Es3

Niccolò Puccinelli

2022-05-10

```
data <- read.csv("Hartnagel.txt", sep=",")
View(data)
```

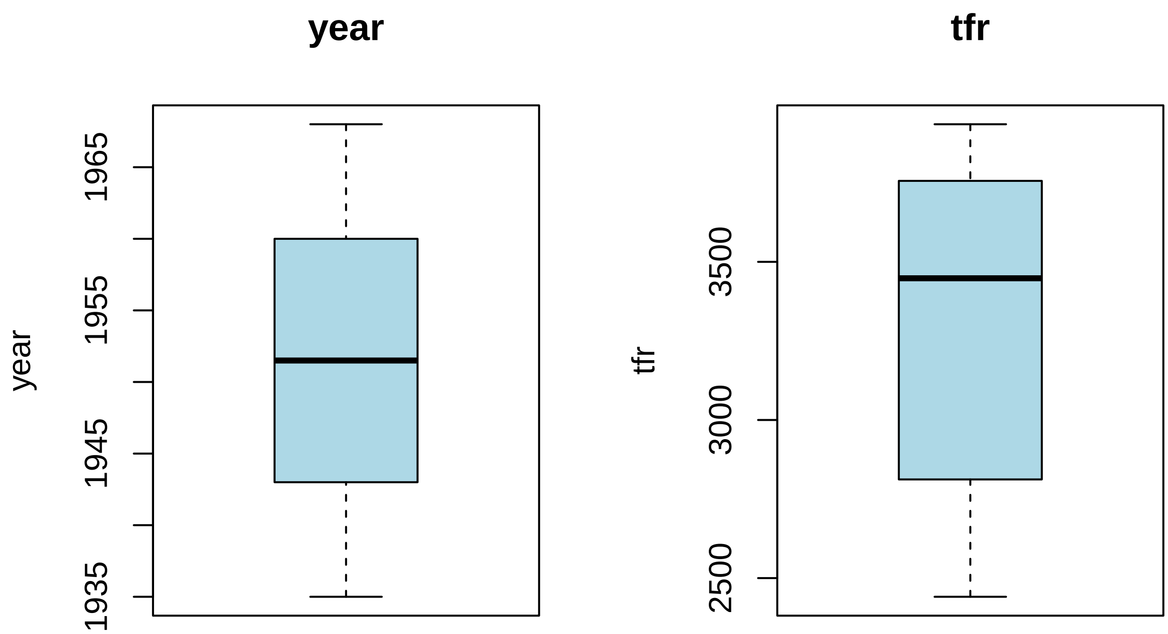
Il dataset presenta variabili numeriche. Occupiamoci anzitutto delle statistiche descrittive, dopo aver ordinato i dati.

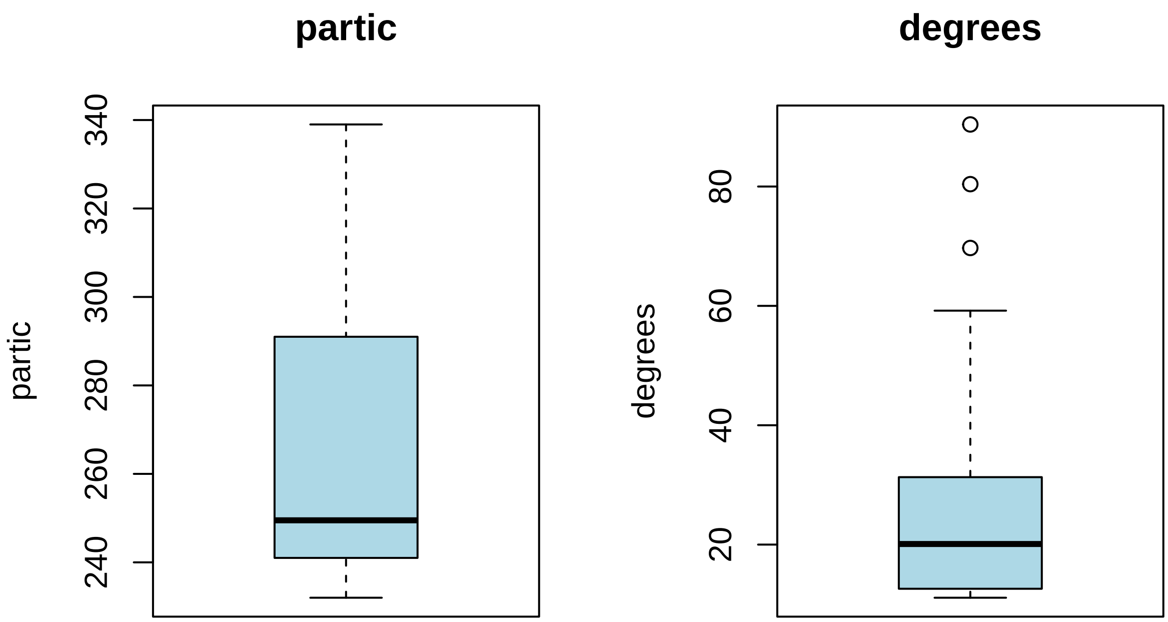
Statistiche descrittive

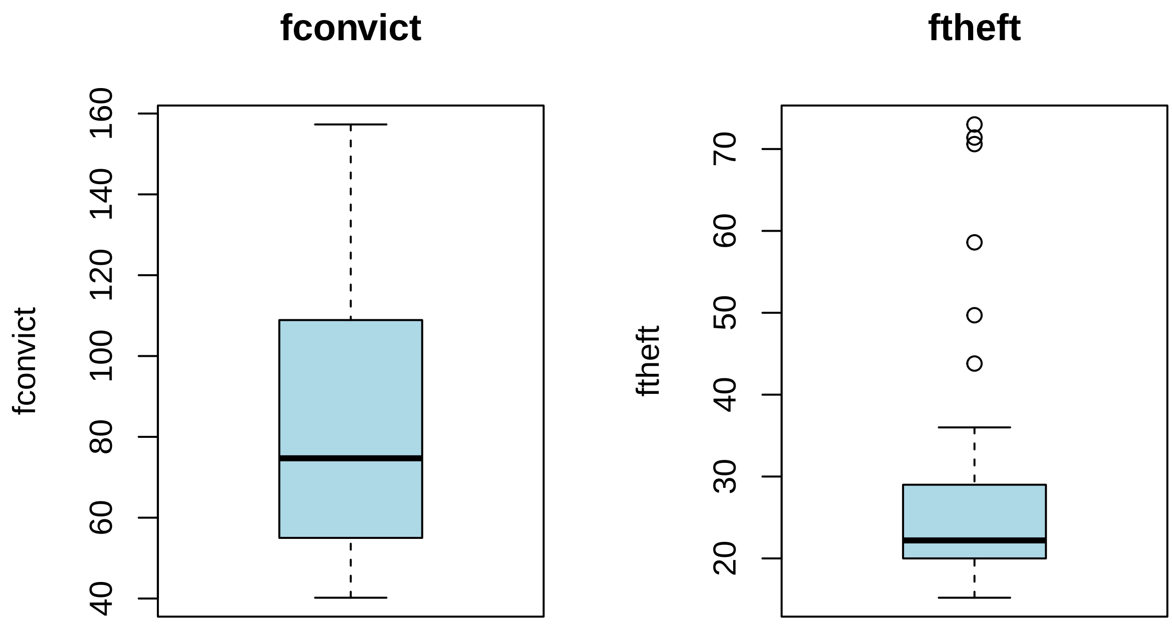
```
var = c("year", "tfr", "partic", "degrees", "fconvict", "ftheft", "mconvict", "mtheft")
summary(data[, var])
```

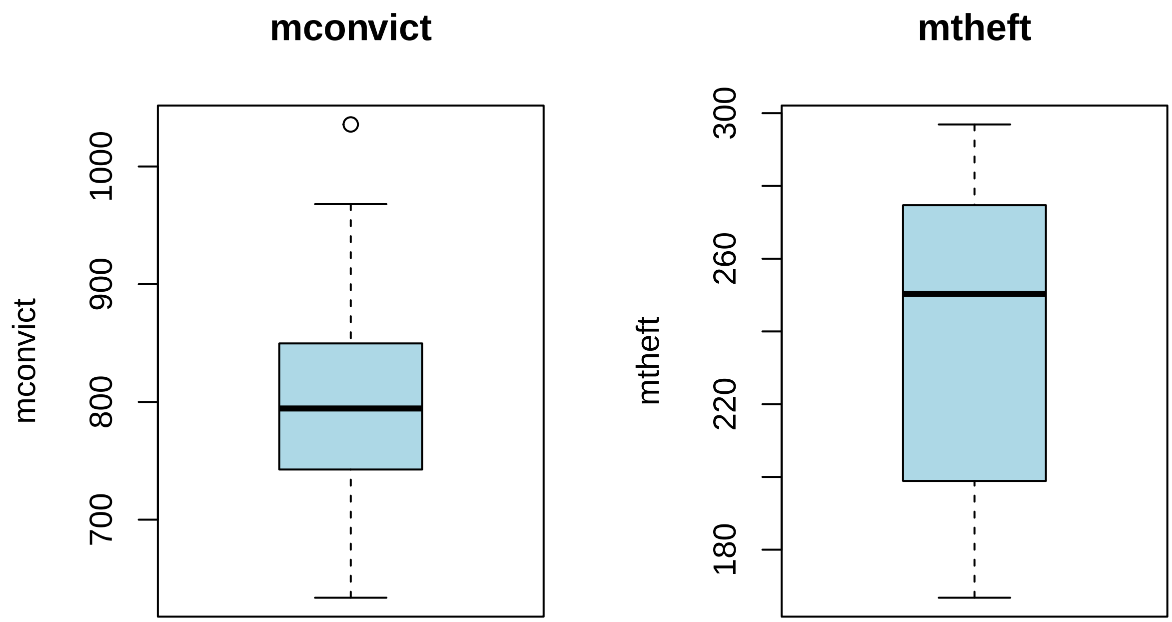
```
##       year           tfr            partic          degrees          fconvict     
##  Min.   :1935   Min.   :2441   Min.   :232.0   Min.   :11.10   Min.   : 40.20  
##  1st Qu.:1943   1st Qu.:2817   1st Qu.:241.2   1st Qu.:12.75   1st Qu.: 55.00  
##  Median :1952   Median :3448   Median :249.5   Median :20.10   Median : 74.70  
##  Mean   :1952   Mean   :3298   Mean   :269.4   Mean   :27.02   Mean   : 84.23  
##  3rd Qu.:1960   3rd Qu.:3747   3rd Qu.:290.8   3rd Qu.:30.60   3rd Qu.:107.00  
##  Max.   :1968   Max.   :3935   Max.   :339.0   Max.   :90.40   Max.   :157.30  
##      ftheft         mconvict          mtheft     
##  Min.   :15.20   Min.   : 633.7   Min.   :166.8  
##  1st Qu.:20.10   1st Qu.: 747.2   1st Qu.:199.0  
##  Median :22.20   Median : 794.5   Median :250.3  
##  Mean   :29.13   Mean   : 803.7   Mean   :237.2  
##  3rd Qu.:28.88   3rd Qu.: 847.8   3rd Qu.:274.1  
##  Max.   :73.00   Max.   :1035.7   Max.   :296.9  
```

```
data<-data[order(data$year),]
# Box-plot
par(mfrow=c(1,2))
for(i in var){
  boxplot(data[,i],main=i,col="lightblue",ylab=i)
}
```


year
tfr
year
tfr
1935
1945
1955
1965
2500
3000
3500


partic
partic
340
320
300
280
260
240
degrees
degrees
80
60
40
20

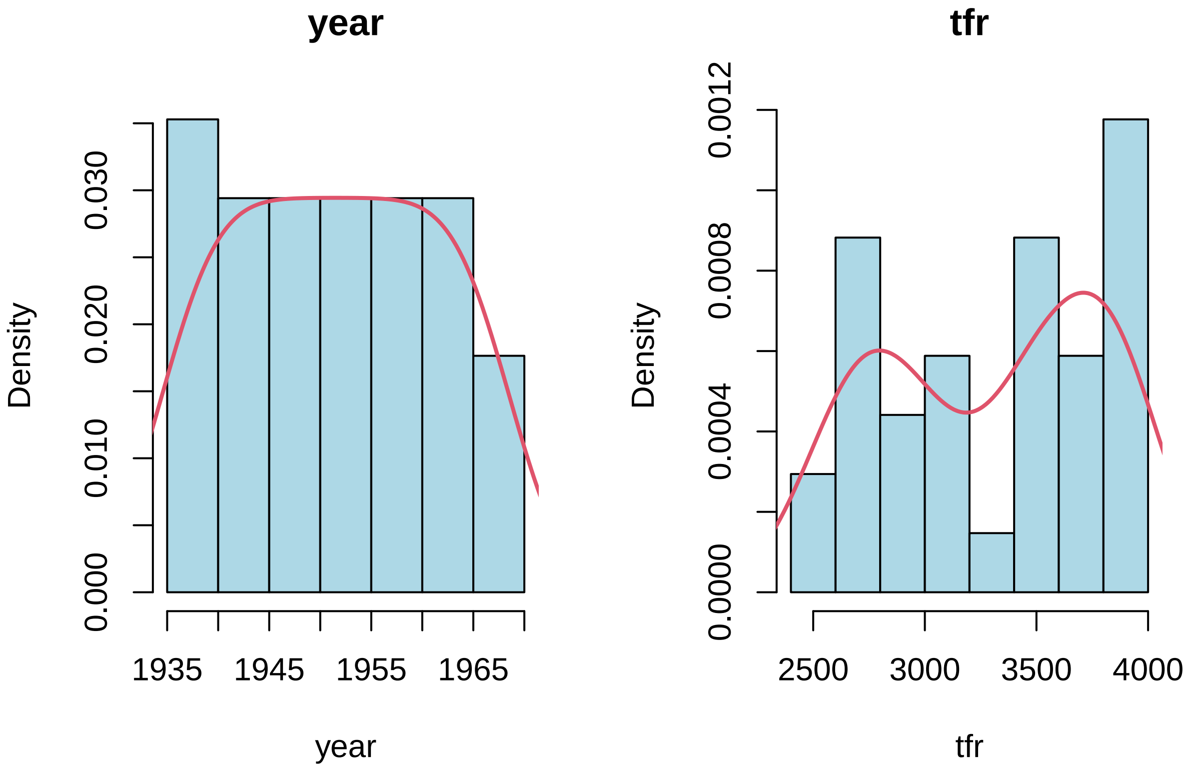

fconvict
fconvict
40
60
80
100
120
140
160
ftheft
ftheft
20
30
40
50
60
70

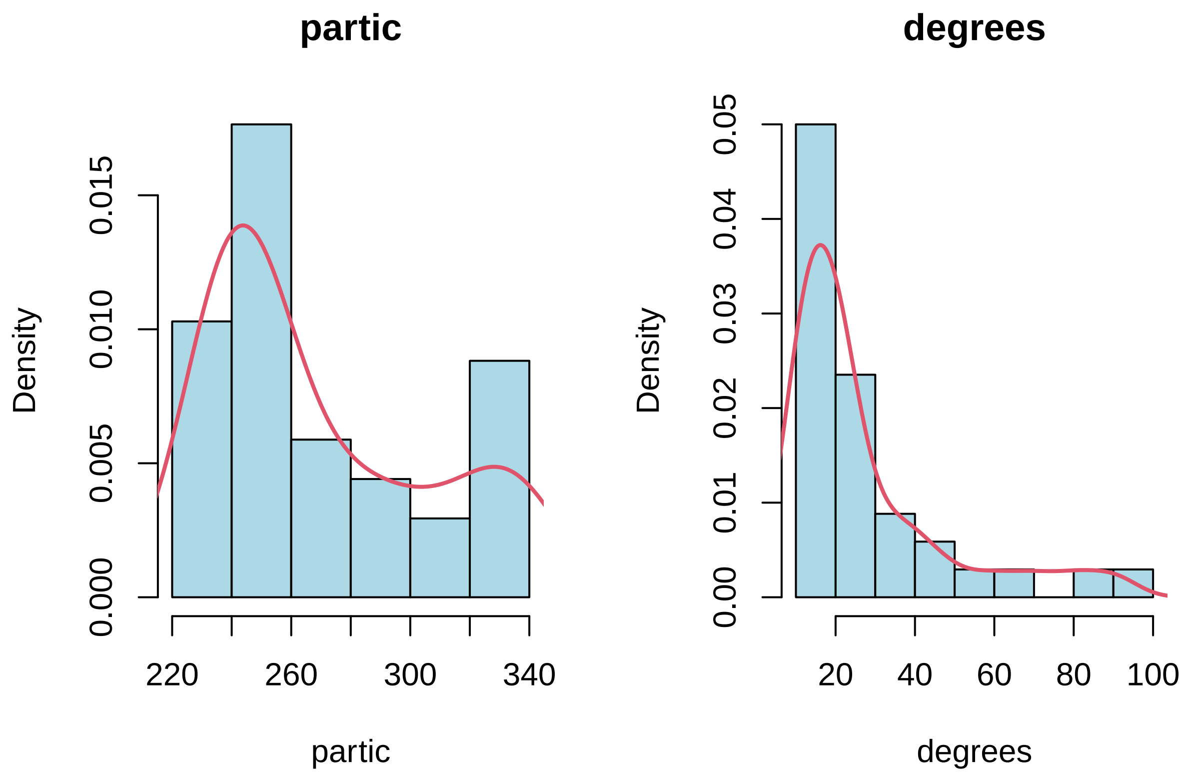


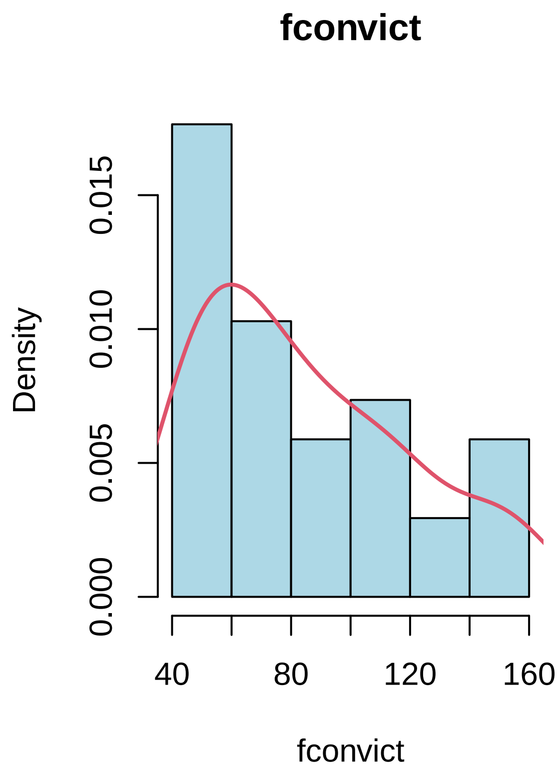
Dai box-plot notiamo che per quasi tutte le variabili la media si discosta dalla mediana, indicando una possibile non-normalità. Inoltre possiamo già vedere i primi outlier, presenti nelle variabili degrees, ftheft e fconvict.

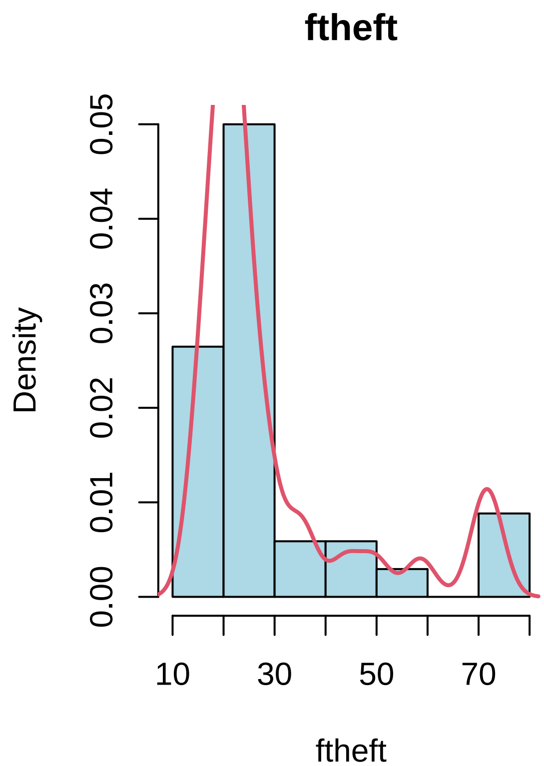
Andiamo a vedere simmetria e possibile non-normalità anche con degli istogrammi.

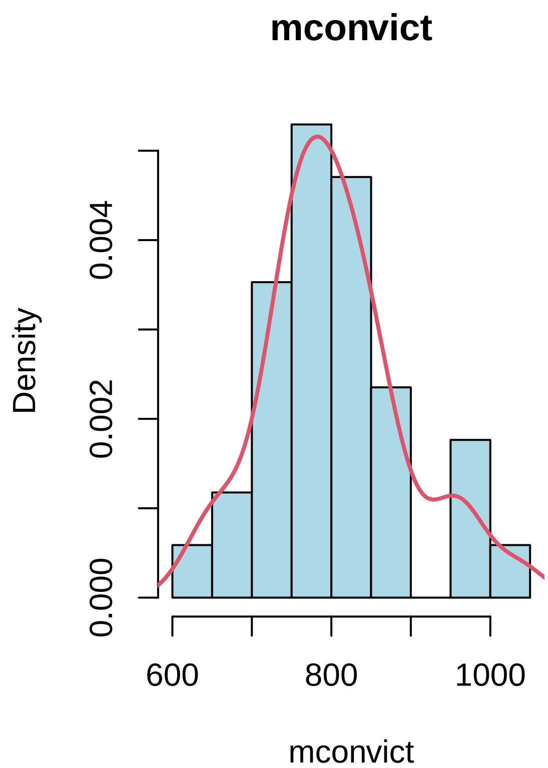
```
# Istogrammi
par(mfrow=c(1,2))
for(i in var){
  hist(data[,i],main=i,col="lightblue",xlab=i,freq=F,prob=TRUE)
  lines(density(data[,i]), col=2, lwd=2)
}
```

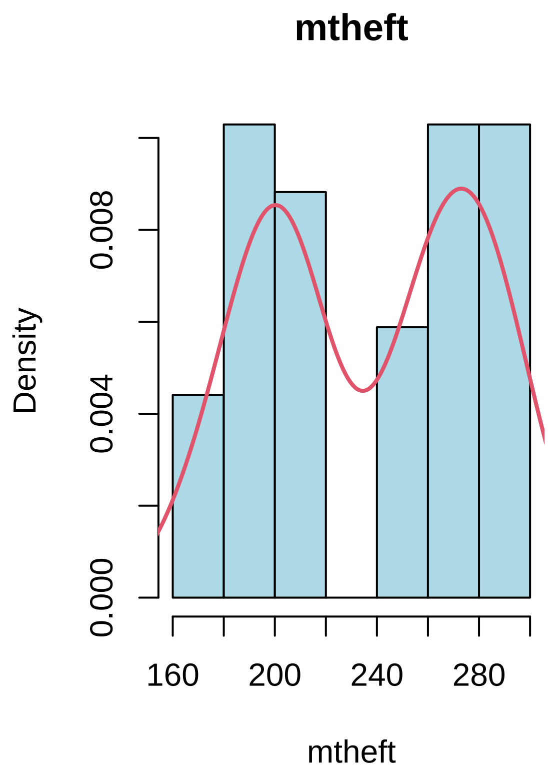

year
Density
0.000
0.010
0.020
0.030
1935
1945
1955
1965
year
tfr
Density
0.0000
0.0004
0.0008
0.0012
2500
3000
3500
4000
tfr


partic
Density
0.000
0.005
0.010
0.015
220
260
300
340
partic
degrees
Density
0.00
0.01
0.02
0.03
0.04
0.05
20
40
60
80
100
degrees


fconvict
Density
0.000
0.005
0.010
0.015
40
80
120
160
fconvict


ftheft
Density
0.00
0.01
0.02
0.03
0.04
0.05
10
30
50
70
ftheft

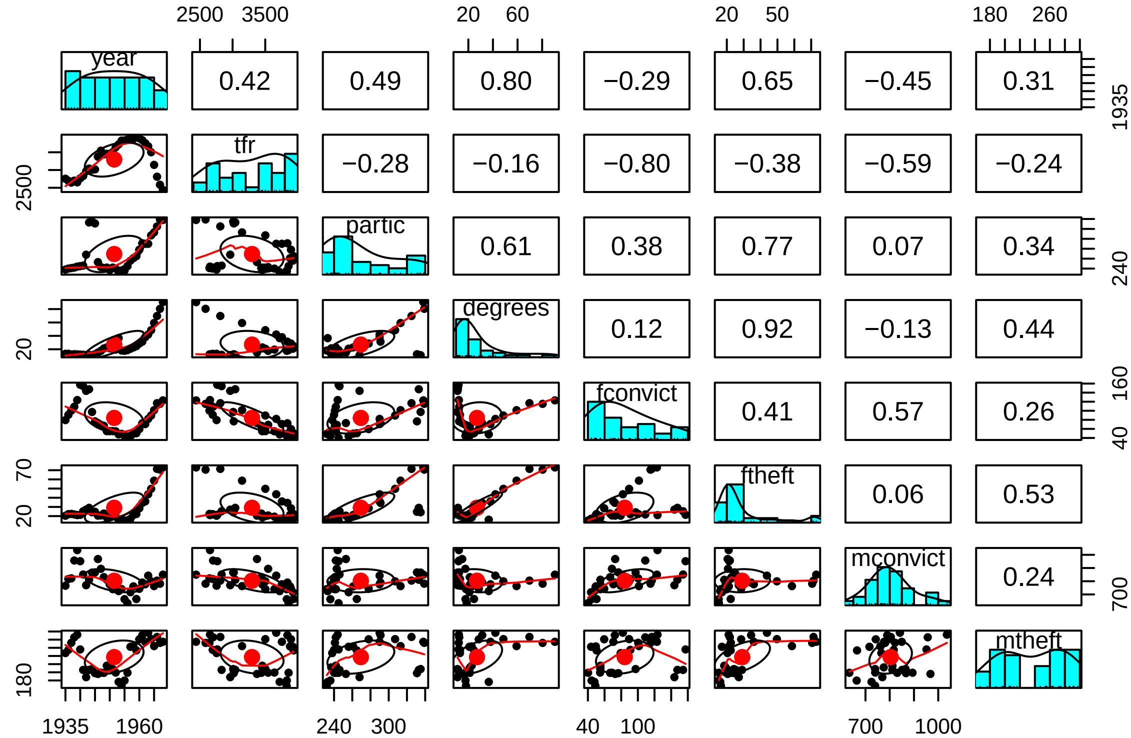




Le variabili partic, degrees, fconvict e ftheft presentano un'evidente asimmetria positiva. Inoltre diverse distribuzioni sono poli-modali (e.g. tfr e mtheft sono bimodali).

Andiamo a studiare le correlazioni.

```
# Studio delle correlazioni
pairs.panels(data[,var])
```




```
cor(data[,var])
```

```
##                year        tfr      partic    degrees   fconvict      ftheft
## year      1.0000000  0.4247077  0.48641461  0.8021857 -0.2895718  0.64662306
## tfr       0.4247077  1.0000000 -0.27662346 -0.1561002 -0.7997838 -0.38259325
## partic    0.4864146 -0.2766235  1.00000000  0.6078014  0.3818579  0.76559363
## degrees   0.8021857 -0.1561002  0.60780137  1.0000000  0.1180351  0.91782180
## fconvict -0.2895718 -0.7997838  0.38185791  0.1180351  1.0000000  0.41208835
## ftheft    0.6466231 -0.3825933  0.76559363  0.9178218  0.4120884  1.00000000
## mconvict -0.4499574 -0.5921274  0.07262595 -0.1327749  0.5698410  0.06224752
## mtheft    0.3134786 -0.2373750  0.33843500  0.4447531  0.2557102  0.53439068
##             mconvict     mtheft
## year     -0.44995741  0.3134786
## tfr      -0.59212741 -0.2373750
## partic    0.07262595  0.3384350
## degrees  -0.13277490  0.4447531
## fconvict  0.56984105  0.2557102
## ftheft    0.06224752  0.5343907
## mconvict  1.00000000  0.2409680
## mtheft    0.24096796  1.0000000
```

Le variabili maggiormente correlate sono year con degrees (0.8), tfr con fconvict (0.8), partic con ftheft (0.77) e degrees con ftheft (0.92).

Consideriamo da ora in avanti esclusivamente le variabili di nostro interesse (i.e. ftheft, partic, degrees e mtheft).

Regressione di ftheft su partic, degrees e mtheft

Effetttuiamo la regressione e commentiamone i risultati.

```r
m1<-lm(ftheft~partic+degrees+mtheft,data)
summary(m1)
```

```
##
## Call:
## lm(formula = ftheft ~ partic + degrees + mtheft, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8242 -3.2892  0.5658  2.5798  9.3956
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.06091    8.11357  -4.445 0.000111 ***
## partic        0.14143    0.02849   4.965 2.57e-05 ***
## degrees       0.53861    0.05380  10.012 4.45e-11 ***
## mtheft        0.05282    0.02279   2.318 0.027467 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.754 on 30 degrees of freedom
## Multiple R-squared:  0.9244, Adjusted R-squared:  0.9168
## F-statistic: 122.2 on 3 and 30 DF,  p-value: < 2.2e-16
```

Tutte le variabili risultano significative. Il modello risulta significativo, poiché il p-value associato alla statistica F è molto basso. Infine l'R quadro risulta molto elevato e spiega circa il 92% della variabilità totale, mentre l'R quadro aggiustato è leggermente minore.

Andiamo a verificare eventuali episodi di multicollinearità.

Multicollinearità

Consideriamo gli appositi indici.

```r
# Se VIF >= 10 considero multicollinearità
vif(m1)
```

```
##   partic  degrees   mtheft
## 1.600527 1.766661 1.258119
```

```r
ols_vif_tol(m1)
```

```
##   Variables Tolerance      VIF
## 1    partic 0.6247941 1.600527
## 2   degrees 0.5660397 1.766661
## 3    mtheft 0.7948372 1.258119
```

```r
# Se Condition index >= 10 e la quota di varianza di ogni variabile associata ai
# valori elevati dell'indice è anch'essa elevata (>80%), considero multicollinearità
ols_eigen_cindex(m1)
```

```
##    Eigenvalue Condition Index    intercept       partic    degrees      mtheft
## 1 3.726335349         1.00000 0.0007004976 0.0007841084 0.01172856 0.001553380
## 2 0.251492599         3.84927 0.0051988108 0.0019275187 0.62156045 0.004717551
## 3 0.016435124        15.05756 0.0446910169 0.1830183185 0.00539348 0.881834576
## 4 0.005736927        25.48598 0.9494096747 0.8142700543 0.36131752 0.111894492
```

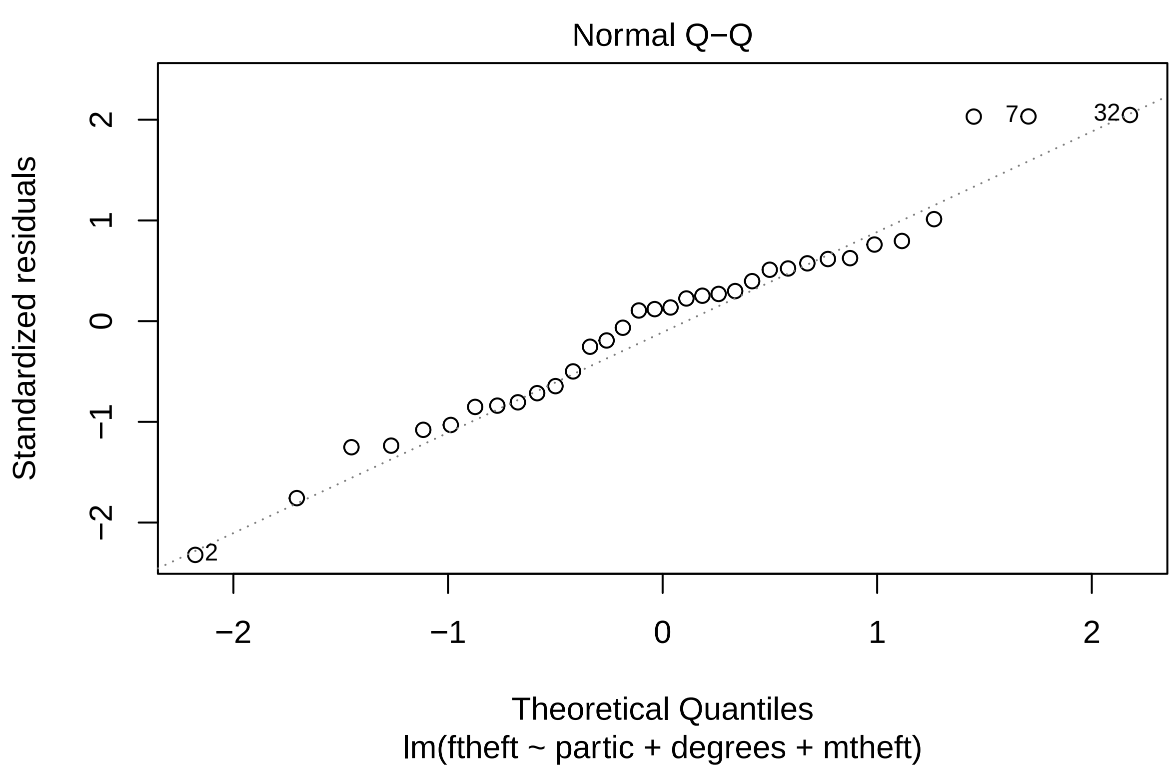
Sia l'indice di tolleranza, sia il VIF non mostrano episodi di evidente multicollinearità. Il VIF, in particolare, è sotto 10 per ogni variabile esplicativa. Per quanto riguarda il condition index invece, questo risulta maggiore di 10 per i due autovalori più piccoli. Ciò nonostante, manteniamo le variabili così come sono.

Poiché abbiamo a che fare con un dataset temporale, manteniamo l'ordine e lasciamo stare gli outlier.

Normalità

Adesso visualizziamo il Normal Q-Q Plot per verificare la normalità.

```r
plot(m1, which=2)
```



La distribuzione si discosta leggermente dall'andamento della normale nella parte centrale. Potrebbe esserci non normalità dei residui. Andiamo a vedere coi relativi test.

```r
ols_test_normality(m1)
```

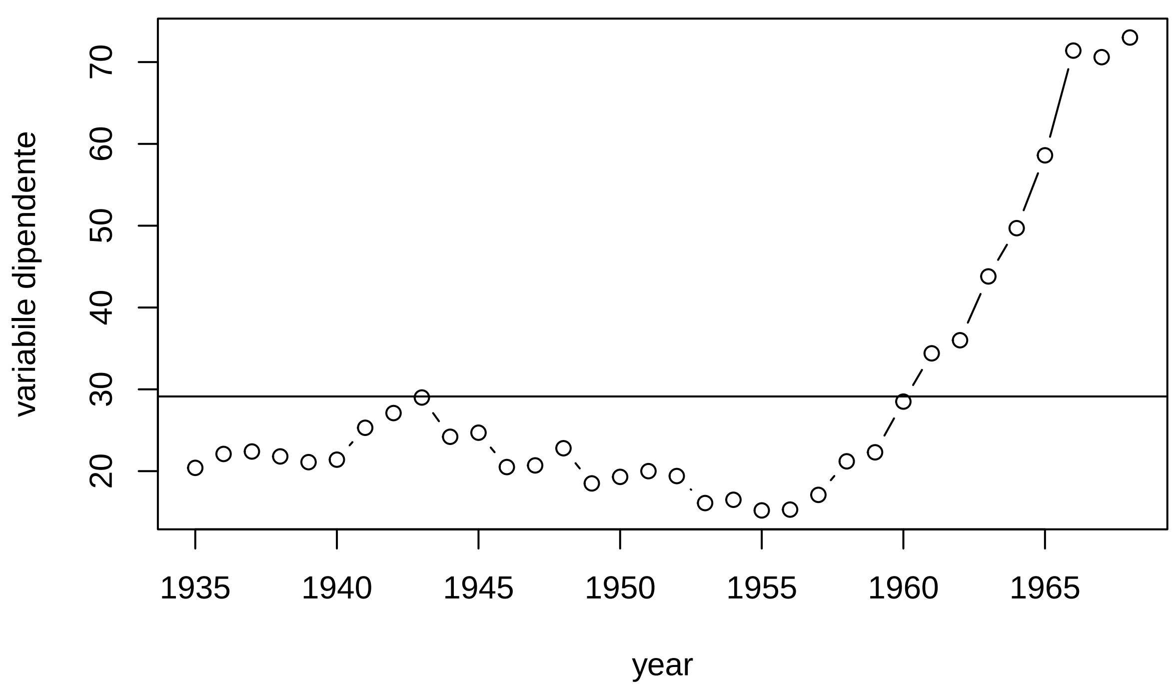
```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9712         0.4949
## Kolmogorov-Smirnov        0.1015         0.8399
## Cramer-von Mises          2.0822         0.0000
## Anderson-Darling          0.3657         0.4154
## -----------------------------------------------
```

Dai test risulta la normalità dei residui e l'ipotesi è pertanto verificata.

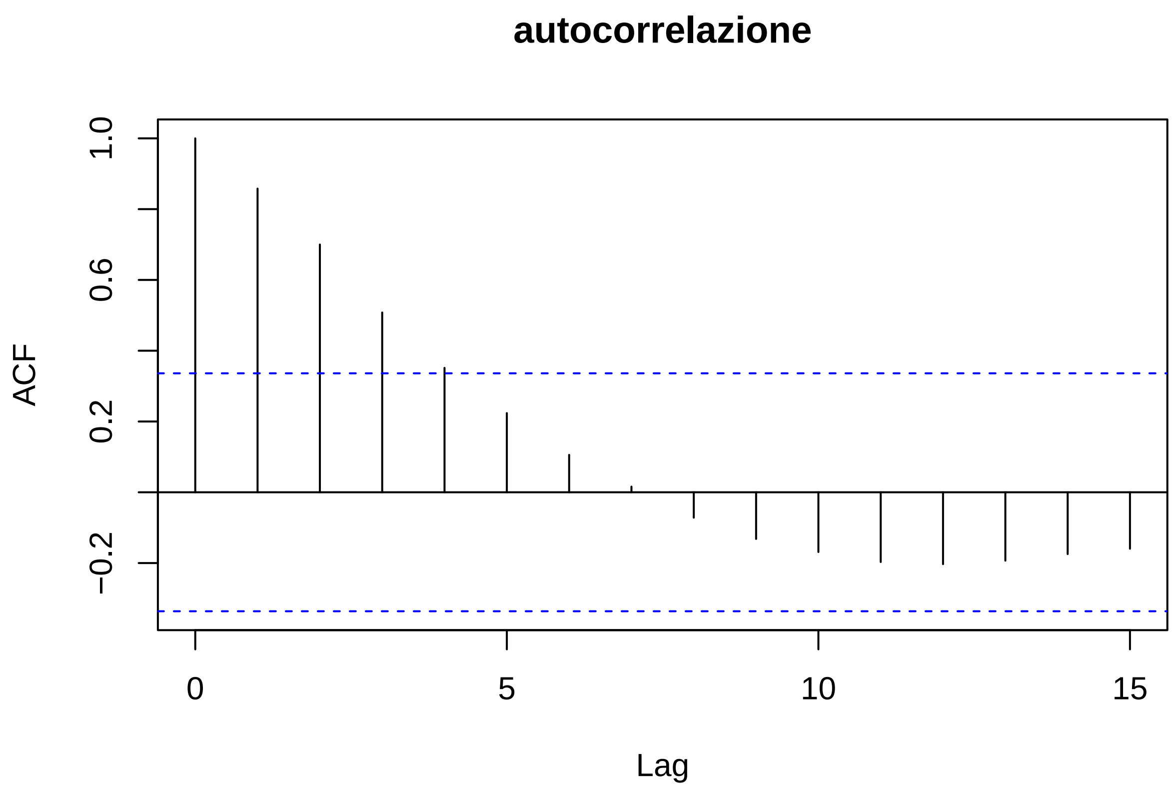
Autocorrelazione

Verifichiamo ora se c'è o meno autocorrelazione, tramite i seguenti grafici.

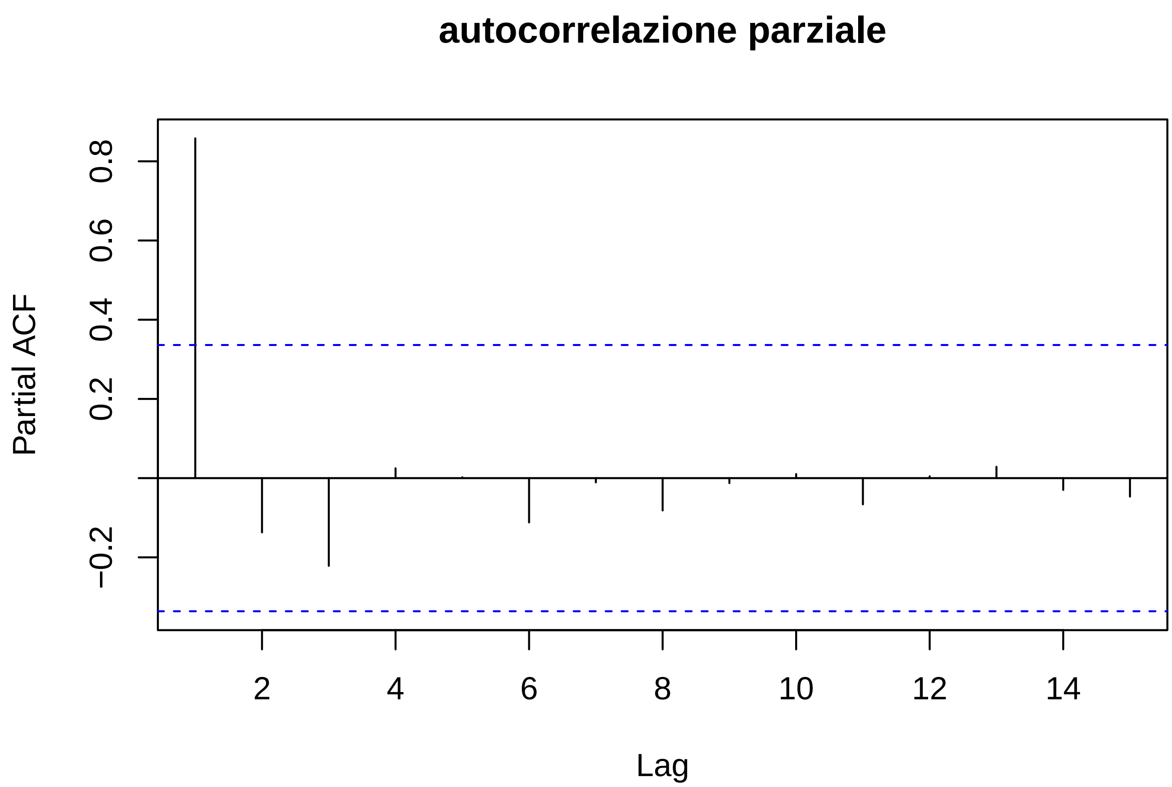
```r
plot(data$year, data$ftheft, ylab="variabile dipendente", xlab="year", type="b")
abline(h=mean(data$ftheft))
```




```r
# Autocorrelazione (correlogramma): grafico tra variabile dipendente e valori ritardati per tutti i rit
acf(data$ftheft, main="autocorrelazione")
```



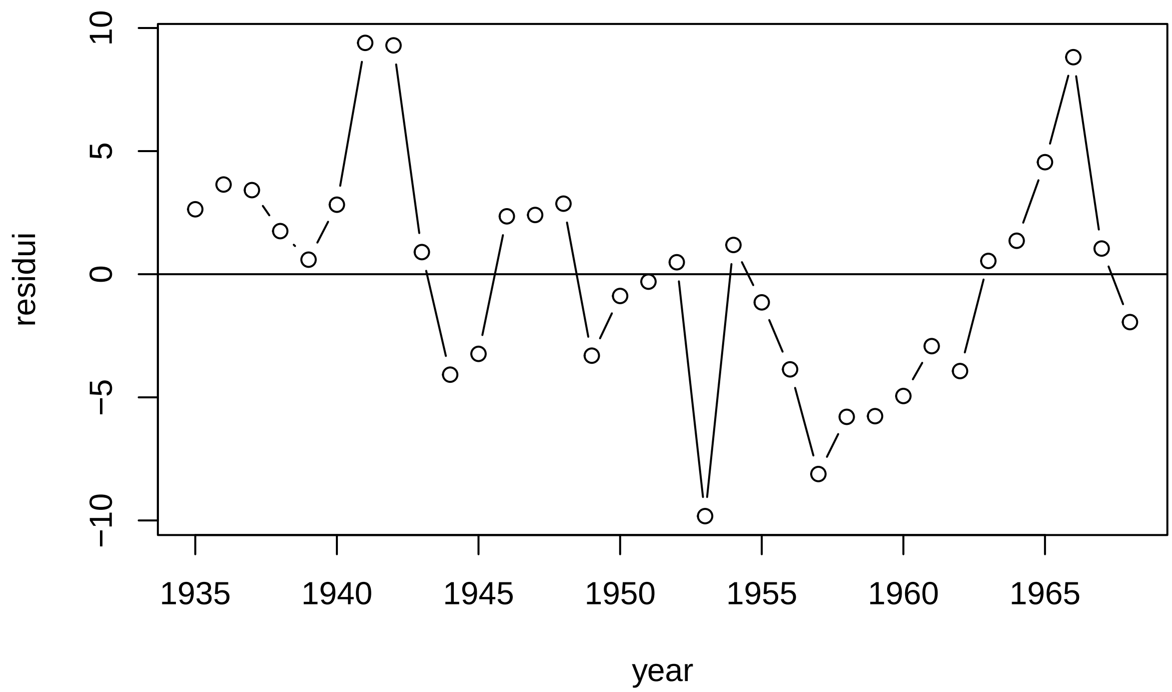

```r
# Autocorrelazione parziale: studia correlazioni di ordine superiore a parità
# delle correlazioni di ordine inferiore
pacf(data$ftheft, main="autocorrelazione parziale")
```



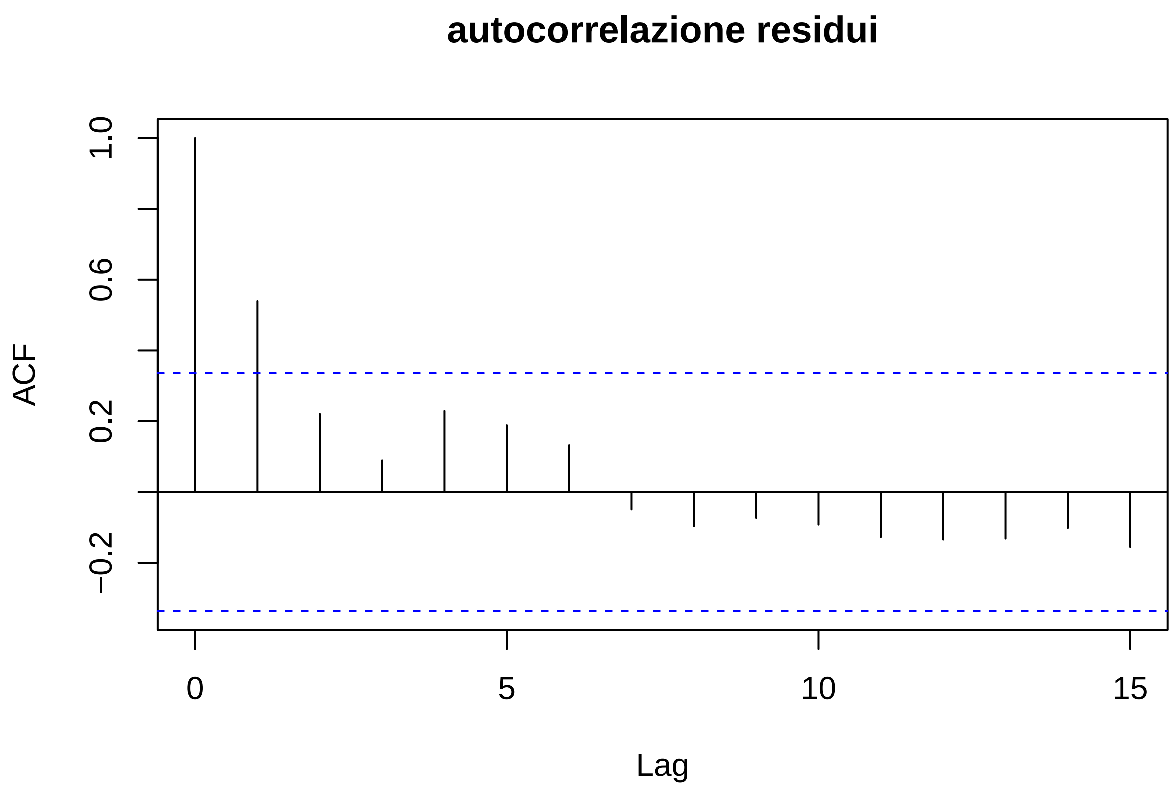
Dai grafici si evince che sussiste evidente autocorrelazione di ordine 1.

Guardiamo anche i residui nel tempo.

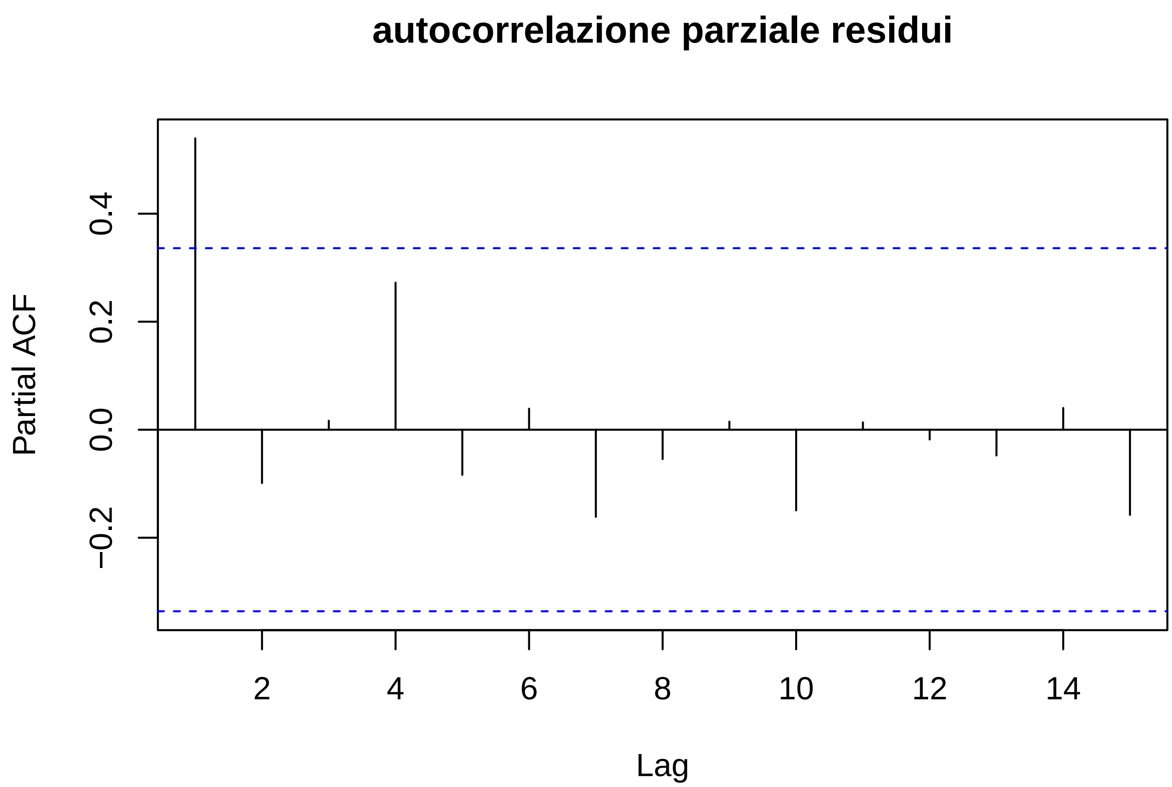
```
plot(data$year, m1$residuals, xlab="year", ylab="residui", type="b")
abline(h=0)
```




```
acf(m1$residuals, main="autocorrelazione residui")
```




```
pacf(m1$residuals, main="autocorrelazione parziale residui")
```



Notiamo che vi è autocorrelazione di ordine 1 anche per quanto riguarda i residui.

Confermiamo la nostra ipotesi con il test di Durbin-Watson.

```
durbinWatsonTest(m1, max.lag=5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.53952640     0.9051118   0.000
##    2      0.22105333     1.5208872   0.110
##    3      0.08943157     1.6523470   0.364
##    4      0.22930795     1.3375370   0.124
##    5      0.18871357     1.4154863   0.302
##  Alternative hypothesis: rho[lag] != 0
```

Il test e il relativo p-value ci conferma le nostre ipotesi di autocorrelazione di ordine 1. In particolare, abbiamo autocorrelazione positiva.

Proviamo dunque a risolvere il problema, mediante trasformazione delle variabili.

Anzitutto regrediamo i residui sui residui laggati e calcoliamo il coefficiente di correlazione di ordine 1.

```
data$res<-m1$residuals
data<-slide(data=data, Var='res', TimeVar = 'year', NewVar='res_lag')
```

```
##
## Lagging res by 1 time units.
```

```r
# Regrediamo residui su residui laggati
aux<-lm(res~res_lag,data)
summary(aux)
```

```
##
## Call:
## lm(formula = res ~ res_lag, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9764 -2.3988  0.0085  2.3227  7.9746
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1119     0.6795  -0.165  0.87024
## res_lag       0.5429     0.1503   3.612  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.903 on 31 degrees of freedom
##   (1 osservazione eliminata a causa di un valore mancante)
## Multiple R-squared:  0.2961, Adjusted R-squared:  0.2734
## F-statistic: 13.04 on 1 and 31 DF,  p-value: 0.001061
```

```r
# Si calcola il coefficiente di autocorrelazione di ordine 1
rho<-aux$coefficients[2]
```

Adesso occupiamoci della trasformazione vera e propria, sottraendo ad ogni variabile il prodotto tra il coefficiente di correlazione e la variabile laggata.

```r
# Variabili laggate di ordine 1
data<-slide(data=data, Var='ftheft', TimeVar = 'year', NewVar='ftheft_lag')
```

```
##
## Lagging ftheft by 1 time units.
```

```r
data<-slide(data=data, Var='partic', TimeVar = 'year', NewVar='partic_lag')
```

```
##
## Lagging partic by 1 time units.
```

```r
data<-slide(data=data, Var='degrees', TimeVar = 'year', NewVar='degrees_lag')
```

```
##
## Lagging degrees by 1 time units.
```

```r
data<-slide(data=data, Var='mtheft', TimeVar = 'year', NewVar='mtheft_lag')
```

```
##
## Lagging mtheft by 1 time units.
```

```r
# Si sottrae ad ogni variabile il prodotto tra il coefficiente di correlazione e la variabile laggata
data$ftheft_t<-data$ftheft-rho*data$ftheft_lag
data$partic_t<-data$partic-rho*data$partic_lag
data$degrees_t<-data$degrees-rho*data$degrees_lag
data$mtheft_t<-data$mtheft-rho*data$mtheft_lag
data$interc_t<-1-rho
```

Adesso ristimiamo il modello con le variabili trasformate e confrontiamolo col precedente.

```r
m2<-lm(ftheft_t~0+interc_t+partic_t+degrees_t+mtheft_t,data)
summary(m2)
```

```
##
## Call:
## lm(formula = ftheft_t ~ 0 + interc_t + partic_t + degrees_t +
##     mtheft_t, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6504  -2.0545  -0.1914   2.0164   8.0233
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## interc_t  -26.82350   10.27163  -2.611  0.01413 *
## partic_t    0.11814    0.03537   3.340  0.00231 **
## degrees_t   0.51676    0.06954   7.431 3.45e-08 ***
## mtheft_t    0.04257    0.03074   1.385  0.17672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.944 on 29 degrees of freedom
##   (1 osservazione eliminata a causa di un valore mancante)
## Multiple R-squared:  0.9517, Adjusted R-squared:  0.945
## F-statistic: 142.8 on 4 and 29 DF,  p-value: < 2.2e-16
```

```r
summary(m1)
```

```
##
## Call:
## lm(formula = ftheft ~ partic + degrees + mtheft, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8242 -3.2892  0.5658  2.5798  9.3956
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.06091    8.11357  -4.445 0.000111 ***
## partic        0.14143    0.02849   4.965 2.57e-05 ***
## degrees       0.53861    0.05380  10.012 4.45e-11 ***
## mtheft        0.05282    0.02279   2.318 0.027467 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.754 on 30 degrees of freedom
## Multiple R-squared:  0.9244, Adjusted R-squared:  0.9168
## F-statistic: 122.2 on 3 and 30 DF,  p-value: < 2.2e-16
```

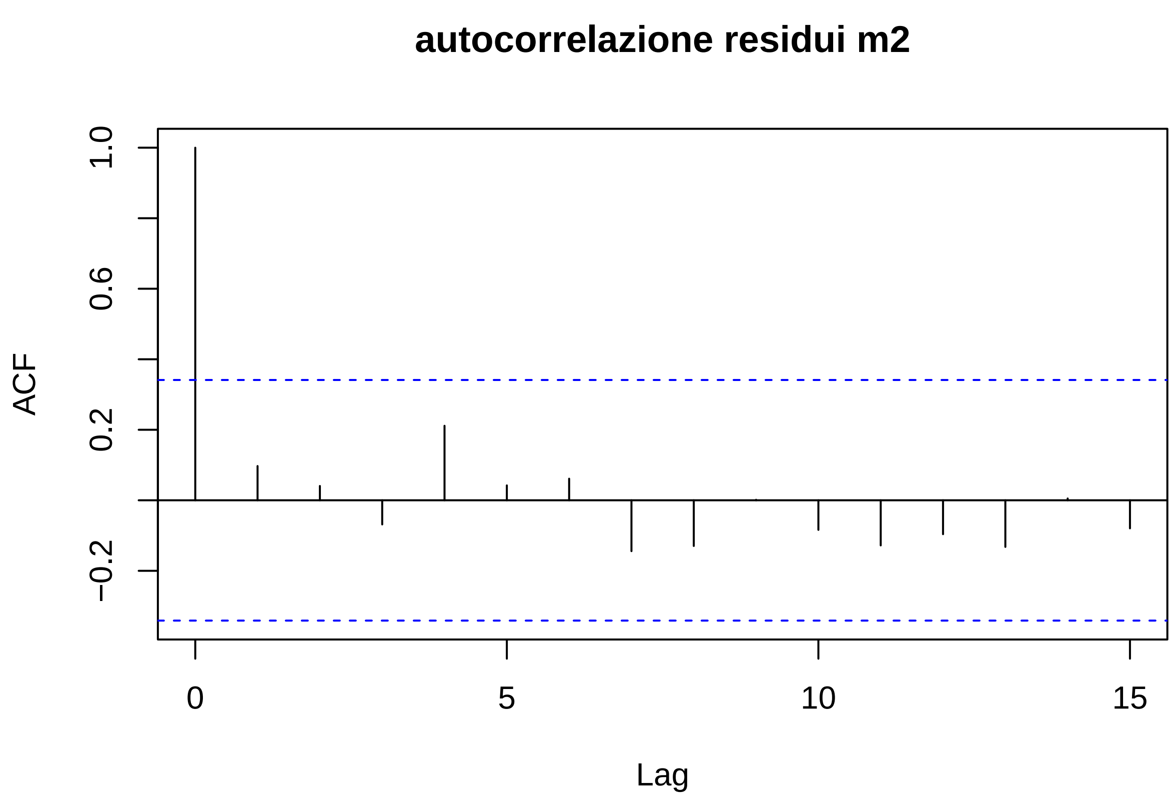
Il modello con le variabili trasformate presenta gli indici R quadro e R quadro aggiustato leggermente più alti del precedente. Tuttavia, nel nuovo modello mtheft non risulta più significativo.

Comunque, a noi interessa vedere se l'autocorrelazione si è risolta. Pertanto utilizziamo di nuovo i grafici e il test di Durbin-Watson.

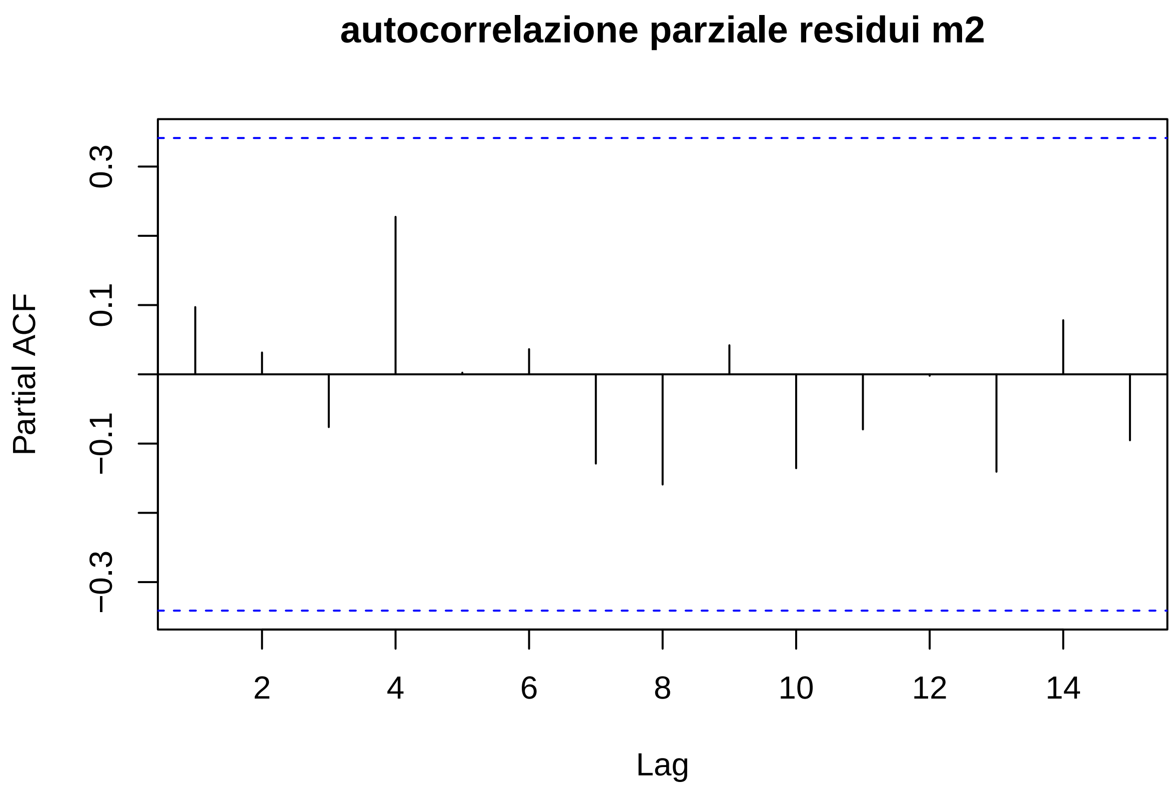
```r
# Durbin-Watson
durbinWatsonTest(m2, max.lag = 5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.09711383      1.795933   0.276
##    2      0.04064272      1.895891   0.738
##    3     -0.06849701      1.971350   0.860
##    4      0.21131119      1.358931   0.174
##    5      0.04234949      1.677867   0.848
##  Alternative hypothesis: rho[lag] != 0
```

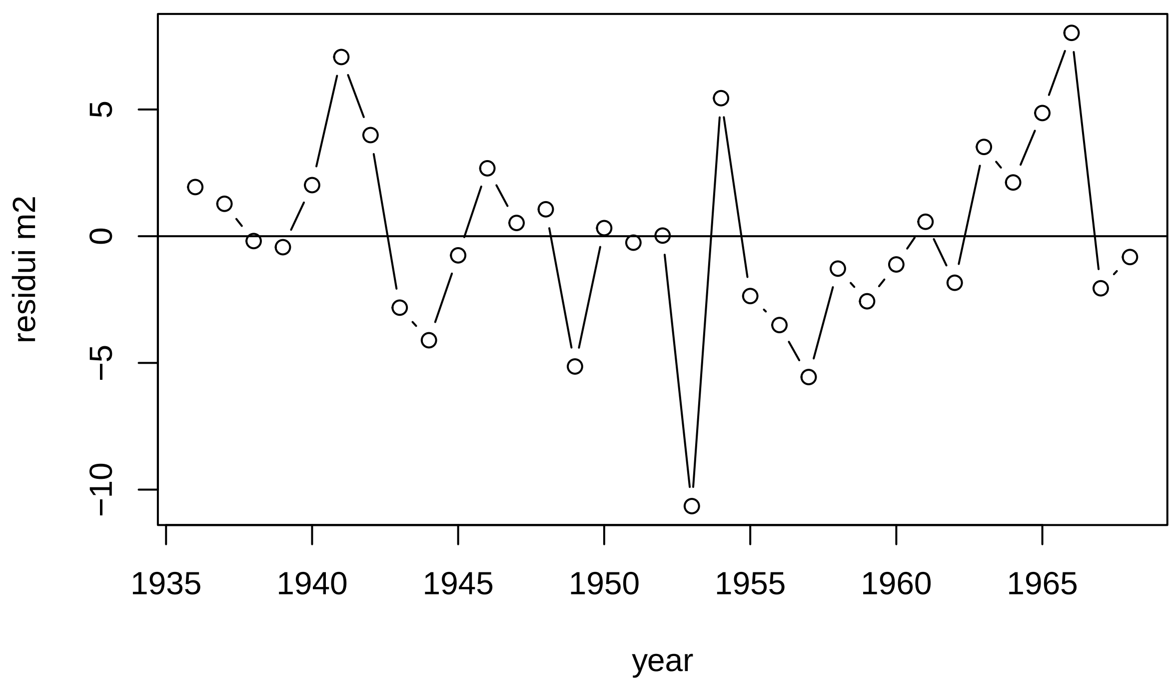
```r
# Correlogrammi
acf(m2$residuals, main="autocorrelazione residui m2")
```



```r
pacf(m2$residuals, main="autocorrelazione parziale residui m2")
```




```r
# Residui nel tempo
plot(data$year[-1], m2$residuals, xlab="year", ylab="residui m2", type="b")
abline(h=0)
```



Come si evince dai grafici e dal test, l'autocorrelazione è risolta.