GLS 4 - Data set: LONGLEY

INTRODUZIONE

I dati sono relativi ad una serie di dati macroeconomici che forniscono un esempio di regressione altamente collineare. I dati si riferiscono al periodo 1947-1962. Le variabili sono le seguenti:

1. GNP: prodotto nazionale lordo
2. UNEMPLOYED: numero di disoccupati
3. ARMED.FORCES: numero di persone nelle forze armate
4. POPULATION: numero di popolazione non istituzionalizzata con più di 14 anni
5. EMPLOYED: numero di persone impiegate
6. TIME: anni trascorsi dal 1947

Analisi proposte:

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

```r
#-- R CODE

library(Hmisc)
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)\\4.Error-GLS\\longley.

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("GNP.deflator","GNP","Unemployed","Armed.Forces","Population","Employed")
```

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

Table 1: Table continues below

year	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year
1,947	83	234.3	235.6	159	107.6	1,947
1,948	88.5	259.4	232.5	145.6	108.6	1,948
1,949	88.2	258.1	368.2	161.6	109.8	1,949
1,950	89.5	284.6	335.1	165	110.9	1,950
1,951	96.2	329	209.9	309.9	112.1	1,951
1,952	98.1	347	193.2	359.4	113.3	1,952

Employed
60.32
61.12
60.17
61.19
63.22
63.64

STATISTICHE DESCRITTIVE

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

Table 3: Table continues below

GNP.deflator	GNP	Unemployed	Armed.Forces	Population
Min. : 83.00	Min. :234.3	Min. :187.0	Min. :145.6	Min. :107.6
1st Qu.: 94.53	1st Qu.:317.9	1st Qu.:234.8	1st Qu.:229.8	1st Qu.:111.8
Median :100.60	Median :381.4	Median :314.4	Median :271.8	Median :116.8
Mean :101.68	Mean :387.7	Mean :319.3	Mean :260.7	Mean :117.4
3rd Qu.:111.25	3rd Qu.:454.1	3rd Qu.:384.2	3rd Qu.:306.1	3rd Qu.:122.3
Max. :116.90	Max. :554.9	Max. :480.6	Max. :359.4	Max. :130.1

Employed
Min. :60.17
1st Qu.:62.71
Median :65.50
Mean :65.32
3rd Qu.:68.29
Max. :70.55

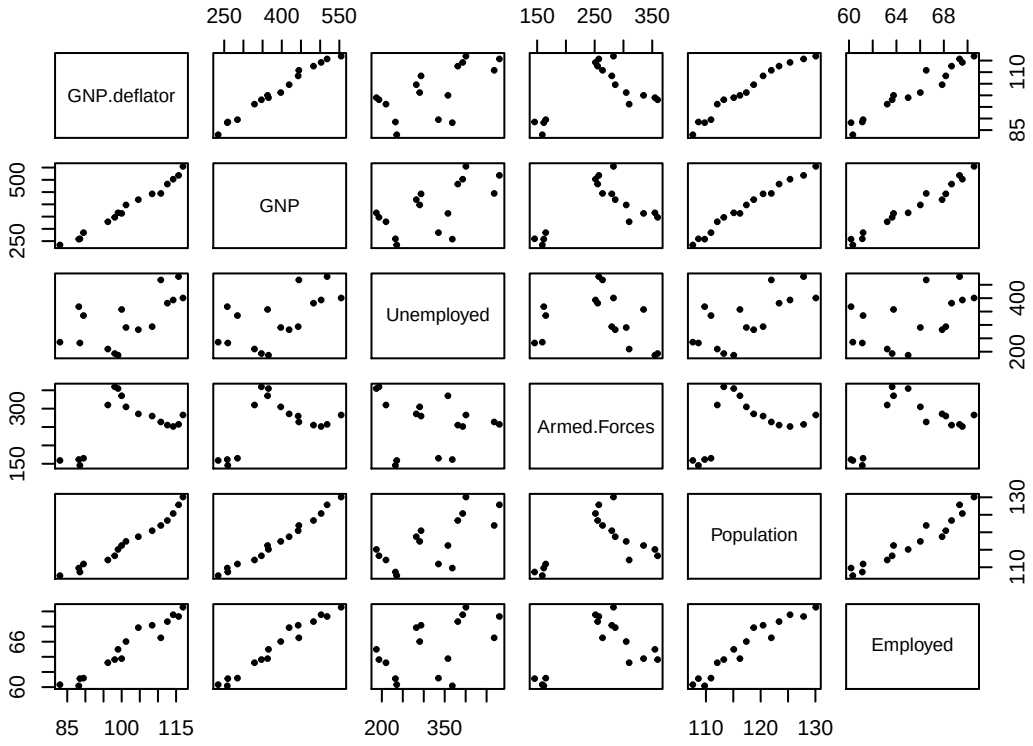
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

Table 5: Table continues below

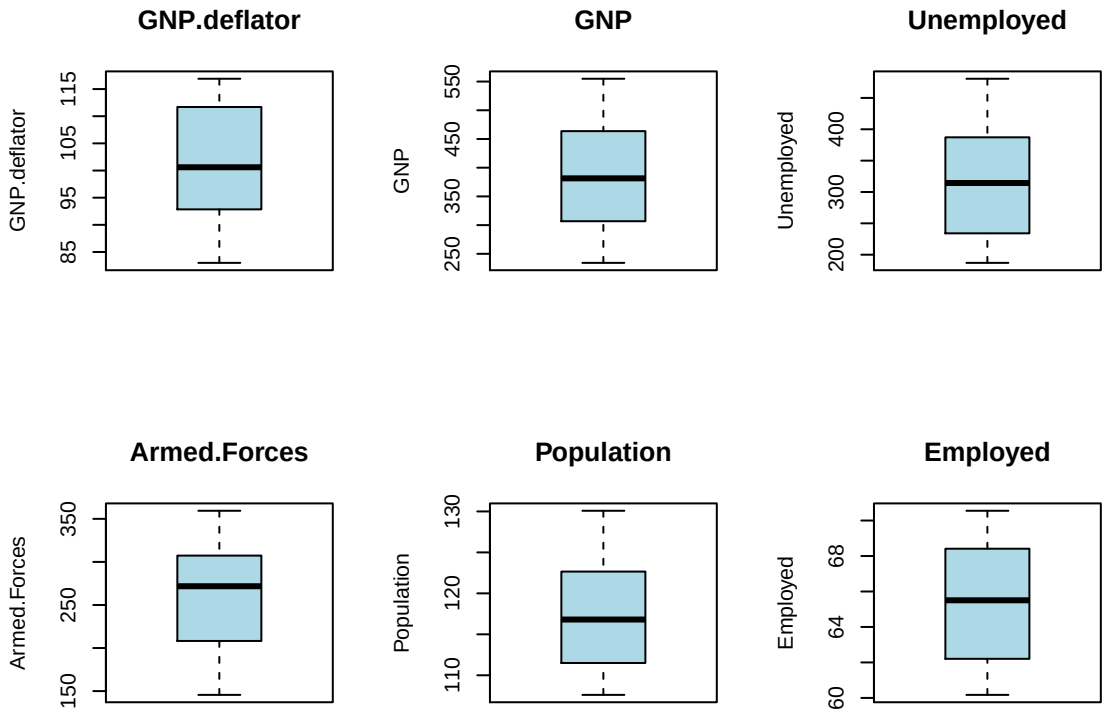
	GNP.deflator	GNP	Unemployed	Armed.Forces
GNP.deflator	1	0.9916	0.6206	0.4647
GNP	0.9916	1	0.6043	0.4464
Unemployed	0.6206	0.6043	1	-0.1774
Armed.Forces	0.4647	0.4464	-0.1774	1
Population	0.9792	0.9911	0.6866	0.3644
Employed	0.9709	0.9836	0.5025	0.4573

	Population	Employed
GNP.deflator	0.9792	0.9709
GNP	0.9911	0.9836
Unemployed	0.6866	0.5025
Armed.Forces	0.3644	0.4573
Population	1	0.9604
Employed	0.9604	1

```
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)
}
```



REGRESSIONE

Data questa situazione si utilizzano come variabili esplicative rispetto a "GNP" solo "Unemployed", "Armed Forces", "Population", "Employed". Gli errori sono omoschedastici secondo il test di White. Il fitting è altissimo ma le uniche variabili veramente significative sono "Population" e "Employed".

```r
#-- R CODE
mod1 <- lm(GNP ~ Unemployed + Armed.Forces + Population + Employed, d) #-- stima modello lineare sempli
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1,339	35.12	-38.14	4.871e-13
Unemployed	0.000343	0.04055	0.008459	0.9934
Armed.Forces	0.08245	0.02955	2.79	0.01759
Population	9.338	1.504	6.211	6.615e-05
Employed	9.321	2.397	3.888	0.002527

Table 8: Fitting linear model: GNP ~ Unemployed + Armed.Forces + Population + Employed

Observations	Residual Std. Error	R^2	Adjusted R^2
16	5.683	0.9976	0.9967

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

Table 9: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Unemployed	1	54,109	54,109	1,676	2.247e-13
Armed.Forces	1	46,900	46,900	1,452	4.908e-13
Population	1	46,338	46,338	1,435	5.243e-13
Employed	1	488.2	488.2	15.12	0.002527
Residuals	11	355.2	32.29	NA	NA

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

Test.statistic	P.value
0.7167	0.6988

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

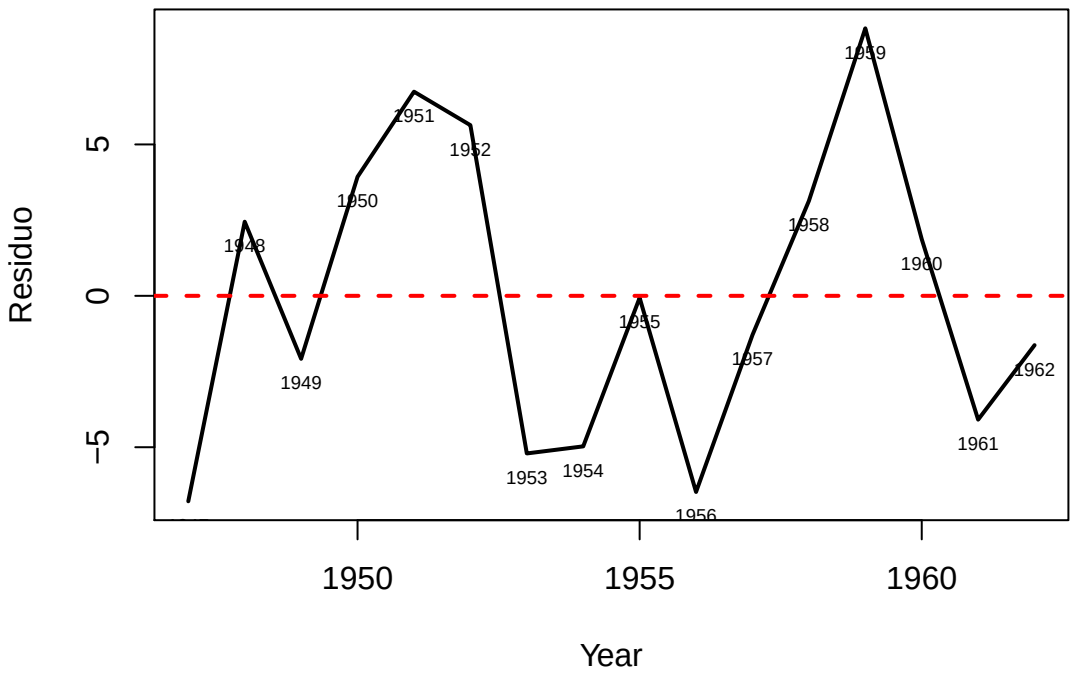
Table 11: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.416	0.01406 *	true autocorrelation is greater than 0

Il modello interpreta bene la variabile dipendente e il fitting è molto elevato. I parametri significativi sono quelli relativi a "partic" e "degrees". Gli errori sono normali come si evince dalla distribuzione dei residui.

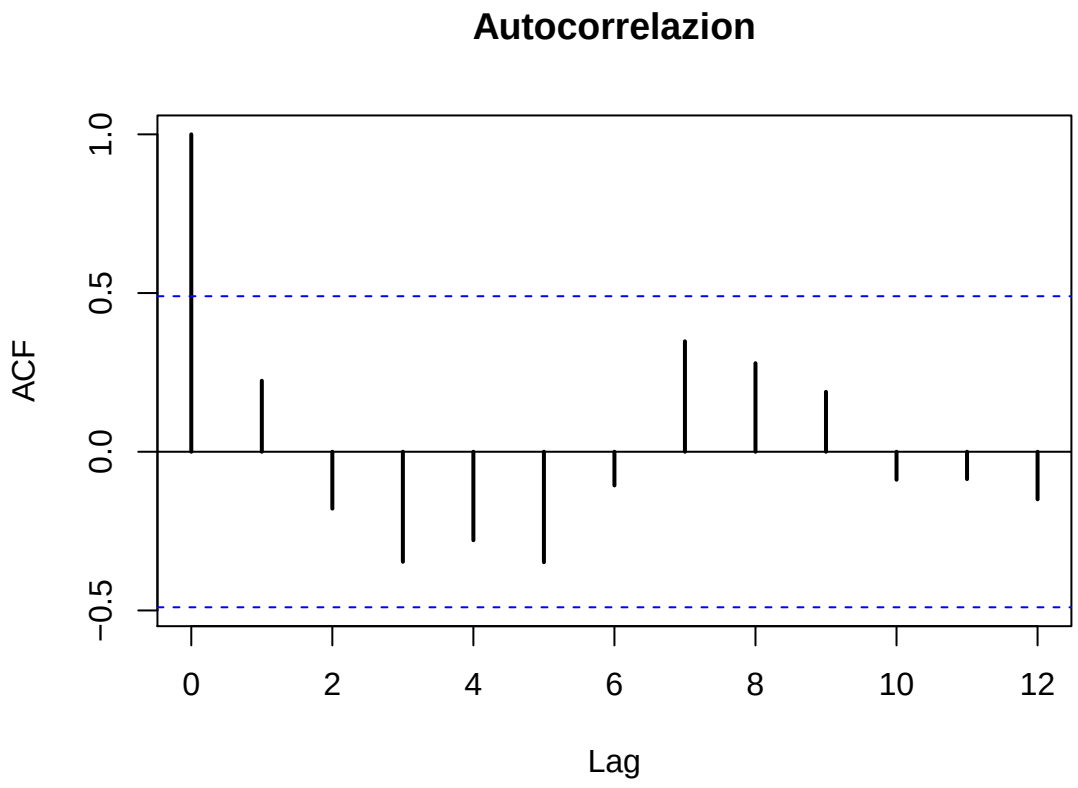
Dal grafico residui-anni in cui i residui non fluttuano intorno allo 0 si intuisce autocorrelazione di 1° grado positiva in quanto si vede un andamento sistematico di dipendenza tra gli errori e errori ritardati nello stesso verso. Poi si intuisce la presenza di un ciclo che presuppone autocorrelazione di ordine superiore. In ogni caso l'autocorrelazione positiva di ordine 1 è confermata dal test Durbin Watson che respinge l'ipotesi di non autocorrelazione di ordine 1.

```
#-- R CODE
plot(d$year,resid(mod1),pch=19,xlab="Year",ylab="Residuo",type="l",col=1,lwd=2)
text(d$year,resid(mod1),d$year,pos=1,cex=.6)
abline(h=0,lwd=2,lty=2,col=2)
```



Si propone quindi un modello con errori incorrelati costruendo variabili e errori incorrelati.

```
#-- R CODE
autocorr <- acf(resid(mod1),main="Autocorrelazion",lwd=2)
```




```
pander(data.frame(LAG=autocorr$lag,VALUE=autocorr$acf)[1:5,])
```

LAG	VALUE
0	1
1	0.2237
2	-0.1792
3	-0.3467
4	-0.2787

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

d1$GNP_t <- d1$GNP-0.2237*Lag(d1$GNP,1)
d1$Unemployed_t <- d1$Unemployed-0.2237*Lag(d1$Unemployed,1)
d1$Armed.Forces_t <- d1$Armed.Forces-0.2237*Lag(d1$Armed.Forces,1)
d1$Population_t <- d1$Population-0.2237*Lag(d1$Population,1)
d1$Employed_t <- d1$Employed-0.2237*Lag(d1$Employed,1)

d1$int_tild <- 1-0.2237
```

```
#-- R CODE
mod2 <- lm(GNP_t ~ 0 + int_tild + Unemployed_t + Armed.Forces_t + Population_t + Employed_t,d1)
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
int_tild	-1,330	35.44	-37.53	4.299e-12
Unemployed_t	-0.02533	0.03951	-0.6409	0.536
Armed.Forces_t	0.05779	0.03232	1.788	0.104
Population_t	9.876	1.51	6.541	6.546e-05
Employed_t	8.45	2.449	3.451	0.00622

Table 14: Fitting linear model: GNP_t ~ 0 + int_tild + Unemployed_t + Armed.Forces_t + Population_t + Employed_t

Observations	Residual Std. Error	R^2	Adjusted R^2
15	5.106	0.9998	0.9997

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

Table 15: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
int_tild	1	1,476,031	1,476,031	56,620	4.226e-20
Unemployed_t	1	14,793	14,793	567.5	3.861e-10
Armed.Forces_t	1	17,261	17,261	662.1	1.806e-10
Population_t	1	42,993	42,993	1,649	1.962e-12
Employed_t	1	310.4	310.4	11.91	0.00622
Residuals	10	260.7	26.07	NA	NA

```
pander(white.test(mod2),big.mark=",") #-- white test
```

Test.statistic	P.value
2.28	0.3198

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

Table 17: Durbin-Watson test: `mod2`

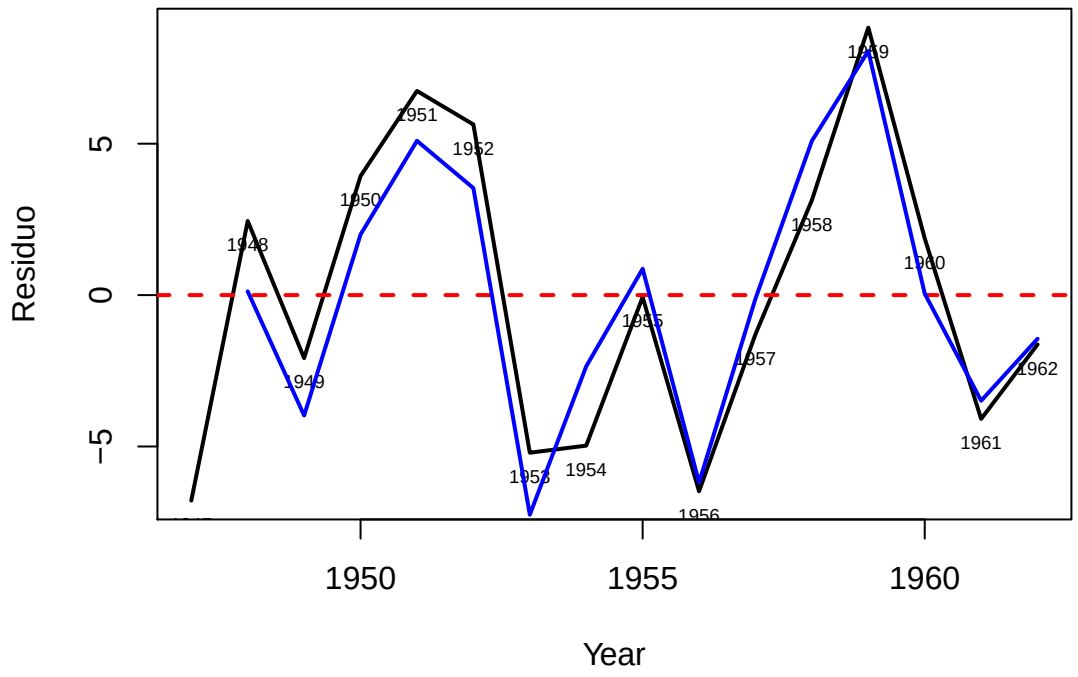
Test statistic	P value	Alternative hypothesis
1.607	0.04634 *	true autocorrelation is greater than 0

Il modello ha sempre un ottimo fitting, "Population" e "Employed" sono significative, il test di White certifica ancora la omoschedasticità e il test di Durbin Watson si muove quindi verso la regione di accettazione della non correlazione ma non è soddisfacente. Infatti sé visto che esistono autocorrelazioni di ordine superiore.

La rappresentazione grafica mostra come non sono state eliminate autocorrelazioni di ordine superiore (il secondo modello AR(1) è evidenziato in blu).

```r
#-- R CODE
plot(d$year,resid(mod1),pch=19,xlab="Year",ylab="Residuo",type="l",col=1,lwd=2)
text(d$year,resid(mod1),d$year,pos=1,cex=.6)
abline(h=0,lwd=2,lty=2,col=2)

lines(d$year[-1],resid(mod2),pch=19,xlab="Year",ylab="Residuo",type="l",col=4,lwd=2)
```



Si prova quindi ad utilizzare la funzione ARIMA (con OLS è assolutamente simile) che tenga conto di tali autocorrelazioni di ordine superiore:

```r
#-- R CODE
mod4 <- arima(d1$GNP, order=c(1,0,0), xreg = d1[,c("Unemployed","Armed.Forces","Population","Employed")]
mod4
```

```
##
## Call:
## arima(x = d1$GNP, order = c(1, 0, 0), xreg = d1[, c("Unemployed", "Armed.Forces",
##     "Population", "Employed")], method = "ML")
##
## Coefficients:
##         ar1   intercept  Unemployed  Armed.Forces  Population  Employed
##       0.283  -1347.5004     -0.0092         0.080      9.5132    9.1855
## s.e.  0.266     31.9761      0.0355         0.027      1.3790    2.2150
##
```

```
## sigma^2 estimated as 20.67:  log likelihood = -46.97,  aic = 107.95
```

```
coeftest(mod4)
```

```
##
## z test of coefficients:
##
##                 Estimate  Std. Error  z value  Pr(>|z|)
## ar1           2.8298e-01  2.6597e-01   1.0639  0.287352
## intercept    -1.3475e+03  3.1976e+01 -42.1409 < 2.2e-16 ***
## Unemployed   -9.1646e-03  3.5505e-02  -0.2581  0.796314
## Armed.Forces  8.0043e-02  2.6951e-02   2.9699  0.002979 **
## Population    9.5132e+00  1.3790e+00   6.8988 5.246e-12 ***
## Employed      9.1855e+00  2.2150e+00   4.1469 3.370e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
durbinWatsonTest(as.numeric(mod4$residuals))
```

```
## [1] 1.750303
```