -2.26904  8.52641
##
## The covariance matrix of the residuals
##          e1       e2
## e1  3.93941 -2.38459
## e2 -2.38459  8.52642
##
## The correlations of the residuals
##           e1        e2
## e1  1.000000 -0.411447
## e2 -0.411447  1.000000
##
##
## SUR estimates for 'e1' (equation 1)
## Model Formula: packpc_CA ~ income_CA
##
##                 Estimate   Std. Error  t value   Pr(>|t|)
## (Intercept)  1.67088e+02  3.58932e+00  46.5514 4.8765e-12 ***
## income_CA   -1.44988e-07  5.67744e-09 -25.5375 1.0421e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.984794 on 9 degrees of freedom
## Number of observations: 11 Degrees of Freedom: 9
## SSR: 35.454665 MSE: 3.939407 Root MSE: 1.984794
## Multiple R-Squared: 0.985804 Adjusted R-Squared: 0.984227
##
##
## SUR estimates for 'e2' (equation 2)
## Model Formula: packpc_TX ~ avgprs_TX
##
##                Estimate  Std. Error  t value   Pr(>|t|)
## (Intercept) 156.4789624   4.0839748  38.3154 2.7903e-11 ***
## avgprs_TX    -0.4096167   0.0245072 -16.7141 4.3931e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.920004 on 9 degrees of freedom
## Number of observations: 11 Degrees of Freedom: 9
## SSR: 76.737807 MSE: 8.526423 Root MSE: 2.920004
## Multiple R-Squared: 0.968361 Adjusted R-Squared: 0.964846
```

I valori dei parametri non cambiano quasi per niente. Viene respinta invece l'ipotesi che i parametri relativi a "income" e "avgprs" rispettivamente della prima e seconda equazione siano identici.