LSM

Bojana Milošević

10/28/2019 i 11/04/2019

Provera korektnosti modela

- nekorelisanost grešaka
- centriranost
- normalnost i jednaka raspodeljenost
- ispitivanje uticaja tačaka
- ispitivanje prisustva autlajera

Šta sve možemo da zaključimo na osnovu vizualnog prikaza?

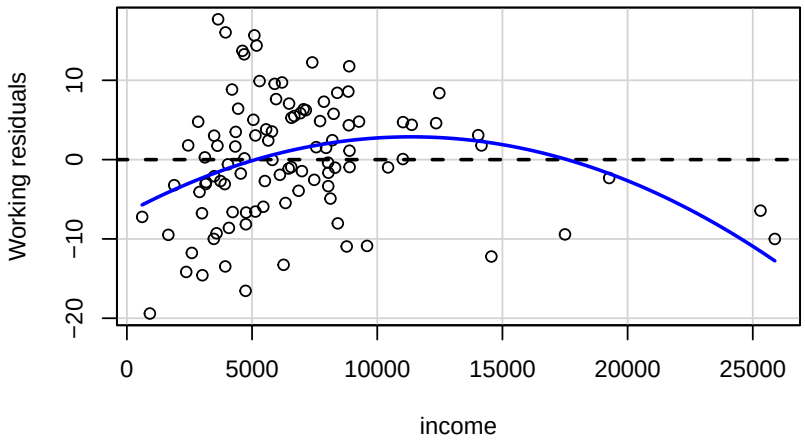
1. $e \sim x$

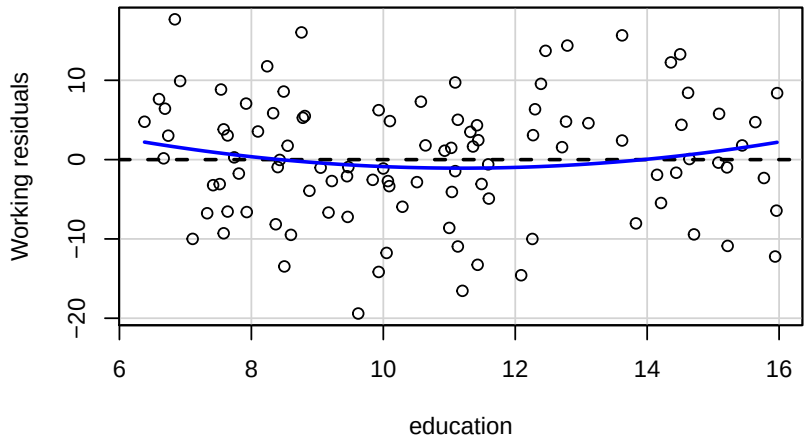
p različitih grafika

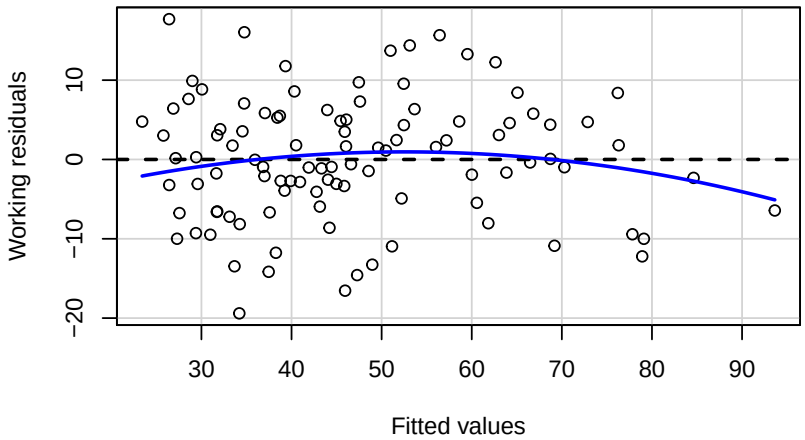
2. $e^2 \sim x$
3. Q-Q plot
4. $e \sim \hat{Y}$

Primer

```
residualPlots(modelPrestige,type='working')
```

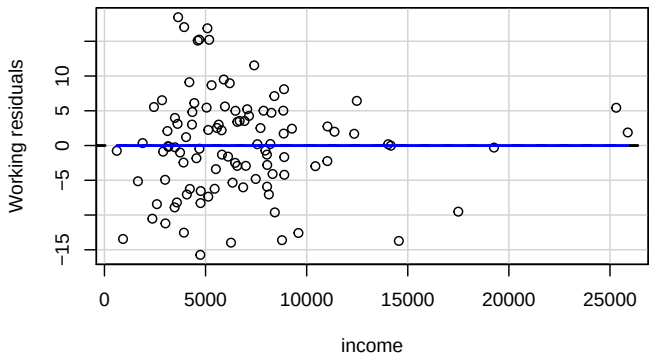


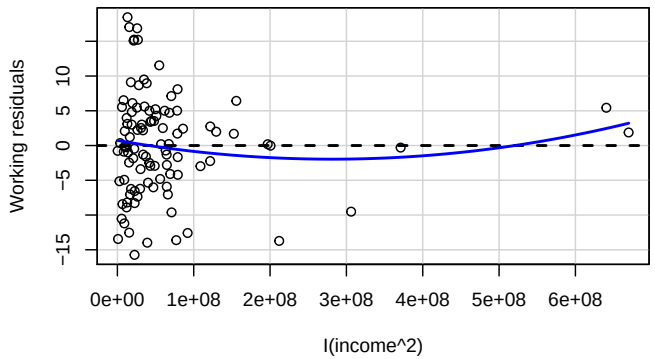


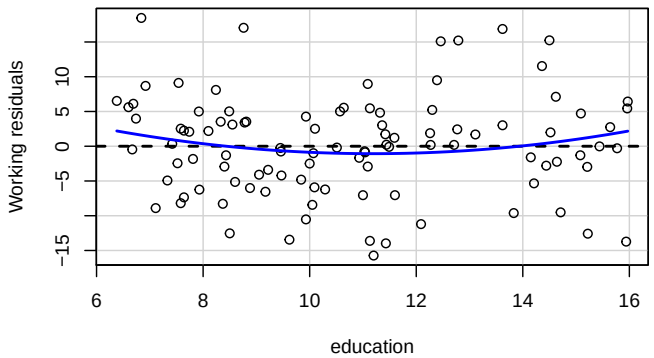


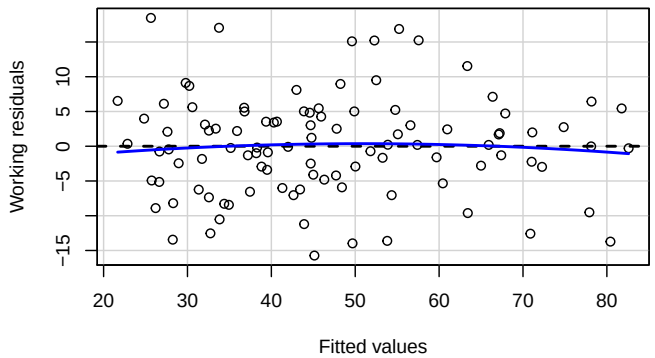
Primer

```
modelPrestige1=lm(prestige~income+I(income^2)+education,
                  data=Prestige)
residualPlots(modelPrestige1,type='working')
```



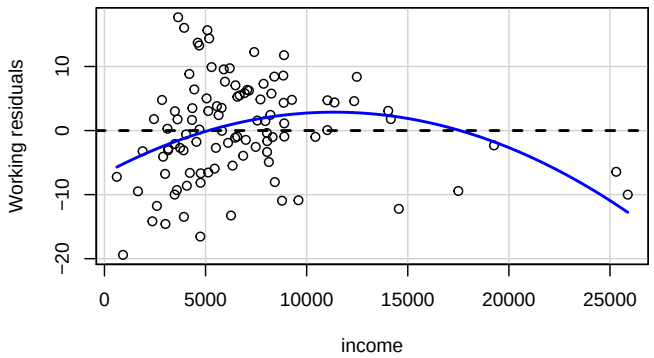


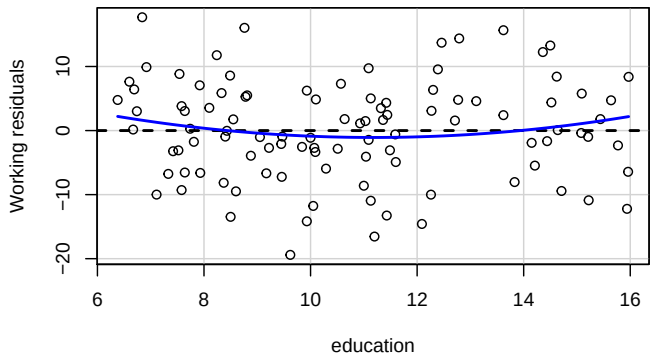


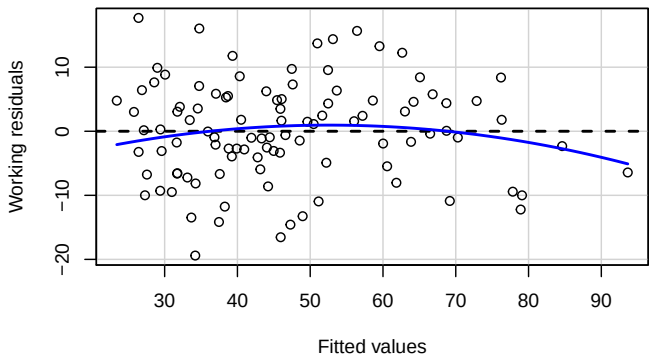


Primer

```
modelPrestige2=lm(prestige~(income^2)+education,
                  data=Prestige)
residualPlots(modelPrestige2,type='working')
```







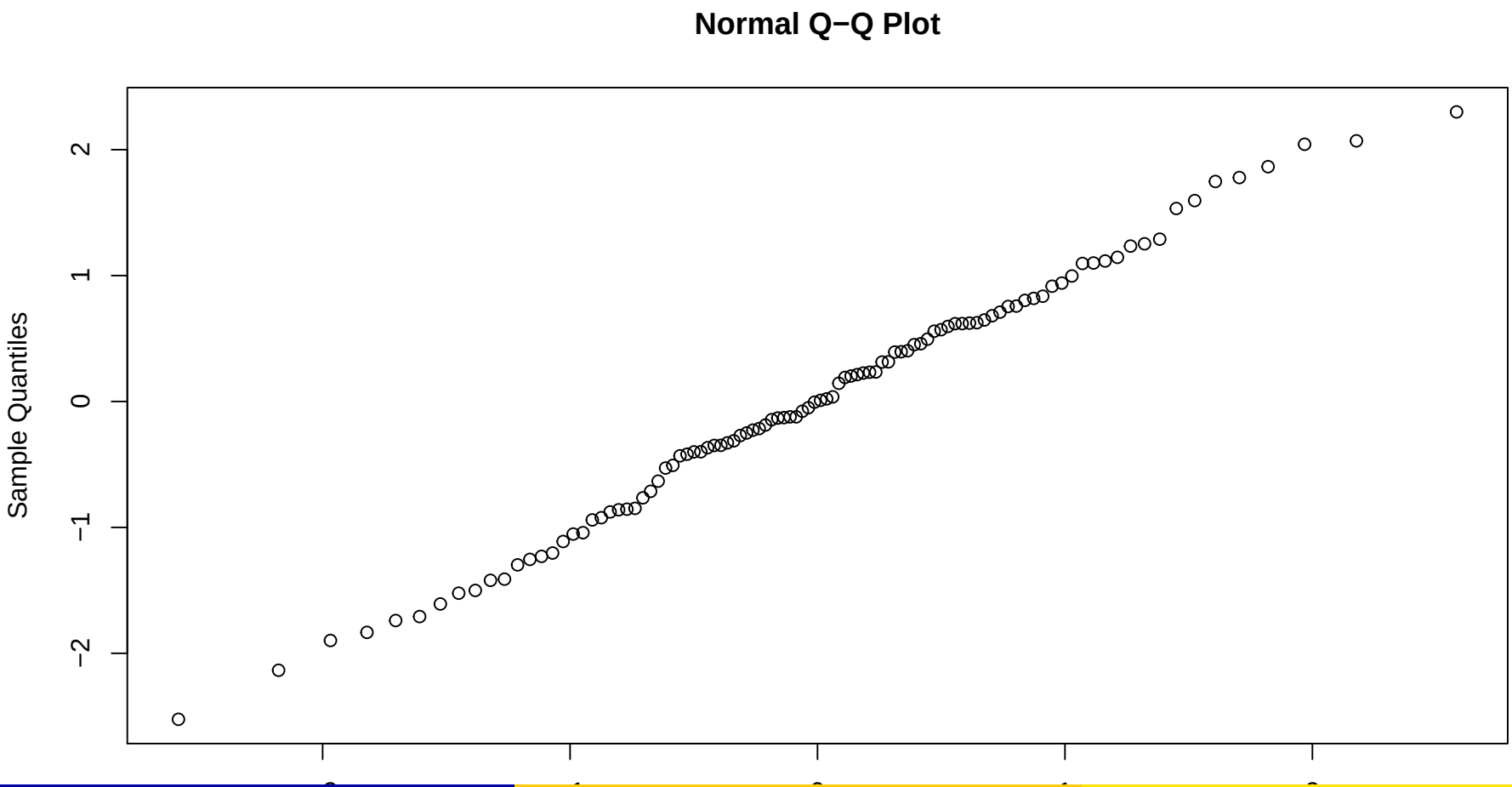
Standardizovani reziduali

$$e_i^s = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}},$$

Ako važi pretostavka modela onda $e_i^s \sim \frac{t_{n-p-1}\sqrt{n-p-1}}{t_{n-p-1}^2+n-p-2}$ i nalaze se u intervalu $[-\sqrt{n-p-1}, \sqrt{n-p-1}]$. Asimptotski imaju standardnu normalnu raspodelu. Ovi reziduali nisu pogodni za proveru nekorelisanosti sa prediktorom. Zašto?

Q-Q plot standardizovanih reziduala

```
rezidualiPrestigeS=rstandard(modelPrestige)
qqnorm(rezidualiPrestigeS)
```



Testovi normalnosti

KS test Test statistika je $KS = \sup_t |F_n(t) - F_0(t)|$. Kritichna oblast za testiranje je $\{KS > c\}$.

AD test Test statistika je $AD = n \int_{-\infty}^{\infty} \frac{(F_n(x)-F_0(x))^2}{F_0(x)(1-F_0(x))} dx$. Kritichna oblast za testiranje je $\{AD > c\}$.

CM test Test statistika je $CM = n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 dx$. Kritichna oblast za testiranje je $\{CM > c\}$.

SW test Test statistika je $W = \frac{\sum_{i=1}^{n} a_i X_{(i)}}{s}$ a_i očekivane vrednosti statistika poretka standardne normalne raspodele. Kritična oblast za testiranje je $\{W < c\}$.

Primer

```
rezidualiPrestigeS1=rstandard(modelPrestige1)
shapiro.test(rezidualiPrestigeS1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rezidualiPrestigeS1
## W = 0.98355, p-value = 0.2367
```

Primer

```
rezidualiPrestigeS2=rstandard(modelPrestige2)
shapiro.test(rezidualiPrestigeS2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rezidualiPrestigeS2
## W = 0.99354, p-value = 0.9129
```

Provera homoskedastičnosti

- uspostavljanje linerane veze izmedju kvadrata ocenjenih reziduala e_i^2 i prediktora
- H_0 svi koeficijenti uz prediktore jednaki nula
- Test statistika je $T = nR^2$ gde je R^2 koeficijent determinacije za taj pomoćni model.
- Ako važi nulta hipoteza onda T ima približno χ_p^2
- Glavna mana ovog testa je što se ova raspodela menja ukoliko reziduali nisu normalno raspodeljeni. Zato je Kroenker predložio modifikaciju ovog testa za koju je pokazao da je robusna na raspodelu reziduala.

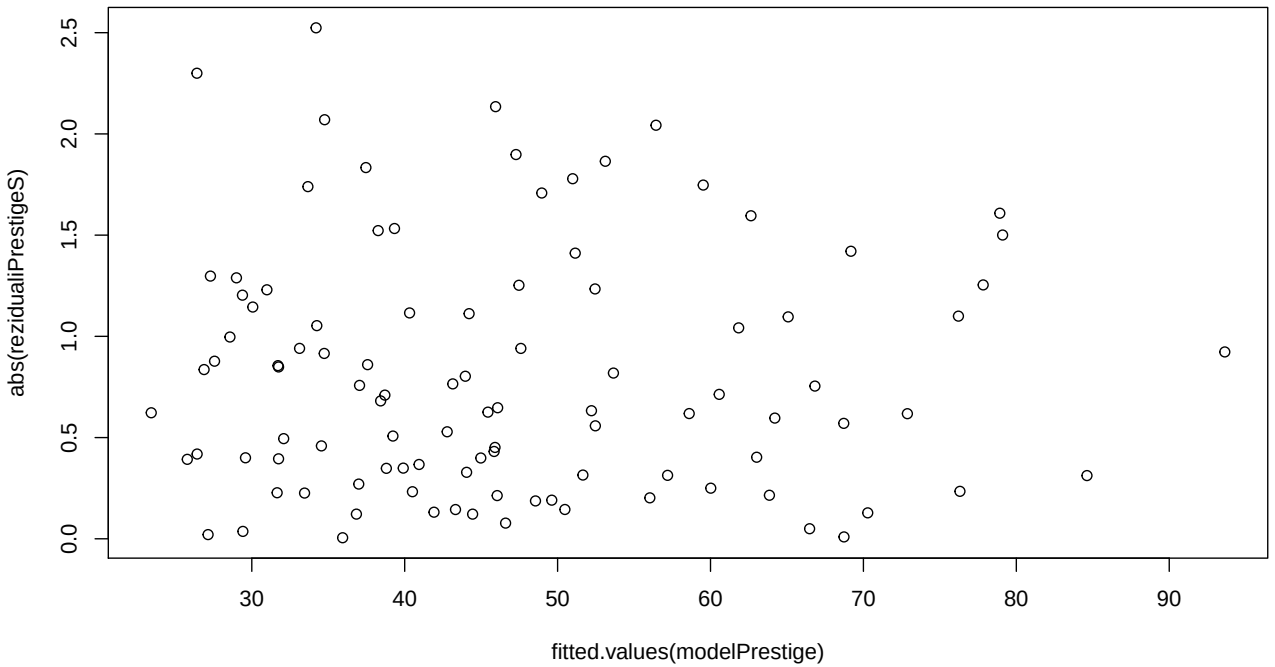
Provera homoskedastičnosti

```
library(lmtest)
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
bptest(modelPrestige)
##
##  studentized Breusch-Pagan test
##
## data:  modelPrestige
## BP = 4.1838, df = 2, p-value = 0.1235
```

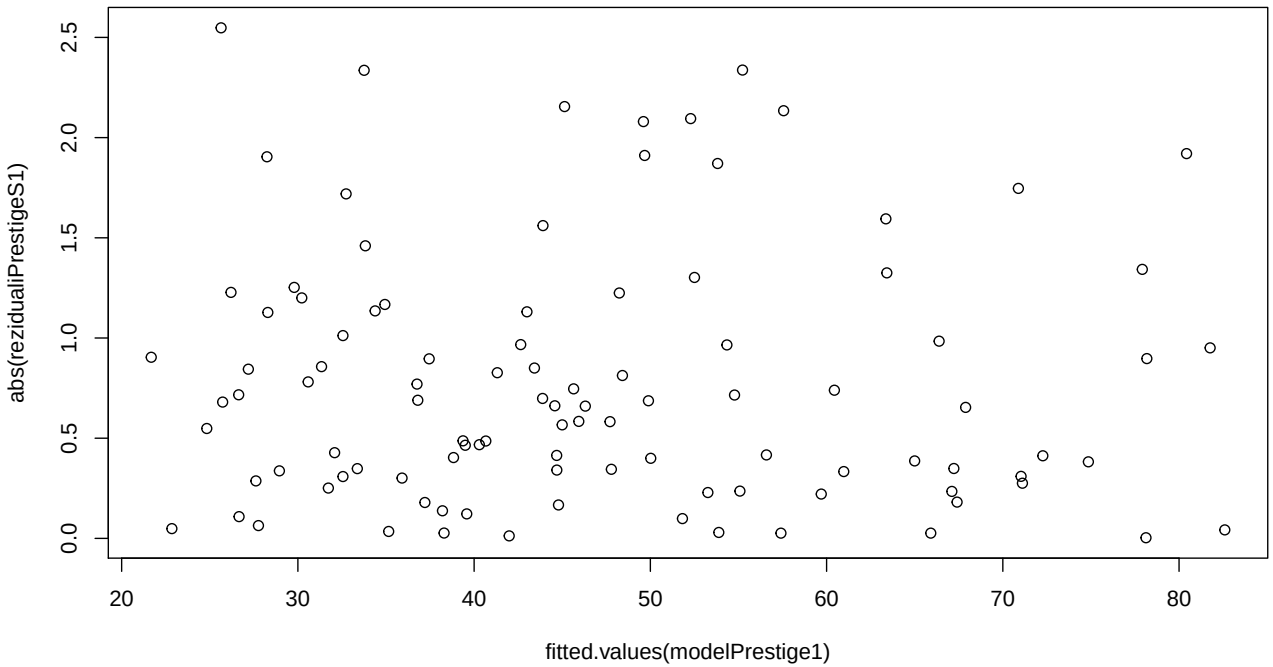
Provera homoskedastičnosti

```
bptest(modelPrestige1)
##
##  studentized Breusch-Pagan test
##
## data:  modelPrestige1
## BP = 7.5236, df = 3, p-value = 0.05695
bptest(modelPrestige2)
##
##  studentized Breusch-Pagan test
##
## data:  modelPrestige2
## BP = 4.1838, df = 2, p-value = 0.1235
```

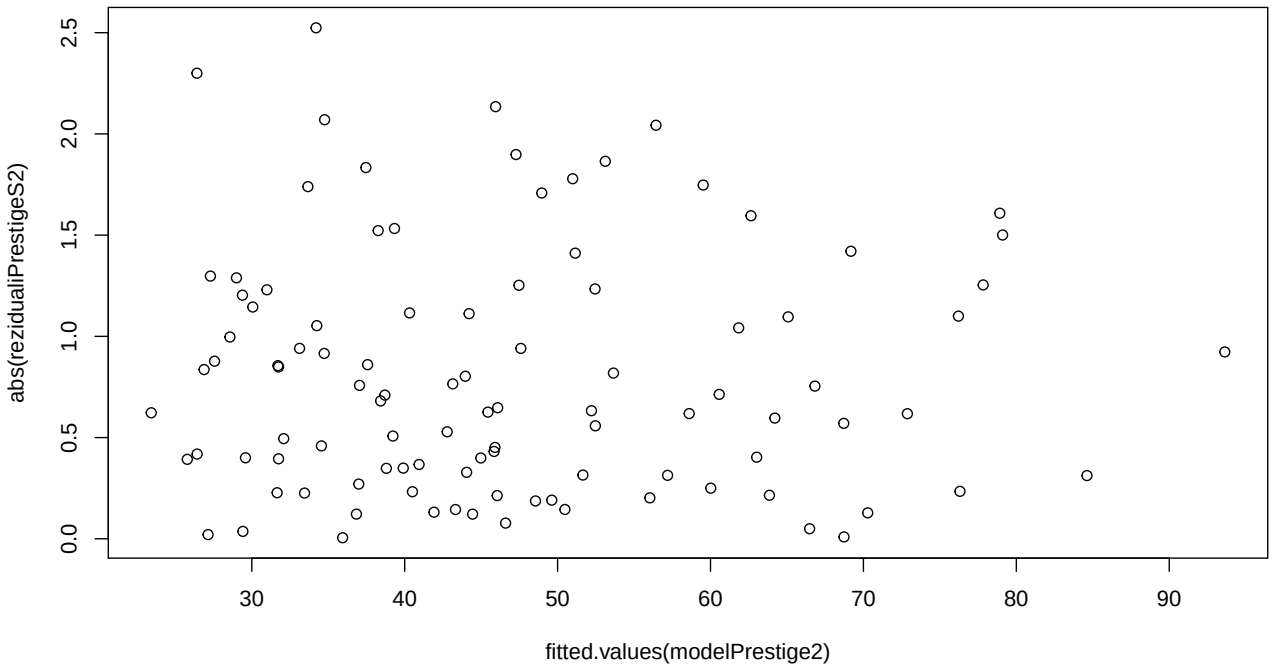
Provera homoskedastičnosti



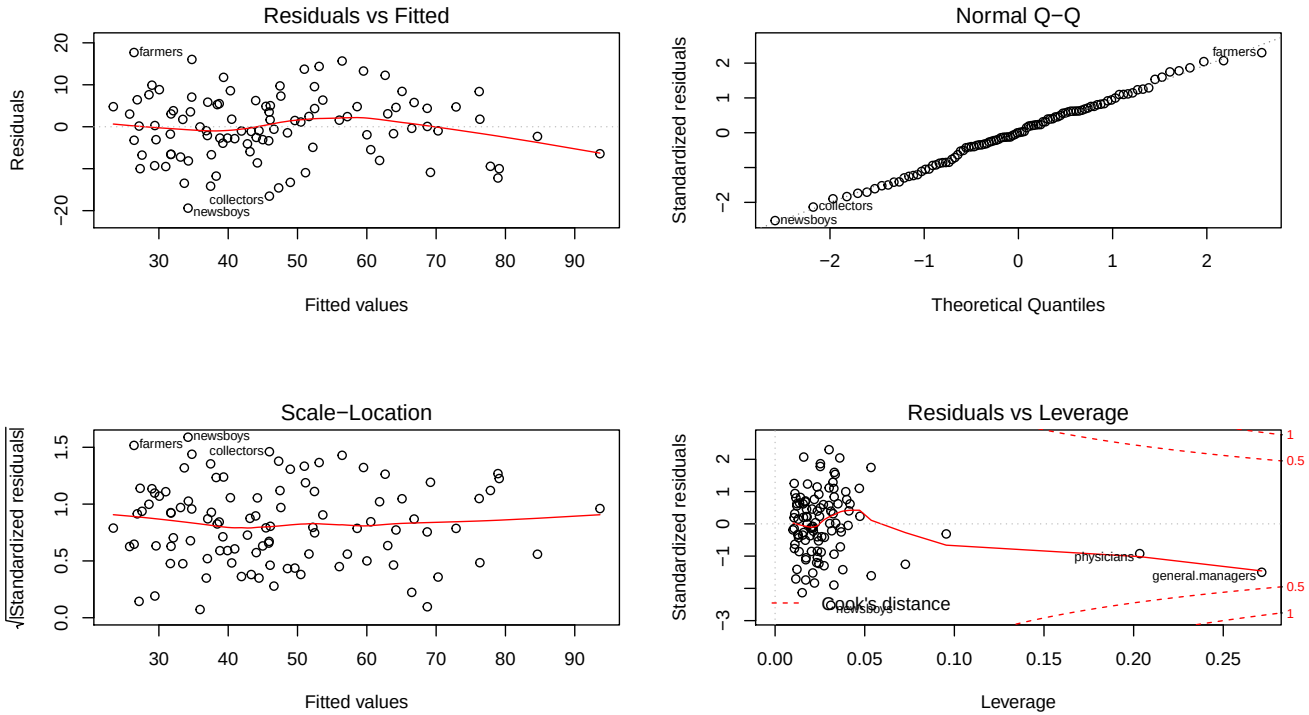
Provera homoskedastičnosti



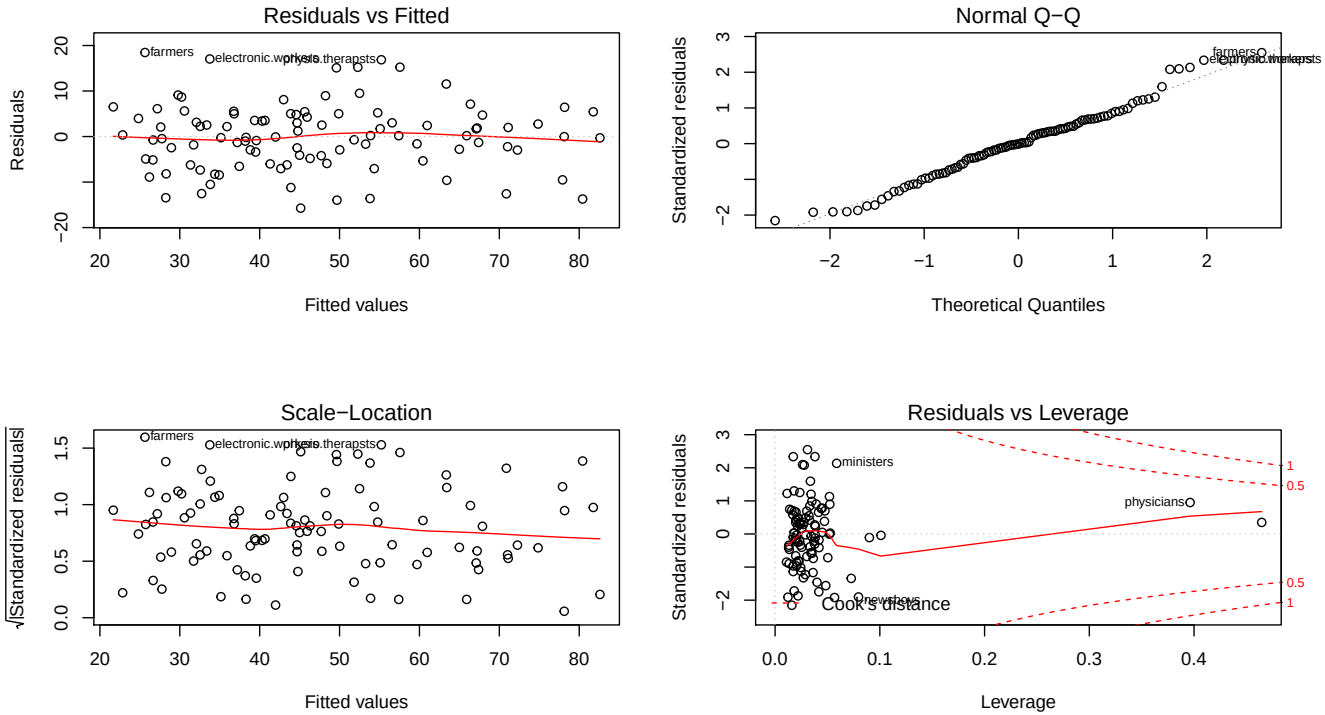
Provera homoskedastičnosti



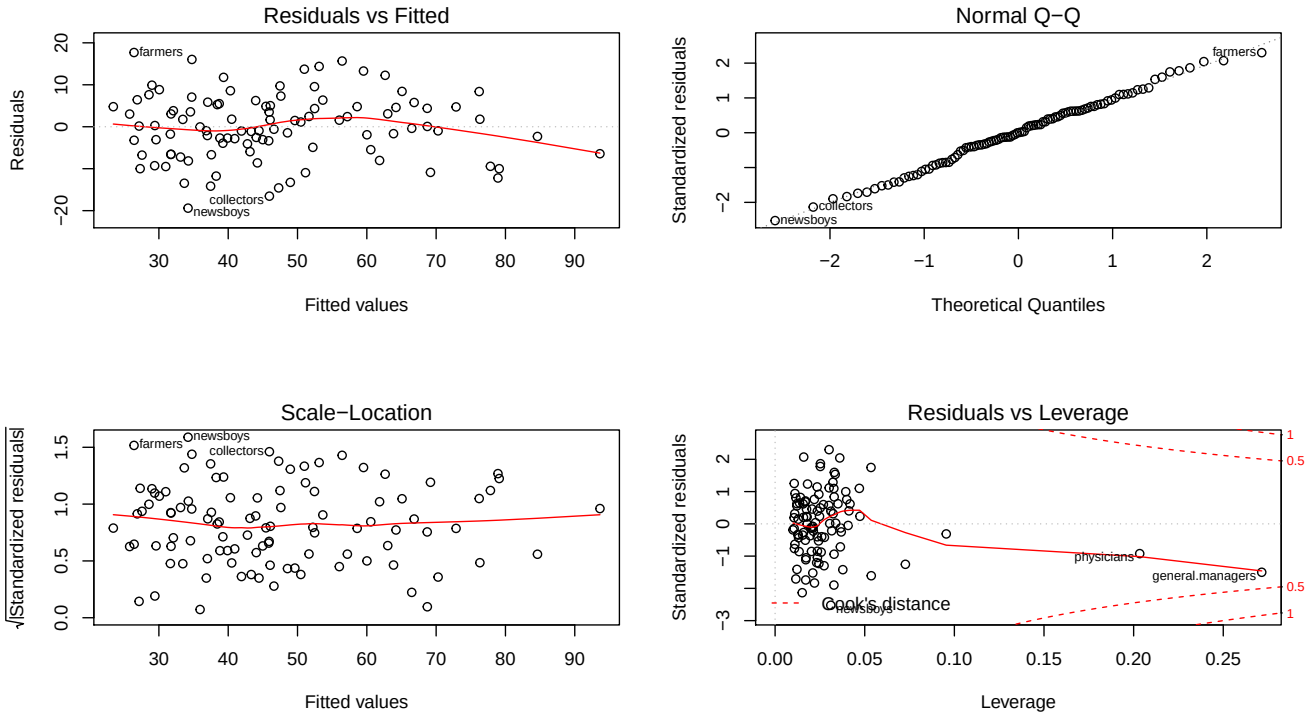
```
par(mfrow=c(2,2))
plot(modelPrestige)
```



```
par(mfrow=c(2,2))
plot(modelPrestige1)
```



```
par(mfrow=c(2,2))
plot(modelPrestige2)
```



Transformacije promenljivih

Neka je $\psi(y)$ transformacija zavisne promenljive. Označimo sa $m = E(Y)$. Tada je

$$\psi(Y) \approx \psi(m) + (Y - m)\psi'(m)$$
$$D(\psi(Y)) \approx (\psi'(m))^2 D(Y).$$

Da bi stabilizovali disperziju potrebno je da je

$$\psi'(m) = \frac{c}{\sqrt{D(y)}}.$$

Ako je $D(Y) \sim m$ onda je $\psi(y) = \sqrt{y}$. Ako je $D(Y) \sim m^2$ onda je $\psi(y) = \log y$.

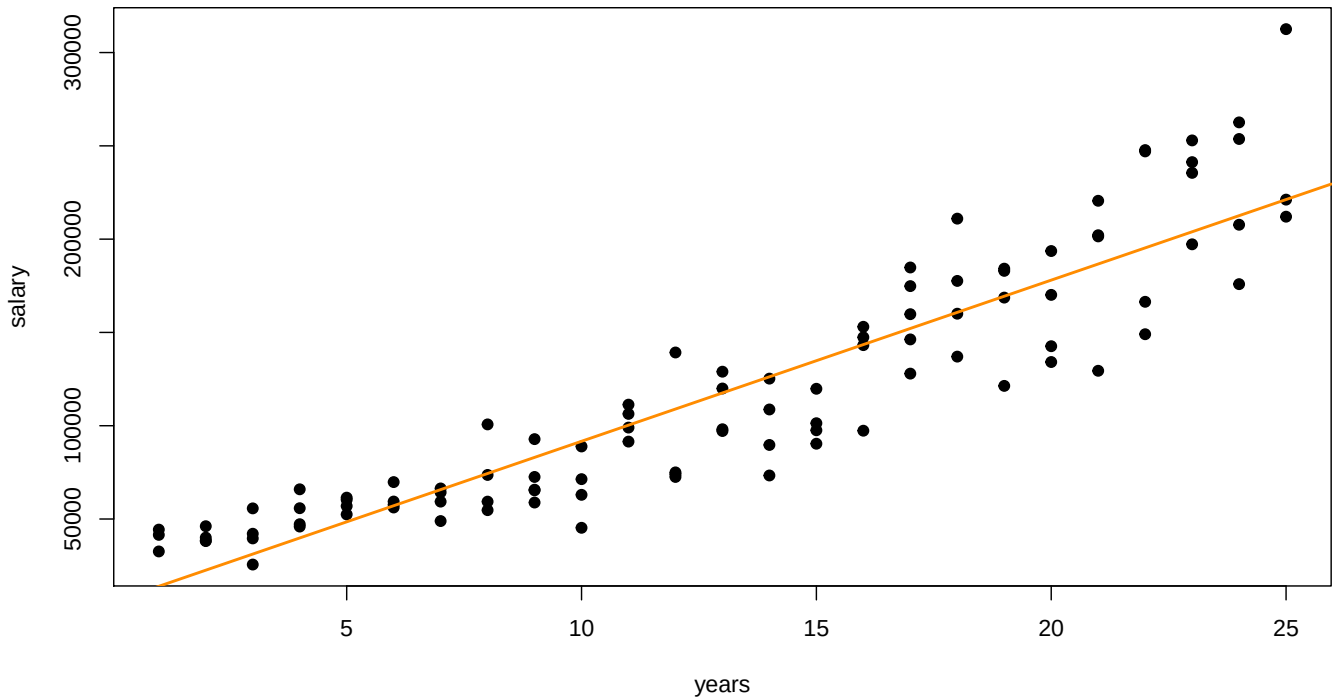
Primer

```
initech = read.csv("initech.csv")
head(initech)
```

```
##   years salary
## 1     1  41504
## 2     1  32619
## 3     1  44322
## 4     2  40038
## 5     2  46147
## 6     2  38447
```

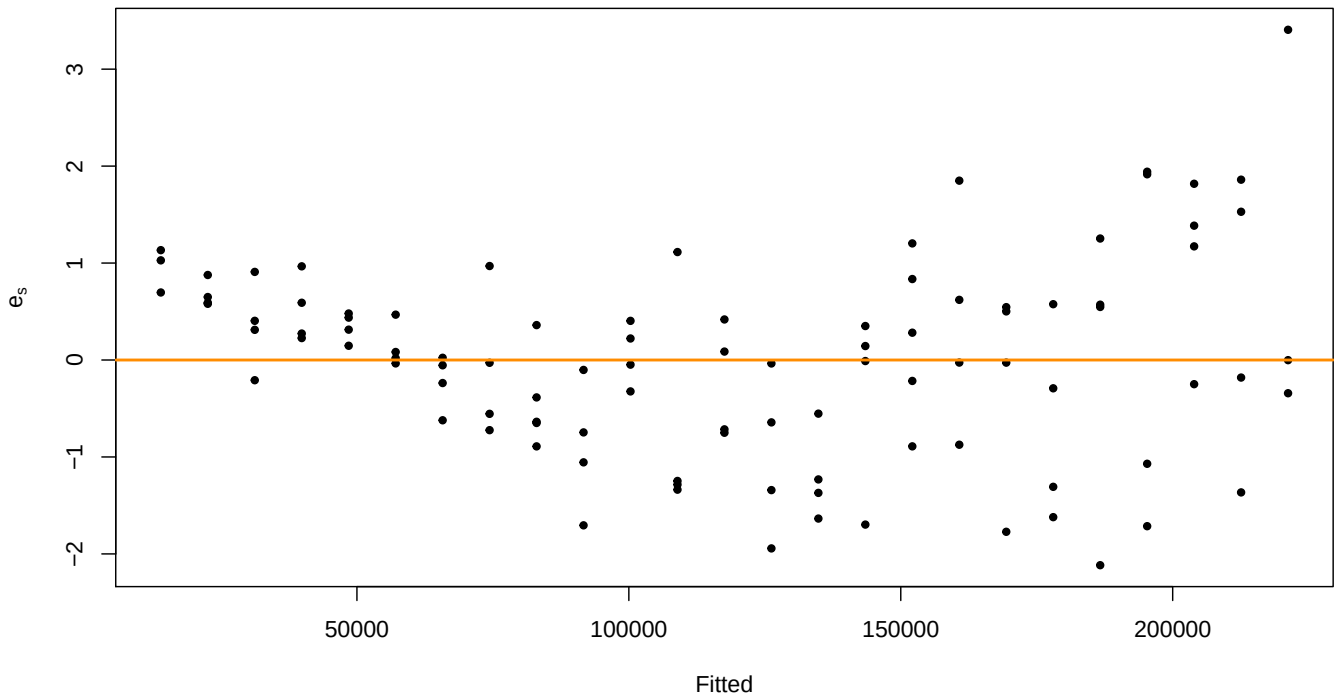
Primer

```
plot(salary ~ years, data = initech, col = "black", pch = 20,
initech_fit = lm(salary ~ years, data = initech)
abline(initech_fit, col = "darkorange", lwd = 2)
```



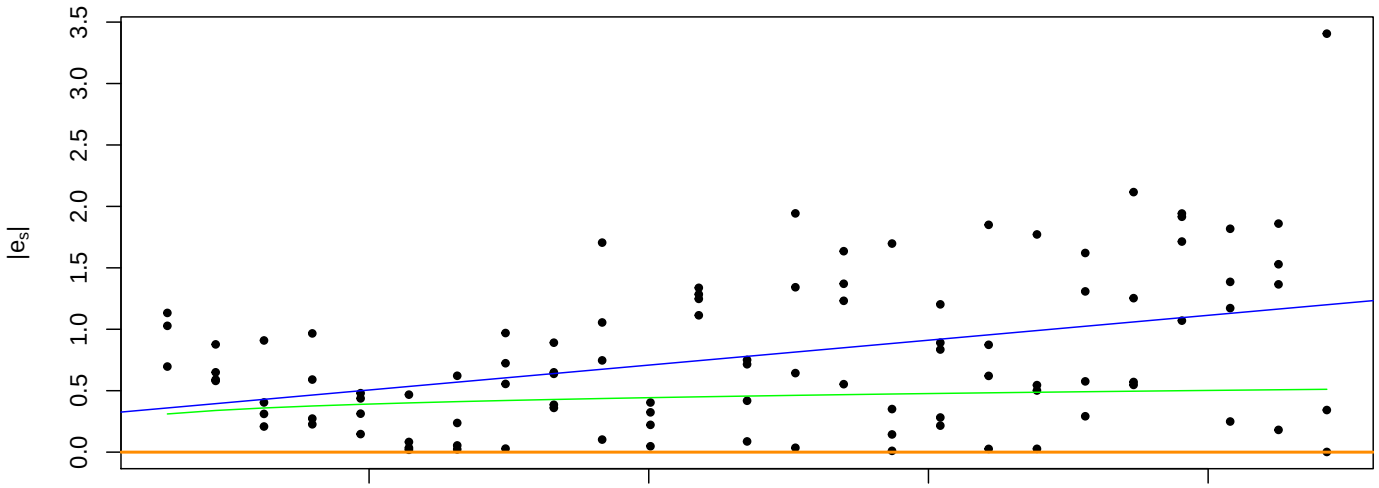
Primer

```
plot(fitted(initech_fit),rstandard(initech_fit), col = "black"
     xlab = "Fitted", ylab = expression(e["s"]))
abline(h = 0, col = "darkorange", lwd = 2)
```



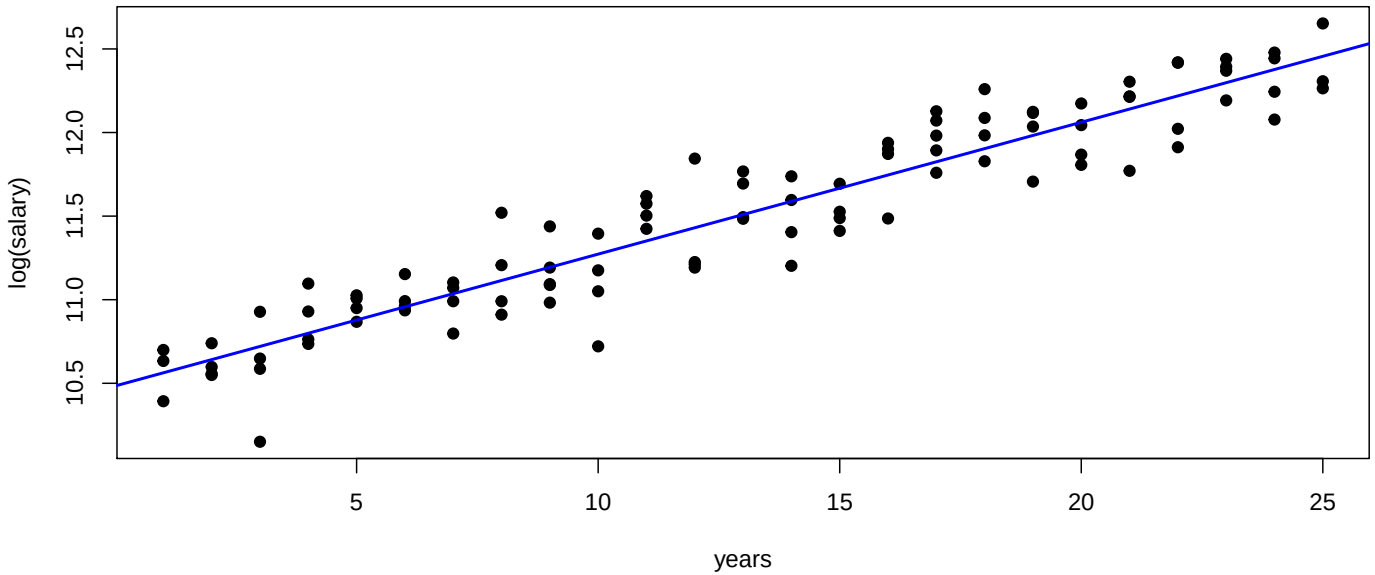
Primer

```
plot(fitted(initech_fit),abs(rstandard(initech_fit)),
     col = "black", pch = 20,
     xlab = "Fitted", ylab =expression(paste("|",e["s"],"|")))
abline(h = 0, col = "darkorange", lwd = 2)
abline(lm(abs(rstandard(initech_fit))~fitted(initech_fit)),
       col="blue")
lines(xrs,xrs^0.18*exp(-2.886),col="green")
```



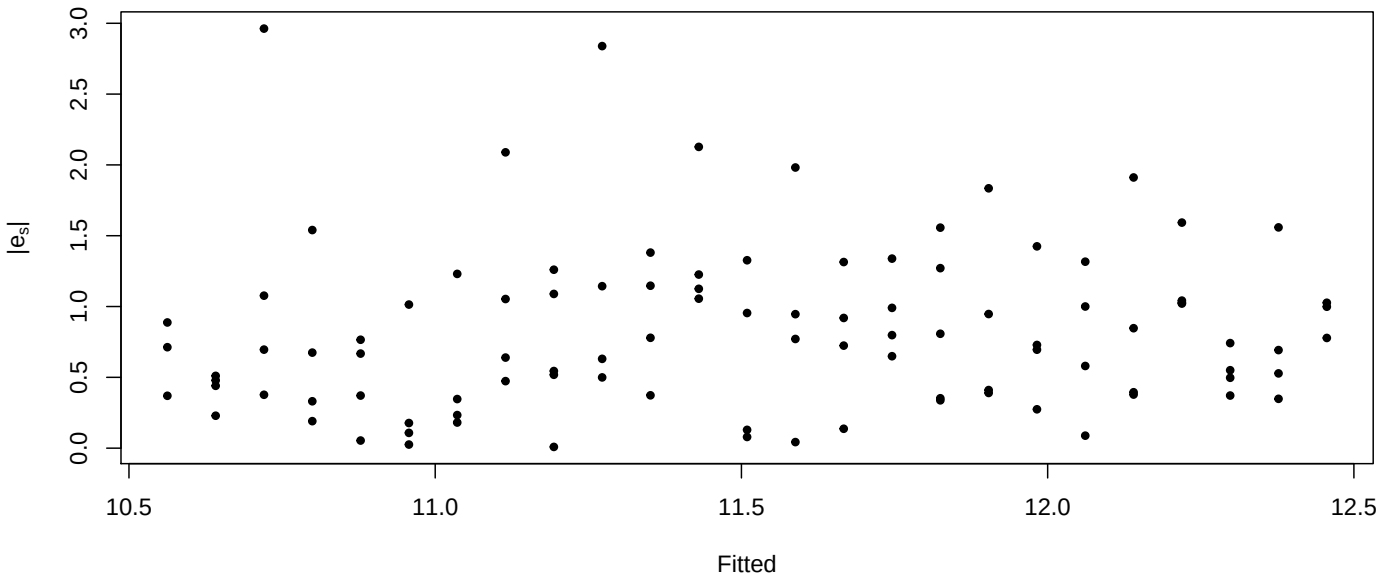
Primer

```
plot(log(salary) ~ years, data = initech, col = "black",
     pch = 20, cex = 1.5)
initech_fitLog = lm(log(salary) ~ years, data = initech)
abline(initech_fitLog, col = "blue", lwd = 2)
```



Primer

```
plot(fitted(initech_fitLog),abs(rstandard(initech_fitLog)),
     col = "black", pch = 20,
     xlab = "Fitted", ylab =expression(paste("|",e["s"],"|")))
```



Boks-Koksova transformacija

- Rešavamo problem odsustva normalnosti grešaka

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda}-1}{\lambda} & \text{za } \lambda \neq 0 \\ \log Y_i & \text{za } \lambda = 0. \end{cases}$$

Funkcija verodostojnosti je

$$L(\lambda, \sigma | Y) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{(Y^{(\lambda)}-X\beta)^T(Y^{(\lambda)}-X\beta)}{2\sigma^2}} (\prod_{i=1}^{n} Y_i)^{\lambda-1}.$$

Boks-Koksova transformacija

$$\log L(\lambda, \sigma | Y) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{(Y^{(\lambda)} - X\beta)^T (Y^{(\lambda)} - X\beta)}{2\sigma^2} + (\lambda - 1)\log(\prod_{i=1}^{n} Y_i)$$

Ocene za β i σ^2 se dobijaju kao i do sada, a λ je ona vrednost koja maksimizira funkciju

$$-\frac{(Y^{(\lambda)} - X\beta)^T (Y^{(\lambda)} - X\beta)}{2\sigma^2} + (\lambda - 1)\log(\prod_{i=1}^{n} Y_i).$$

Korišćenjem Vilksove teoreme dobija se da $2(\log L(\hat{\lambda}) - \log L(\lambda_0))$ ima graničnu χ_1^2 raspodelu.

Boks-Koksova transformacija

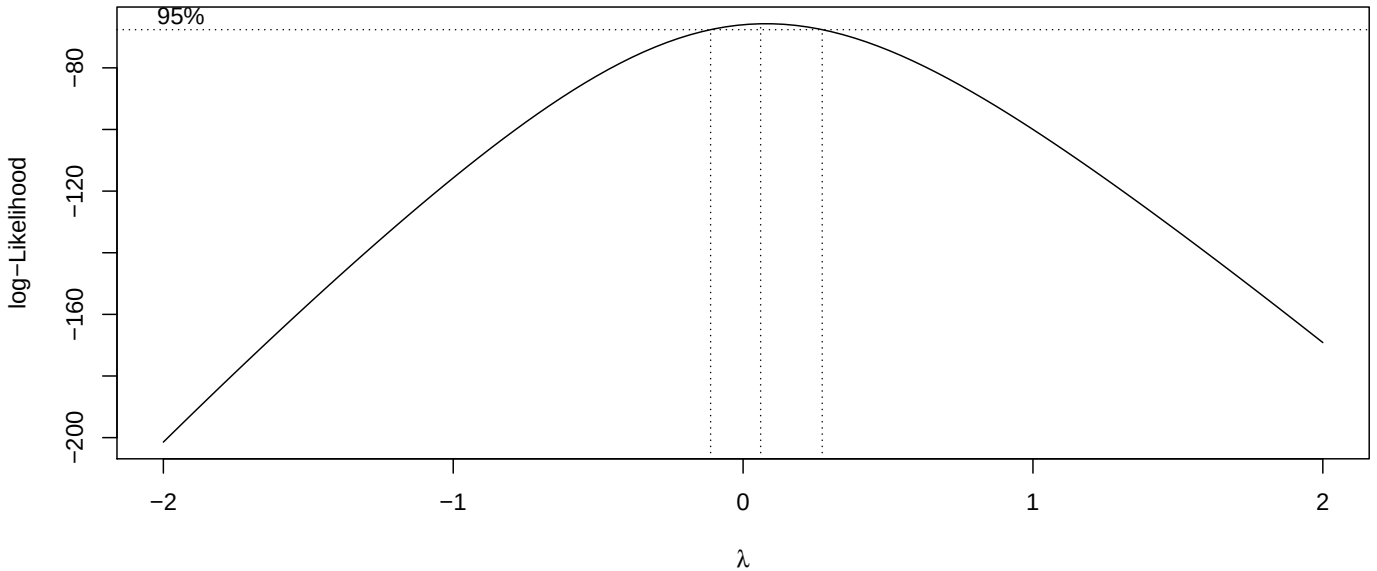
Ako je $Y_i > -a$, za neko $a > 0$ onda se može primeniti transformacija

$$Y_i^{(\lambda)} = \begin{cases} \frac{(Y_i+a)^\lambda - 1}{\lambda} & \text{za } \lambda \neq 0 \\ \log(Y_i + a) & \text{za } \lambda = 0. \end{cases}$$

Ukoliko je a nepoznato može se odrediti metodom maksimalne verodostojnosti. Tada $2(\log L(\hat{\lambda}) - \log L(\lambda_0))$ ima graničnu χ^2_2 raspodelu.

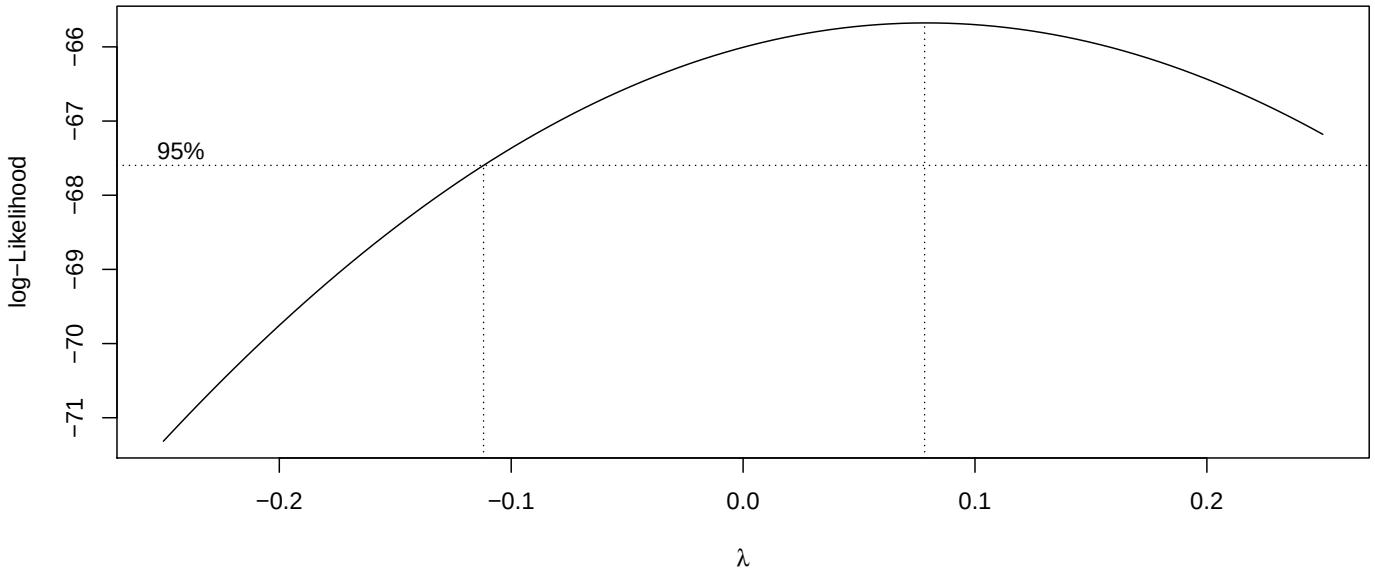
Primer

```
library(MASS)
boxcox(initech_fit)
```



Primer

```
boxcox(initech_fit,lambda = seq(-0.25, 0.25, length = 100))
```



Primer

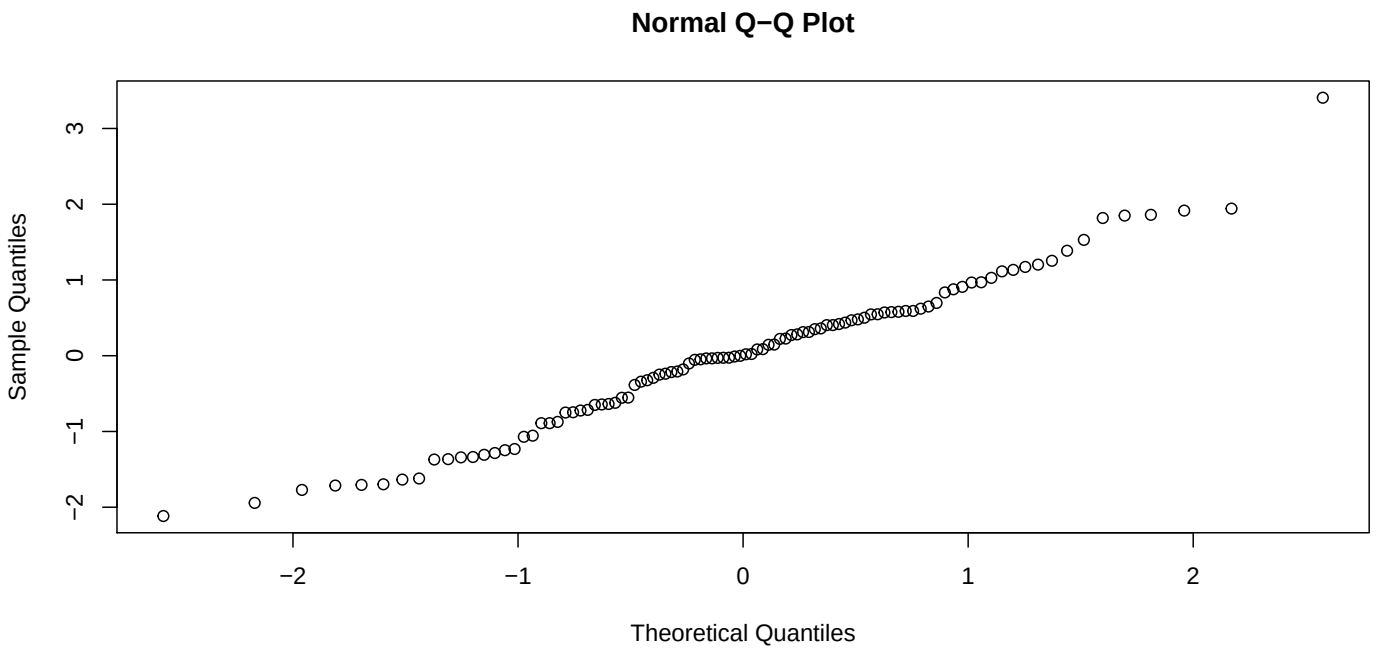
$H_0 : \lambda = 0$ Sada je $2(\log L(\hat{\lambda}) - \log L(\lambda_0))$

```
## [1] 0.6577234
```

$p - vred = 0.42$

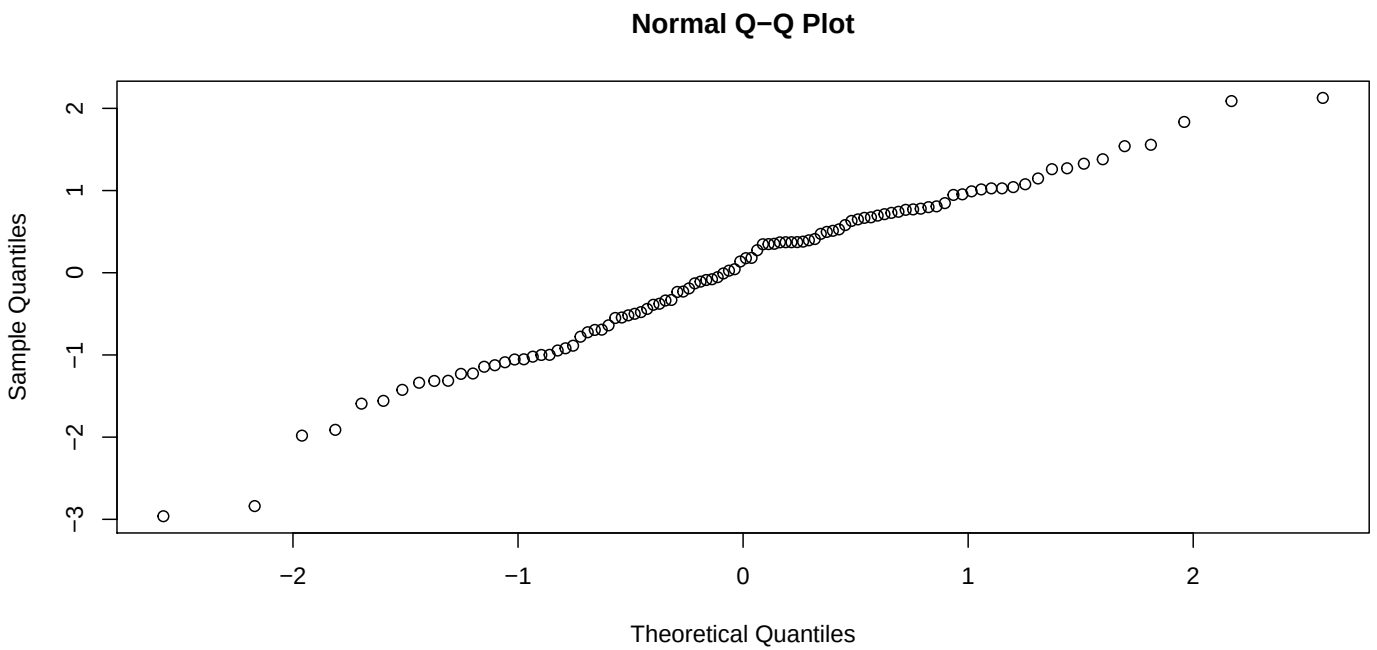
Primer-model bez transformacije

```
qqnorm(rstandard(initech_fit))
```



Primer

```
qqnorm(rstandard(initech_fitLog))
```



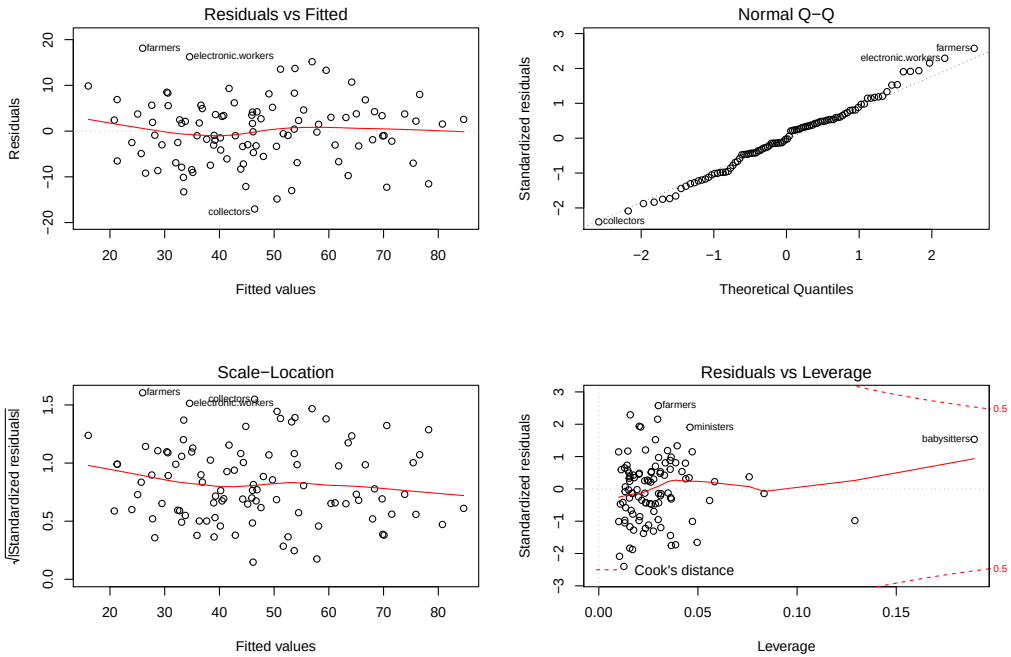
Boks-Tidvelova transformacija

Boks-Koks za prediktore

```
boxTidwell(prestige~income+education,data=Prestige)
```

```
##                MLE of lambda Score Statistic (z)  Pr(>|z|)
## income           -0.0020081             -4.3095 1.637e-05 ***
## education         1.8461610              1.7853   0.07421 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
## iterations =  10
```

```
modelPrestige3=lm(prestige~log(income)+education, data=Prestig
par(mfrow=c(2,2))
plot(modelPrestige3)
```



Težinska regresija

Pretpostavimo da je $D(Y_i) = \frac{\sigma^2}{w_i}$ gde je w_i poznata konstanta. Umesto modela

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad j = 1, 2..., n, \tag{1}$$

gde je $Cov\varepsilon = \Omega\sigma^2$ (Ω je dijagonalna matrica čiji su elementi w_i^{-1}). posmatramo model

$$Y_i\sqrt{w_i} = \beta_0\sqrt{w_i} + \sum_{j=1}^{p} \beta_j X_{ij}\sqrt{w_i} + \varepsilon_i\sqrt{w_i}, \quad j = 1, 2..., n. \tag{2}$$

Kako odabrati težine?

Ukoliko je $D(\varepsilon_i) \sim x_i^2$ onda je najbolje uzeti da je $w_i = \frac{1}{x_i}$. Ukoliko je za Y_i izvršeno n_i merenja $D(Y_i) = \frac{\sigma^2}{n_i}$ odakle je $w_i = n_i$.

MNK ocena se dobija minimiziranjem

$$\sum_{i=1}^{n}(Y_i\sqrt{w_i} - \beta_0\sqrt{w_i} - \sum_{j=1}^{p}\beta_j X_{ij}\sqrt{w_i})^2 = \sum_{i=1}^{n} w_i(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2$$

Vektorski oblik

Ω^{-1} možemo predstaviti u obliku $\Omega^{-1} = CC^T$ gde je C $n \times n$ dijagonalna matrica sa elementima $\sqrt{w_i}$ Tada je

$$CY = CX\beta + C\varepsilon,$$

i $CovC\varepsilon = CC^T Cov\varepsilon = CC^T\Omega\sigma^2 = \sigma^2 I$, odnosno uslovi Gaus-Markova su zadovoljeni pa je

$$\hat{\beta} = (X^T C^T CX)^{-1} X^T C^T CY = (X^T\Omega^{-1}X)^{-1} X^T\Omega^{-1}Y$$

$$Cov(\hat{\beta}) = \sigma^2(X^T\Omega^{-1}X)^{-1}$$

Vektorski oblik

Ocenjene vrednosti su $C\hat{Y} = CX\hat{\beta}$ pa je vektor reziduala novog modela 2

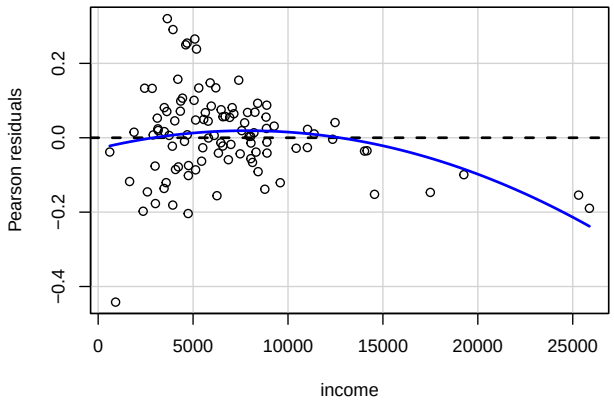
$$r = CY - CX\hat{\beta} = C(Y - \hat{Y}).$$

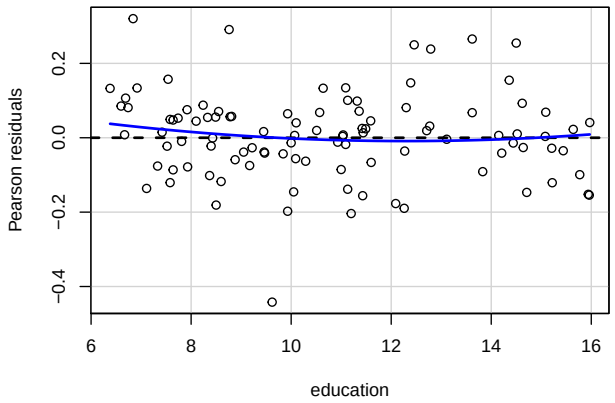
Nepristrasna ocena za σ^2 je $\hat{\sigma}^2 = \frac{1}{n-p-1} r^T r$

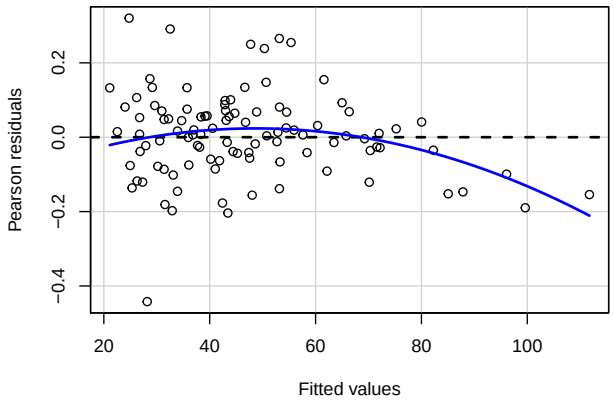
Sada se može primeniti Gaus-Markova teorema odakle vidimo stvarni značaj korišćenja ove regresije.

Primer

```
modelPrestige4=lm(prestige~income+education, data=Prestige,wei
residualPlots(modelPrestige4)
```







Testiranje nekorelisanosti

Alternativni model

$$\varepsilon_i = a\varepsilon_{i-1} + u_i, \tag{3}$$

gde niz $\{u_i\}$ zadovoljava uslove Gaus-Markova

Testiranje nekorelisanosti

Ocena metodom najmanjih kvadrata za parametar a je

$$\hat{a} = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2} \tag{4}$$

$$H_0 : a = 0$$

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Ukoliko su reziduali nekorelisani test statistika ima približnu vrednost 2. Vrednosti između 2 i 4 upućuju na negativnu korelisanost, a vrednosti između 0 i 2 na pozitivnu korelisanost.

Generalizovani metod najmanjih kvadrata

Pretpostavimo da je $Cov(\varepsilon) = \Sigma\sigma^2$ gde je Σ simetrična, pozitivno definitna matrica je ocena za β generalizovanom metodom najmanjih kvadrata data sa

$$\hat{\beta}_{GL} = (X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}Y$$

$\Sigma = MDM^T$ i $S = M\sqrt{D}$

$$S^{-1}Y = S^{-1}X\beta + S^{-1}\varepsilon$$
$$D(S^{-1}\varepsilon) = \sigma^2 I.$$

Primer

```
initechAve<-data.frame(years=unique(initech$years),salary=rep(
for(i  in 1:length(initechAve$years))
{
  initechAve[i,2]=mean(initech[initech$years==initechAve$years
  initechAve[i,3]=length(initech[initech$years==initechAve$yea

}
```

Primer

```
initech_fitS = lm(salary ~ years, data = initechAve)
initech_fitW = lm(salary ~ years, data = initechAve,
                  weights = count)
```

Primer

```
summary(initech_fitS)
```

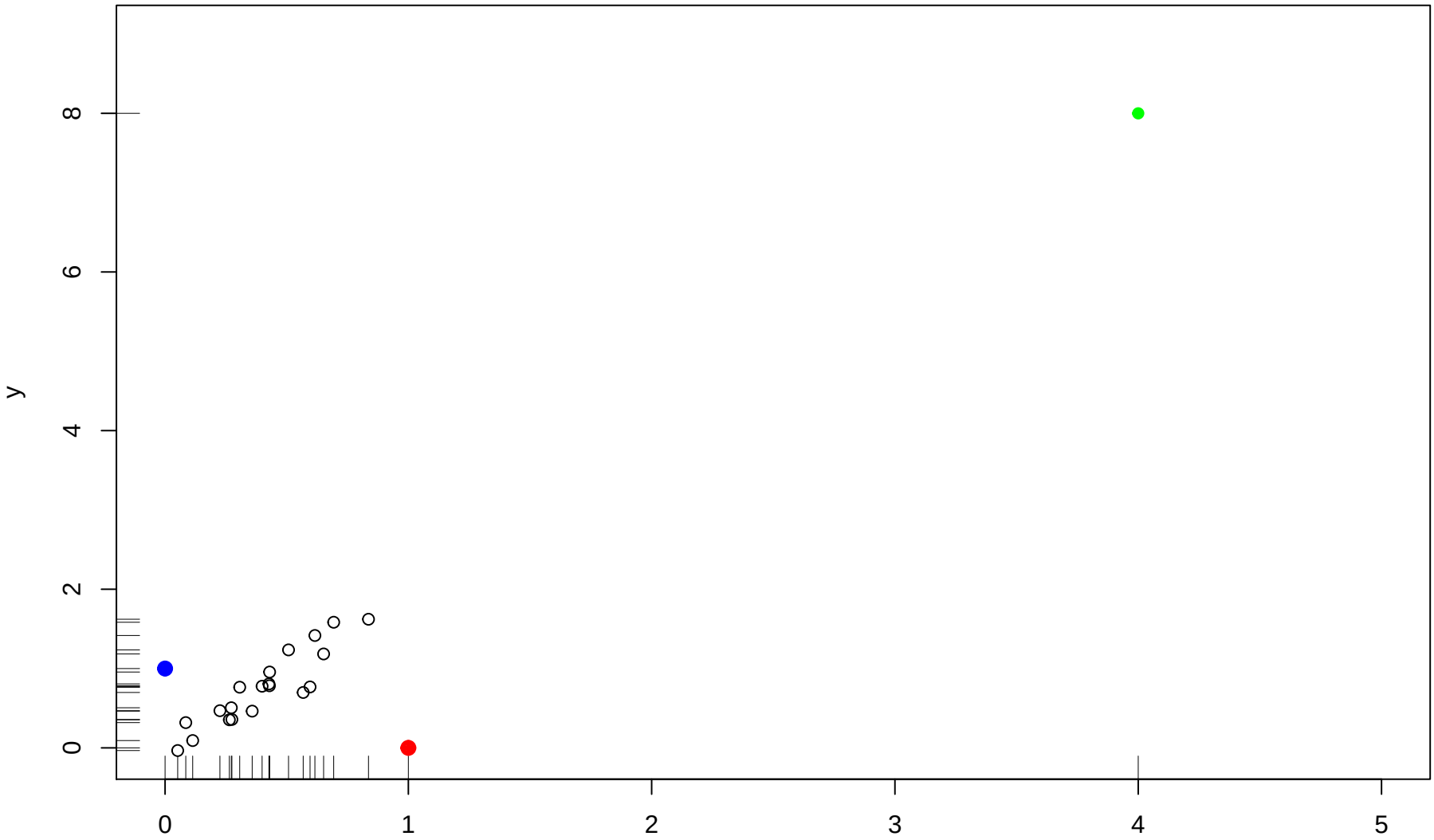
```
##
## Call:
## lm(formula = salary ~ years, data = initechAve)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -33145  -8802    1221   10127   27308
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6015.9     7000.8   0.859    0.399
## years         8626.5      470.9  18.318 3.25e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Primer

```
summary(initech_fitW)
```

```
##
## Call:
## lm(formula = salary ~ years, data = initechAve, weights = c
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -65165 -16499   3459  21319  55581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5302.1     6977.5    0.76    0.455
## years         8636.6      472.1   18.29 3.34e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

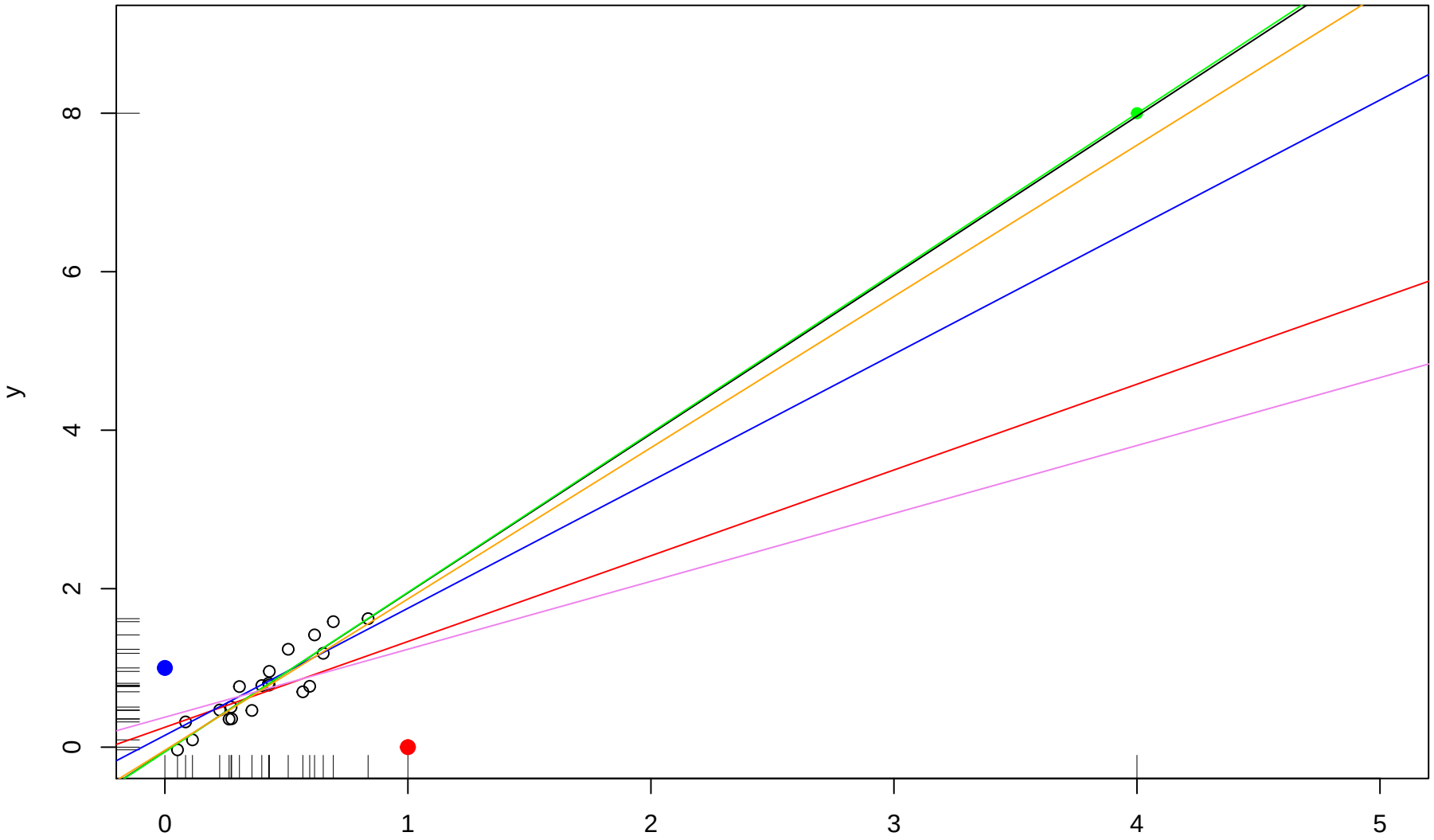
Autlajeri, težinske i uticajne tačke



Autlajeri, težinske i uticajne tačke

```
##                     (Intercept)         x
## crne                -0.05622806 2.0046664
## crne+plava           0.14885515 1.6035588
## crne+crvena          0.25081602 1.0823801
## crne+zelena         -0.06006866 2.0144332
## crne+plava+crvena    0.37809354 0.8570997
## sve                 -0.03876707 1.9087977
```

Autlajeri, težinske i uticajne tačke



Težinske tačke

$$e = (I - H)\varepsilon \quad D(e_i) = \sigma^2(1 - h_i)$$
$$h_i = (1, x_i)^T (XX^T)^{-1}(1, x_i)$$

Sa h_i se meri koliko je i-ta obzervacija odaljena od centra $\bar{X}$.

Kako je $tr(H) = p + 1$, tačke za koje je $h_i > \frac{2(p+1)}{n}$ možemo smatrati teškim.

U slučaju da ima puno obzervacija onda $h_i > 0.5$ smatramo da imaju veliku težinu, a one za koje je $h_i \in [0.2, 0.5]$ srednje težinskim tačkama.

Težinske tačke

```
round(hatvalues(model.sve),3)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## 0.057 0.067 0.904 0.044 0.049 0.045 0.045 0.061 0.052 0.050 0.050 0.044
##    13    14    15    16    17    18    19    20    21    22    23
## 0.045 0.044 0.043 0.059 0.044 0.047 0.045 0.063 0.050 0.046 0.049
```

Težinske tačke - interpretacija

$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$ i $z_{ik} = x_{ik} - \bar{x}_k$

$$
\begin{aligned}
Y_i &= \sum_{k=1}^{p}(z_{ik} + \bar{x}_k)\beta_k + \beta_0 + \varepsilon_i \\
&= \sum_{k=1}^{p} z_{ik}\beta_k + \underbrace{\sum_{k=1}^{p} \beta_k \bar{x}_k + \beta_0}_{\gamma} + \varepsilon_i
\end{aligned}
$$

$$
Y = (1, Z)\begin{pmatrix} \gamma \\ \beta^{\star} \end{pmatrix} + \varepsilon = \mathbf{Z}A + \varepsilon
$$

Težinske tačke - interpretacija

$$\hat{A} = (\mathbf{Z}^T\mathbf{Z})^{-1}Z^TY = \begin{pmatrix} n & \mathbf{0}^T \\ \mathbf{0} & (Z^TZ) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}^T \\ Z^T \end{pmatrix} Y$$

$$= \begin{pmatrix} \frac{1}{n} & \mathbf{0}^T \\ \mathbf{0} & (Z^TZ)^{-1} \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ Z^TY \end{pmatrix} = \begin{pmatrix} \bar{Y} \\ (Z^TZ)^{-1}Z^TY \end{pmatrix}$$

Važno: $\hat{\gamma} = \bar{Y}$ i $\hat{Y} - \bar{Y} = Z\hat{\beta}^\star = Z(Z^TZ)^{-1}Z^TY$

$$Cov(\hat{A}) = \sigma^2 \begin{pmatrix} n & \mathbf{0} \\ \mathbf{0} & (Z^TZ) \end{pmatrix}^{-1}$$

Težinske tačke - interpretacija

$\hat{Y} - \bar{Y} = \widetilde{H}Y$ gde je $\widetilde{H} = Z(Z^TZ)^{-1}Z^T$

Važi: $\hat{Y} = HY$ pa je $\widetilde{H} = H - \frac{1}{n}J$, odnosno

$$\widetilde{h}_i = h_i - \frac{1}{n}$$

$$\widetilde{h}_i = (x_i - \bar{x})^T(Z^TZ)^{-1}(x_i - \bar{x})$$

Težinske tačke - interpretacija

$\frac{1}{n-1}(Z^T Z) = C$ je uzoračka kovarijaciona matrica

Mahalonobiusovo rastojanje

$$MD_i^2 = (x_i - \bar{x})^T C^{-1}(x_i - \bar{x}) = (n-1)(h_i - \frac{1}{n})$$

Kada bi X imalo višedimenzionu normalnu raspodelu onda bi MD_i^2 imalo približno χ_p^2 raspodelu

```
##  [1]  0.30  0.52 18.93  0.01  0.11  0.03  0.02  0.38  0.19  0.14  0.14
## [12]  0.00  0.03  0.01  0.00  0.33  0.00  0.07  0.03  0.43  0.15  0.05
## [23]  0.11
```

```
## [1] 3.841459
```

- robusne modifikacije

Studentizovani reziduali

$$e_i^\star = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}},$$

gde je $\hat{\sigma}_{(i)}$ nepristrasna ocena disperzije kada se iz modela izbaci i-ta obzervacija

Mogu se prikazati u obliku

$$e_i^\star = e_i\Big(\frac{n-p-2}{\hat{\sigma}^2(1-h_i)(n-p-1)-e_i^2}\Big)^{\frac{1}{2}}$$

Mere uticaja

$$DFBETA_i = \hat{\beta} - \hat{\beta}_{(i)} = \frac{(XX^T)^{-1}x_i e_i}{1 - h_i}$$

$$DFFIT_i = \hat{y}_i - \hat{y}_{i,(i)} = \frac{h_i e_i}{1 - h_i},$$

pri čemu je $\hat{y}_{i,(i)}$ je prognoza i-te vrednosti kada je iz modela isključena i-ta obzervacija

Umesto njih se često koriste standardizovane verzije tih mera uticaja

Mere uticaja

Tada su rezudali modela u i-toj obzervaciji

$$e_{i,(i)} = y_i - \hat{y}_{i,(i)} = \frac{e_i}{1 - h_i}$$

$$D(e_{i,(i)}) = \frac{D(e_i)}{(1 - h_i)^2} = \frac{\sigma^2}{1 - h_i}.$$

Primetimo da što je h_i veće, $e_{i,(i)}$ će biti veće u odnosu na polazni rezidual e_i.

Disperzija ovog reziduala se može prikazati i u obliku

$$D(e_{i,(i)}) = \sigma^2(1 + X_i^T(X_{(i)}X_{(i)})^{-1}X_i),$$

gde je $X_{(i)}$ dizajn matrica bez i-te obzervacije.

Mere uticaja

$$\frac{e_{i,(i)}}{\frac{\hat{\sigma}_{(i)}}{\sqrt{1-h_i}}} = \frac{\frac{e_i}{1-h_i}}{\frac{\hat{\sigma}_{(i)}}{\sqrt{1-h_i}}} = e_i^{\star} \sim t_{n-p-2}$$

Sada se može definisati jednostavan test za određivanje autlajera zasnovan na tome da eksterno studentizovane rezudale karakteriše velika aposolutna vrednost.

Mere uticaja

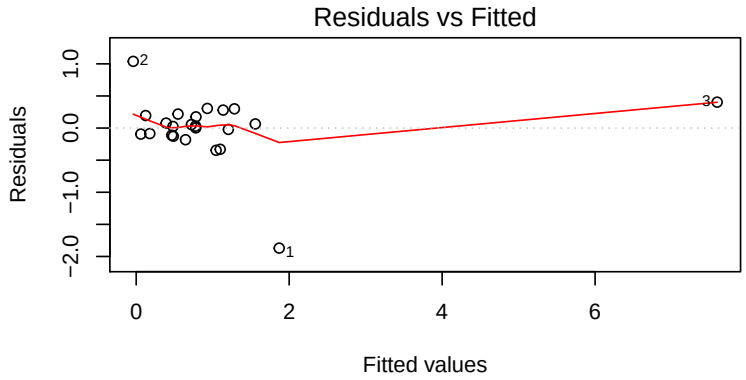
KUKOVO RASTOJANJE

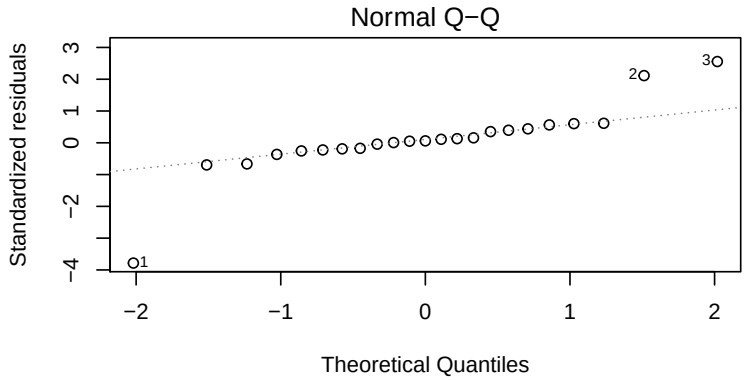
$$
\begin{aligned}
D_i &= \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(i)})}{(p+1)\hat{\sigma}^2} \\
&= \frac{(\hat{Y}_i - \hat{Y}_{(i)})^T (\hat{Y}_i - \hat{Y}_{(i)})}{(p+1)\hat{\sigma}^2} \\
&= \frac{e_i^2 h_i}{(p+1)\hat{\sigma}^2 (1-h_i)^2} = \frac{(e_i^s)^2}{p+1} \cdot \frac{h_i}{1-h_i}
\end{aligned}
$$

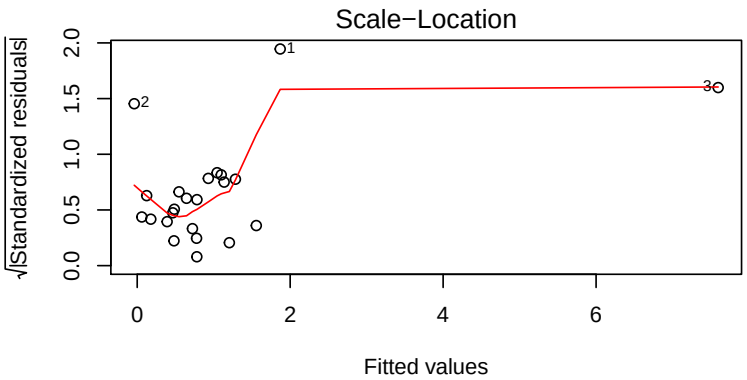
Dogovor je da se tačke za koje je Kukokvo rastojanje veće od 1 smatraju uticajnim, ali da treba obratiti pažnju i na one sa rastojanjem većem od 0.5. Do zaključka se može doći poređenjem sa kvantilima Fišerove $F_{p+1,n-p-1}$. Sve što je veće od 50% kvantila se može smatrati velikim rastojanjem.

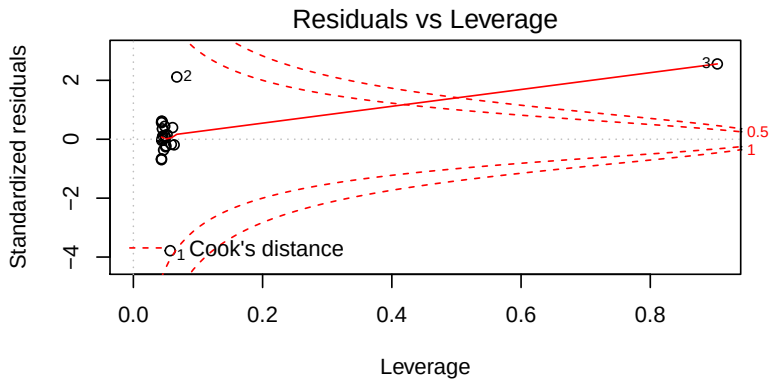
Zaključivanje

```
par(mfrow = c(2, 2))
plot(model.sve)
```









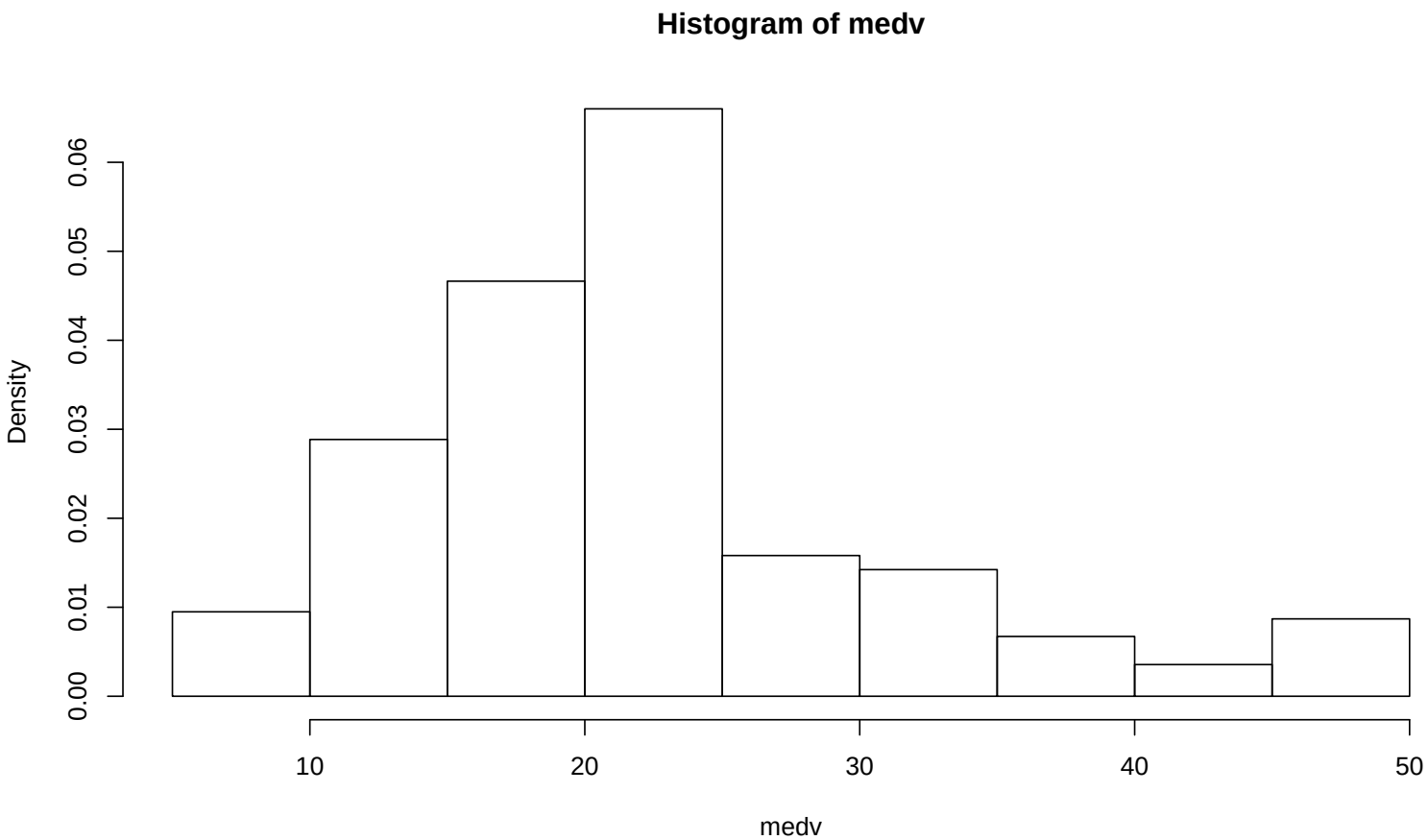
Primer

```
library(MASS)
attach(Boston)
```

- crim: per capita crime rate by town.
- zn: proportion of residential land zoned for lots over 25,000 sq.ft.
- indus: proportion of non-retail business acres per town.
- chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- nox: nitrogen oxides concentration (parts per 10 million).
- rm: average number of rooms per dwelling.
- age: proportion of owner-occupied units built prior to 1940.
- dis: weighted mean of distances to five Boston employment centres.
- rad: index of accessibility to radial highways.
- tax: full-value property-tax rate per $10,000.
- ptratio: pupil-teacher ratio by town.
- black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
- lstat: lower status of the population (percent).
- medv: median value of owner-occupied homes in $1000s

Primer

```
hist(medv,prob=TRUE)
```



Primer

```
(cormat<-round(cor(Boston),2))
```

```
##           crim    zn indus  chas   nox    rm   age   dis   rad   tax
## crim      1.00 -0.20  0.41 -0.06  0.42 -0.22  0.35 -0.38  0.63  0.58
## zn       -0.20  1.00 -0.53 -0.04 -0.52  0.31 -0.57  0.66 -0.31 -0.31
## indus     0.41 -0.53  1.00  0.06  0.76 -0.39  0.64 -0.71  0.60  0.72
## chas     -0.06 -0.04  0.06  1.00  0.09  0.09  0.09 -0.10 -0.01 -0.04
## nox       0.42 -0.52  0.76  0.09  1.00 -0.30  0.73 -0.77  0.61  0.67
## rm       -0.22  0.31 -0.39  0.09 -0.30  1.00 -0.24  0.21 -0.21 -0.29
## age       0.35 -0.57  0.64  0.09  0.73 -0.24  1.00 -0.75  0.46  0.51
## dis      -0.38  0.66 -0.71 -0.10 -0.77  0.21 -0.75  1.00 -0.49 -0.53
## rad       0.63 -0.31  0.60 -0.01  0.61 -0.21  0.46 -0.49  1.00  0.91
## tax       0.58 -0.31  0.72 -0.04  0.67 -0.29  0.51 -0.53  0.91  1.00
## ptratio   0.29 -0.39  0.38 -0.12  0.19 -0.36  0.26 -0.23  0.46  0.46
## black    -0.39  0.18 -0.36  0.05 -0.38  0.13 -0.27  0.29 -0.44 -0.44
## lstat     0.46 -0.41  0.60 -0.05  0.59 -0.61  0.60 -0.50  0.49  0.54
## medv     -0.39  0.36 -0.48  0.18 -0.43  0.70 -0.38  0.25 -0.38 -0.47
##           ptratio black lstat  medv
## crim         0.29 -0.39  0.46 -0.39
## zn          -0.39  0.18 -0.41  0.36
## indus        0.38 -0.36  0.60 -0.48
## chas        -0.12  0.05 -0.05  0.18
## nox          0.19 -0.38  0.59 -0.43
## rm          -0.36  0.13 -0.61  0.70
## age          0.26 -0.27  0.60 -0.38
## dis         -0.23  0.29 -0.50  0.25
## rad          0.46 -0.44  0.49 -0.38
## tax          0.46 -0.44  0.54 -0.47
## ptratio      1.00 -0.18  0.37 -0.51
## black       -0.18  1.00 -0.37  0.33
## lstat        0.37 -0.37  1.00 -0.74
```

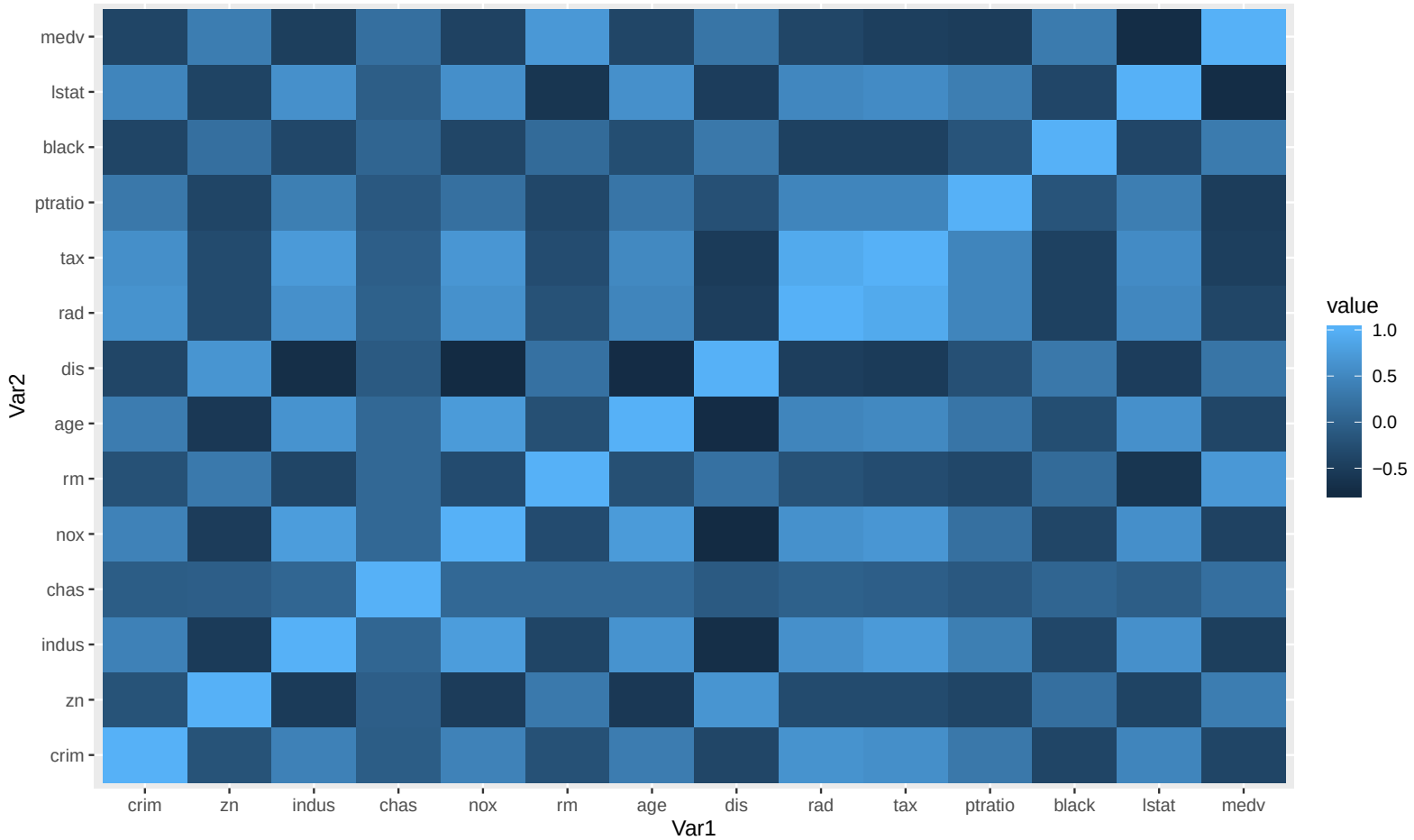
Primer

```
library(reshape2)
melted_cormat <- melt(cormat)
head(melted_cormat)
```

```
##    Var1 Var2 value
## 1  crim crim  1.00
## 2    zn crim -0.20
## 3 indus crim  0.41
## 4  chas crim -0.06
## 5   nox crim  0.42
## 6    rm crim -0.22
```

Primer

```
library(ggplot2)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()
```

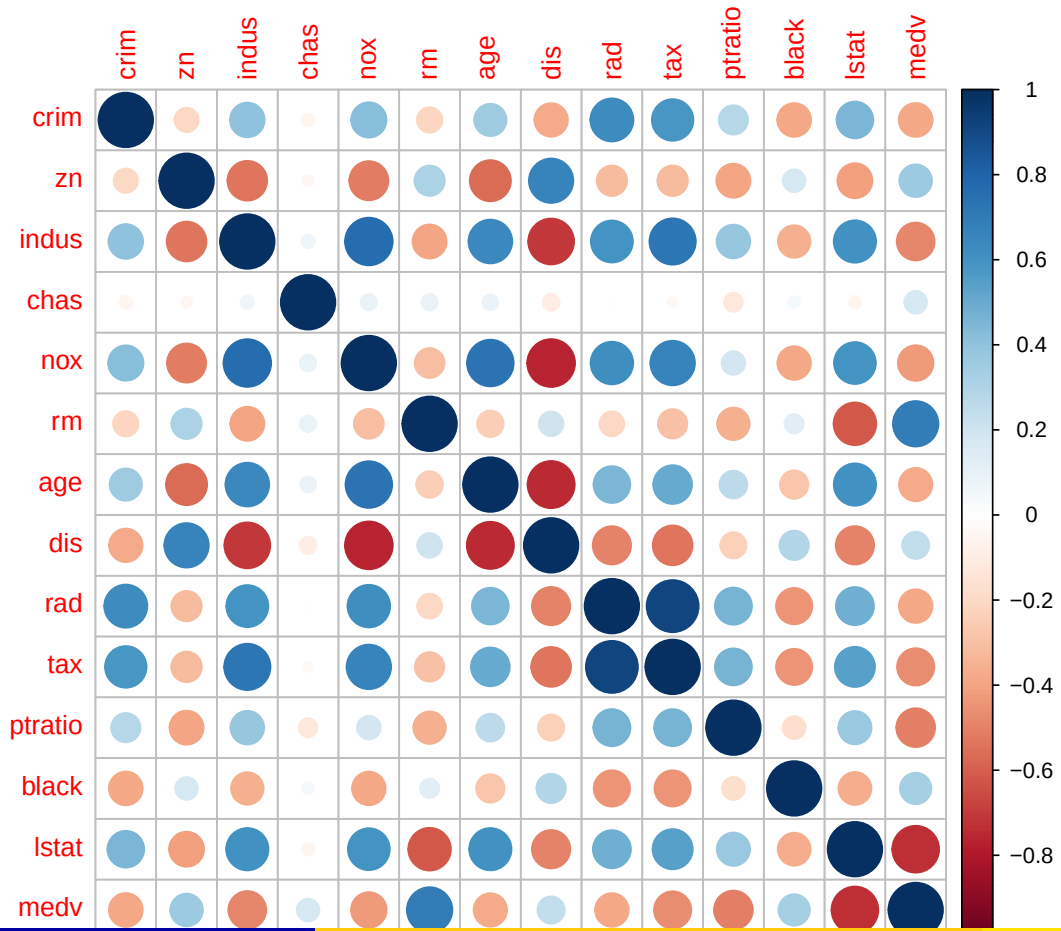


Primer

```
library(corrplot)
```

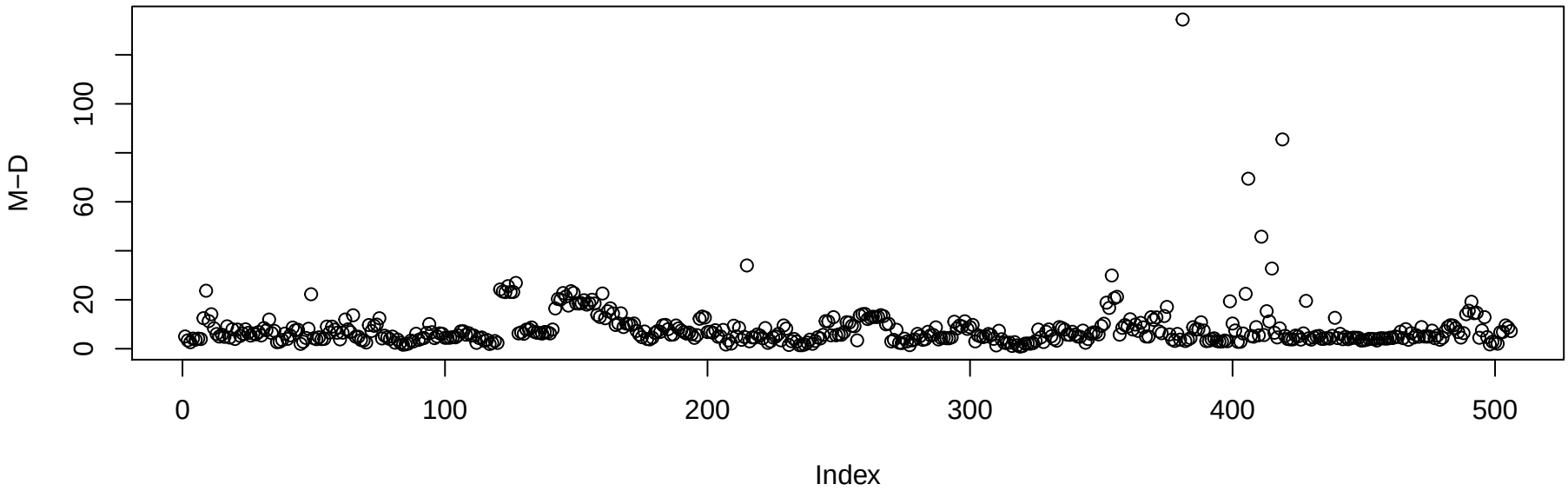
```
## corrplot 0.84 loaded
```

```
corrplot::corrplot(cor(Boston))
```



Primer

```
modelBoston<-lm(medv~lstat+dis+age+ptratio+indus+nox+tax+crim)
xBoston=cbind(lstat,dis,age,ptratio,indus,nox,tax,crim)
plot(mahalanobis(xBoston,center=colMeans(xBoston),cov=cov(xBoston)),ylab='M-D')
```




```
which.max(mahalanobis(xBoston,center=colMeans(xBoston),cov=cov(xBoston)))
```

```
## [1] 381
```

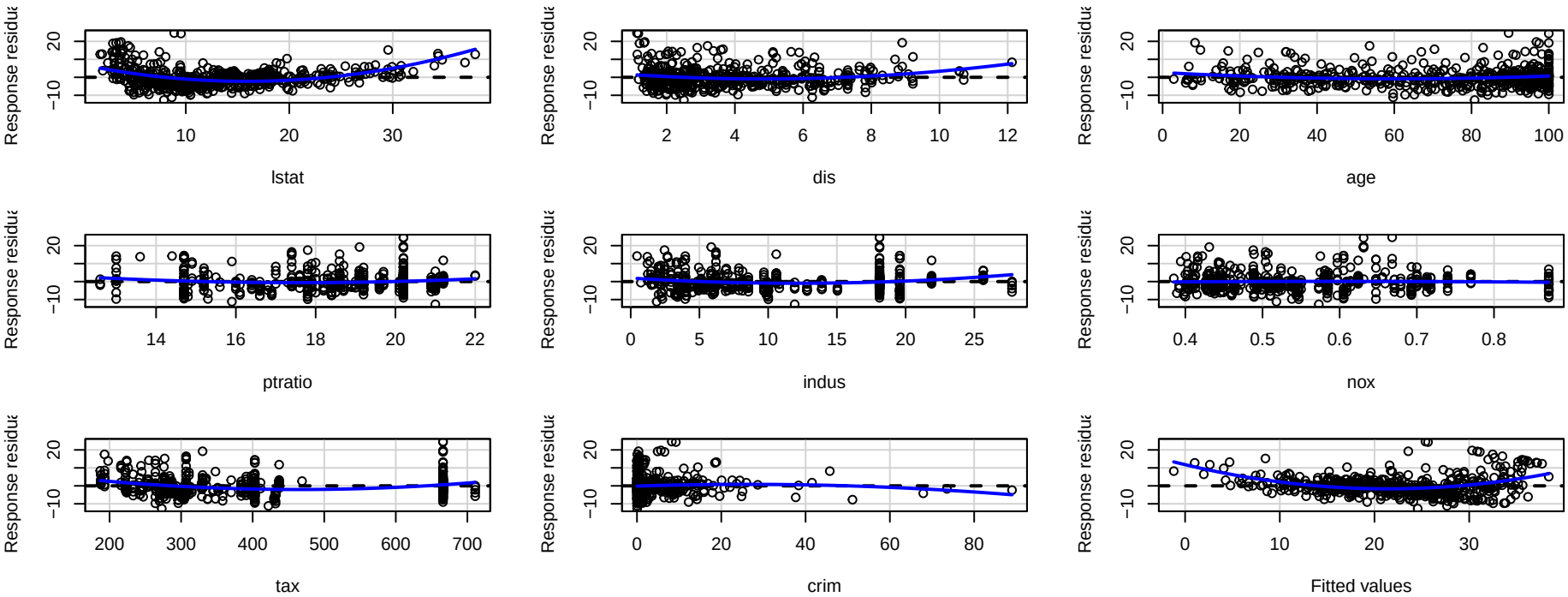
Primer

```
summary(xBoston)
```

```
##      lstat            dis              age            ptratio     
##  Min.   : 1.73   Min.   : 1.130   Min.   :  2.90   Min.   :12.60  
##  1st Qu.: 6.95   1st Qu.: 2.100   1st Qu.: 45.02   1st Qu.:17.40  
##  Median :11.36   Median : 3.207   Median : 77.50   Median :19.05  
##  Mean   :12.65   Mean   : 3.795   Mean   : 68.57   Mean   :18.46  
##  3rd Qu.:16.95   3rd Qu.: 5.188   3rd Qu.: 94.08   3rd Qu.:20.20  
##  Max.   :37.97   Max.   :12.127   Max.   :100.00   Max.   :22.00  
##      indus            nox              tax             crim        
##  Min.   : 0.46   Min.   :0.3850   Min.   :187.0   Min.   : 0.00632  
##  1st Qu.: 5.19   1st Qu.:0.4490   1st Qu.:279.0   1st Qu.: 0.08204  
##  Median : 9.69   Median :0.5380   Median :330.0   Median : 0.25651  
##  Mean   :11.14   Mean   :0.5547   Mean   :408.2   Mean   : 3.61352  
##  3rd Qu.:18.10   3rd Qu.:0.6240   3rd Qu.:666.0   3rd Qu.: 3.67708  
##  Max.   :27.74   Max.   :0.8710   Max.   :711.0   Max.   :88.97620
```

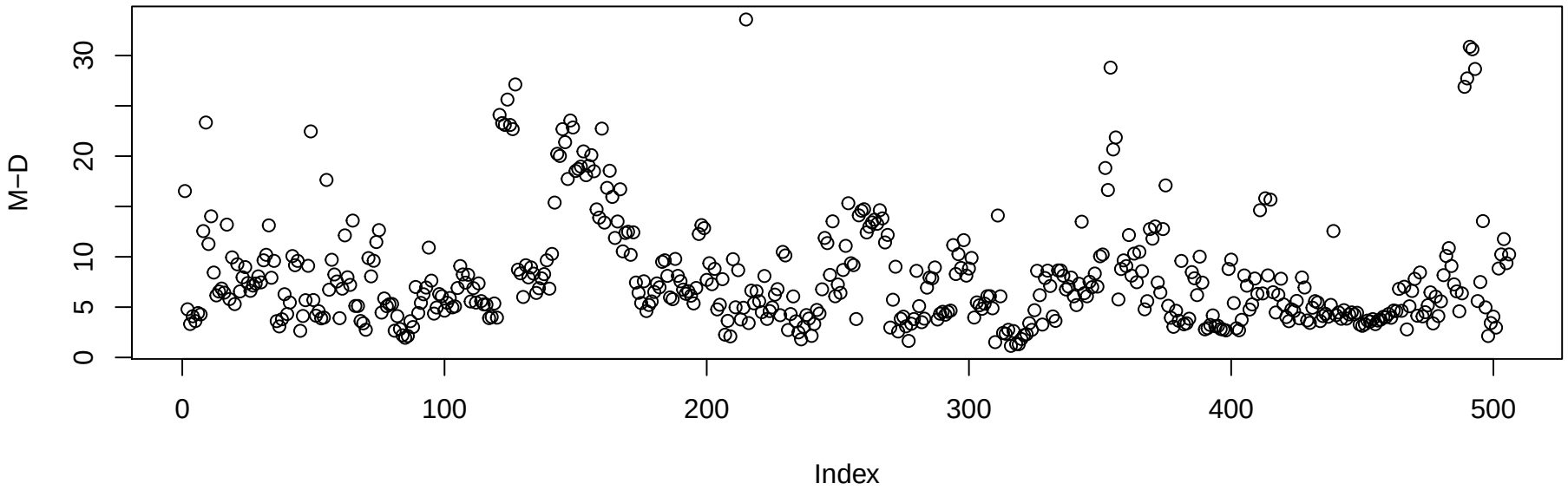
Primer

```
residualPlots(modelBoston,type='response',tests=FALSE)
```



Primer

```
modelBoston.1<-lm(medv~lstat+dis+age+ptratio+indus+nox+tax+log(crim))
xBoston.1=cbind(lstat,dis,age,ptratio,indus,nox,tax,log(crim))
plot(mahalanobis(xBoston.1,center=colMeans(xBoston.1),cov=cov(xBoston.1)),ylab='M-D')
```

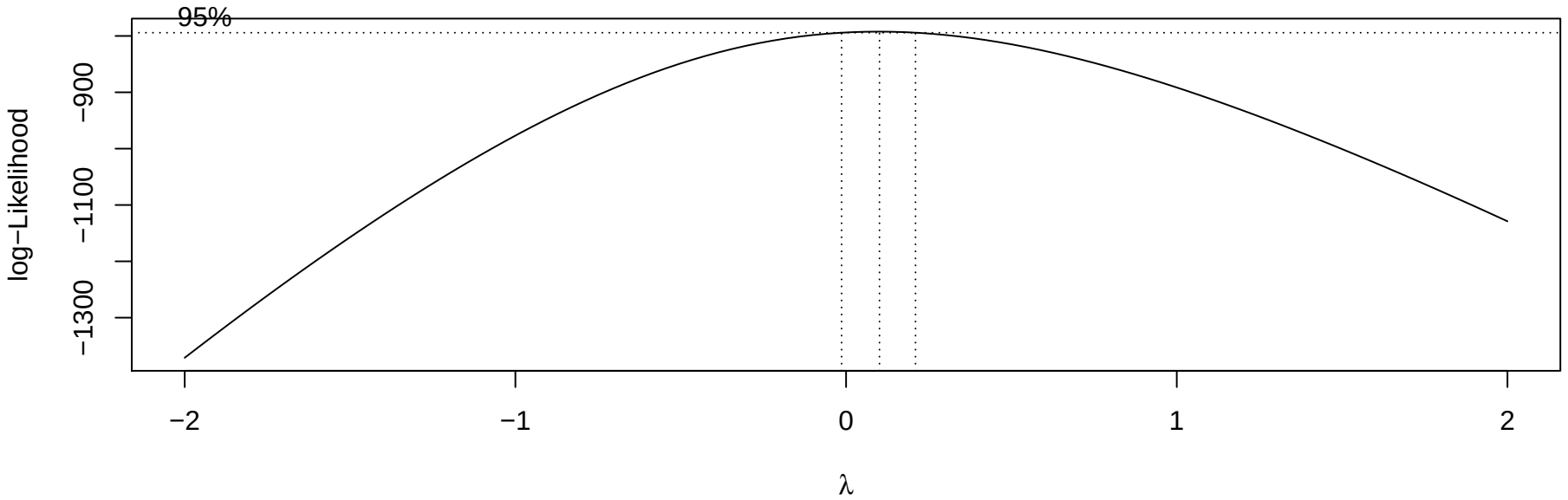



```
which.max(mahalanobis(xBoston.1,center=colMeans(xBoston.1),cov=cov(xBoston.1)))
```

```
## [1] 215
```

Primer

```
boxcox(modelBoston.1)
```



Primer

```
library(MASS)
modelBoston.2<-lm(log(medv)~lstat+dis+age+ptratio+indus+nox+tax+log(crim))
outlierTest(modelBoston.2)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 406 -4.403043         1.3083e-05    0.0066199
```

```
hatvalues(modelBoston.2)[406]
```

```
##         406
## 0.01609512
```

Primer

```
summary(modelBoston.2)
```

```
##
## Call:
## lm(formula = log(medv) ~ lstat + dis + age + ptratio + indus +
##     nox + tax + log(crim))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92349 -0.12514 -0.01653  0.11205  0.80806
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.9790906  0.1608763  30.950  < 2e-16 ***
## lstat       -0.0395664  0.0018998 -20.827  < 2e-16 ***
## dis         -0.0433298  0.0082839  -5.231 2.49e-07 ***
## age          0.0008015  0.0005842   1.372   0.1707
## ptratio     -0.0434987  0.0053673  -8.104 4.15e-15 ***
## indus       -0.0010252  0.0026144  -0.392   0.6951
## nox         -0.7695094  0.1780660  -4.321 1.87e-05 ***
## tax         -0.0002192  0.0001132  -1.936   0.0534 .
## log(crim)    0.0050233  0.0101292   0.496   0.6202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2153 on 497 degrees of freedom
## Multiple R-squared:  0.7269, Adjusted R-squared:  0.7225
## F-statistic: 165.4 on 8 and 497 DF,  p-value: < 2.2e-16
```

Primer

```
summary(modelBoston.1)
```

```
##
## Call:
## lm(formula = medv ~ lstat + dis + age + ptratio + indus + nox +
##     tax + log(crim))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9556  -3.0984  -0.7736   2.2555  24.4776
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  75.244672   4.057249  18.546  < 2e-16 ***
## lstat        -0.859267   0.047912 -17.934  < 2e-16 ***
## dis          -1.320330   0.208918  -6.320 5.83e-10 ***
## age           0.018894   0.014734   1.282   0.2003
## ptratio      -1.274795   0.135362  -9.418  < 2e-16 ***
## indus        -0.123492   0.065935  -1.873   0.0617 .
## nox         -21.965172   4.490768  -4.891 1.36e-06 ***
## tax          -0.001380   0.002855  -0.483   0.6290
## log(crim)     0.608186   0.255456   2.381   0.0177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.43 on 497 degrees of freedom
## Multiple R-squared:  0.6569, Adjusted R-squared:  0.6514
## F-statistic: 118.9 on 8 and 497 DF,  p-value: < 2.2e-16
```

Primer

```
modelBoston.3<-lm(log(medv)~lstat+dis+ptratio+nox)
anova(modelBoston.2,modelBoston.3)
```

```
## Analysis of Variance Table
##
## Model 1: log(medv) ~ lstat + dis + age + ptratio + indus + nox + tax +
##     log(crim)
## Model 2: log(medv) ~ lstat + dis + ptratio + nox
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    497 23.043
## 2    501 23.430 -4  -0.38646 2.0838 0.08175 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

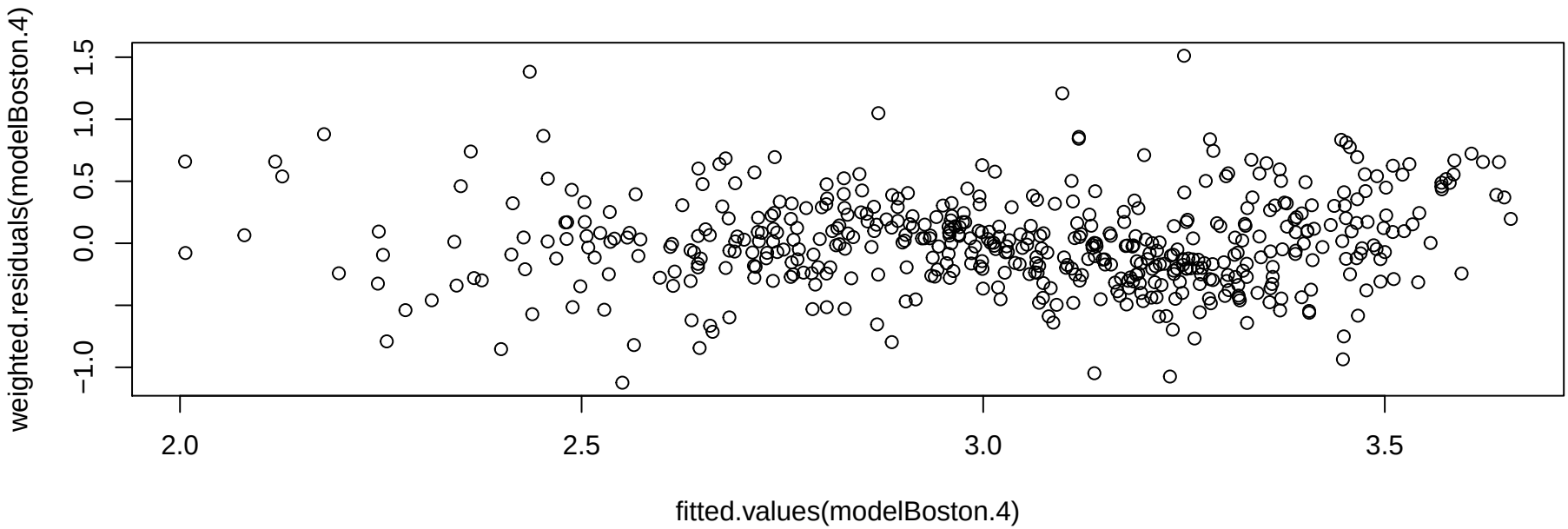
Primer

```
modelBoston.3<-lm(log(medv)~lstat+dis+ptratio+nox)
anova(modelBoston.2,modelBoston.3)
```

```
## Analysis of Variance Table
##
## Model 1: log(medv) ~ lstat + dis + age + ptratio + indus + nox + tax +
##     log(crim)
## Model 2: log(medv) ~ lstat + dis + ptratio + nox
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    497 23.043
## 2    501 23.430 -4  -0.38646 2.0838 0.08175 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

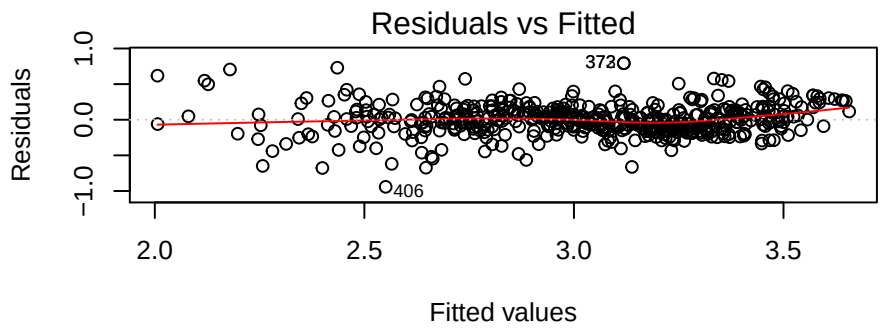
Primer

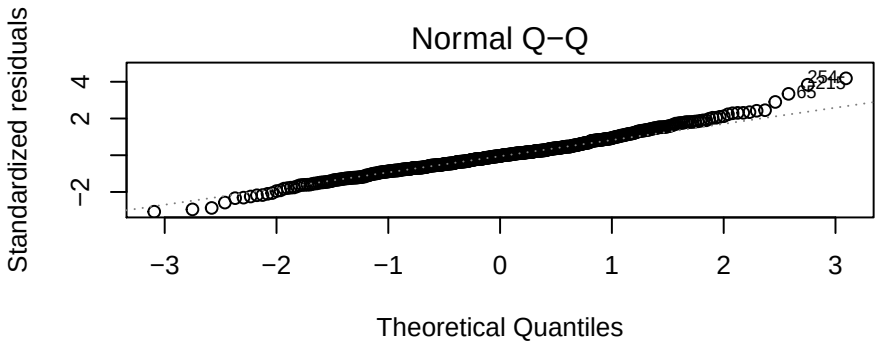
```
modelBoston.4<-lm(log(medv)~lstat+dis+ptratio+nox,weights = dis)
plot(fitted.values(modelBoston.4),weighted.residuals(modelBoston.4))
```

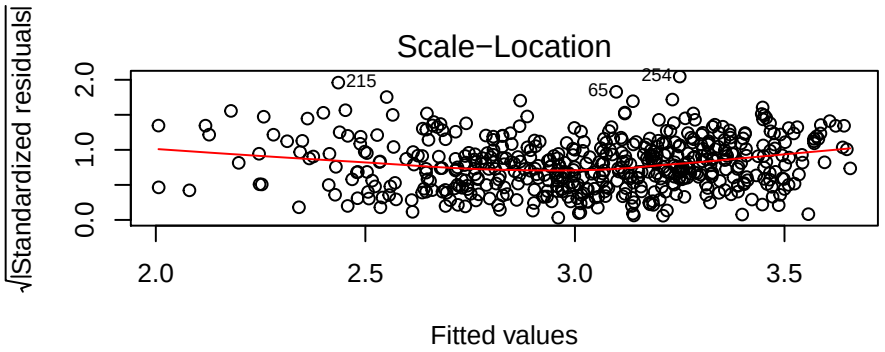


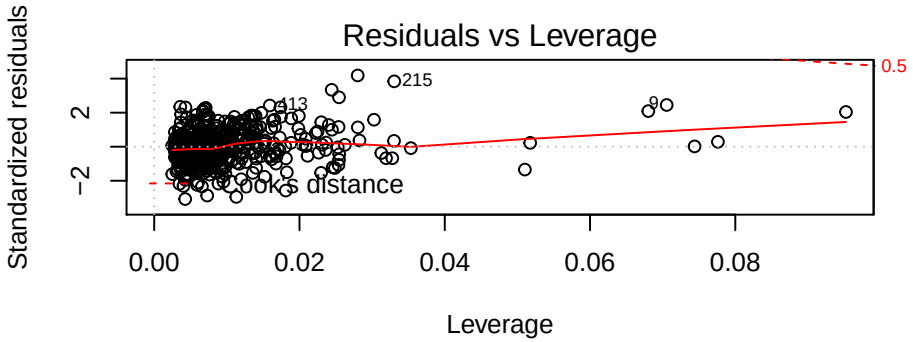
Primer

```
par(mfrow=c(2,2))
plot(modelBoston.4)
```



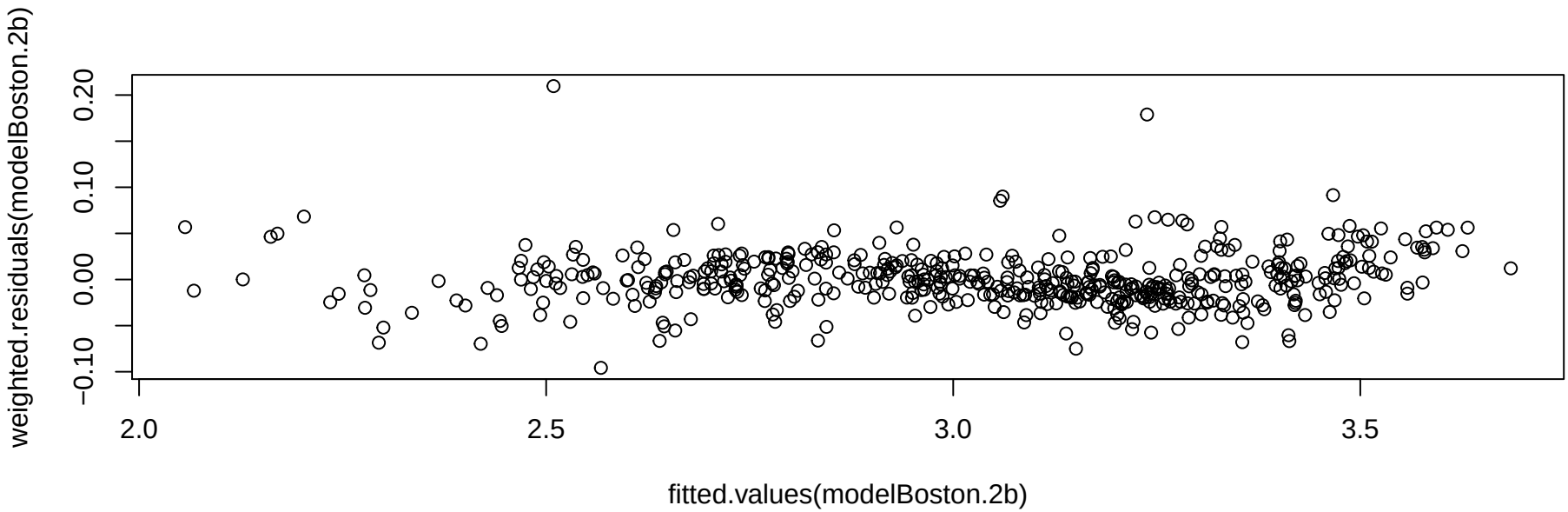






Primer

```
modelBoston.2b<-lm(log(medv)~lstat+dis+age+ptratio+indus+nox+tax+log(crim),weight=1/age)
plot(fitted.values(modelBoston.2b),weighted.residuals(modelBoston.2b))
```



Primer

```
outlierTest(modelBoston.2b)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 215 9.101259         2.1744e-18   1.1002e-15
## 254 6.736077         4.5208e-11   2.2875e-08
```

```
hatvalues(modelBoston.2b)[c(215,254)]
```

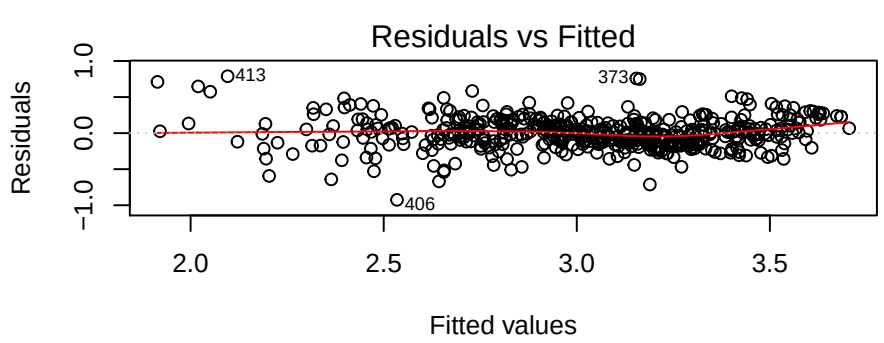
```
##       215       254
## 0.2850254 0.1113287
```

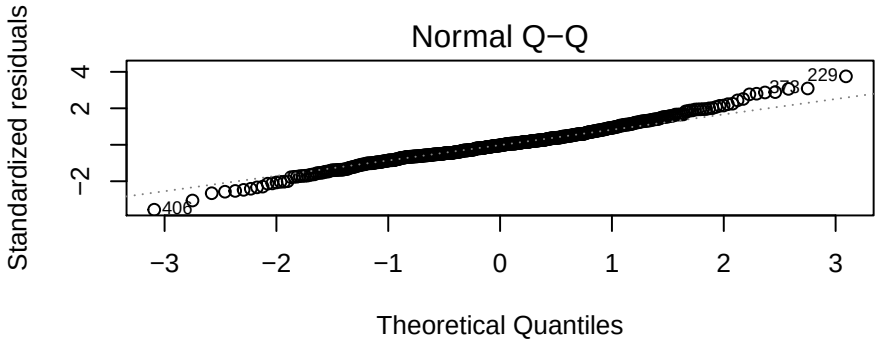
```
modelBoston.3b<-lm(log(medv)~lstat+dis+age+ptratio+indus+nox+tax+log(crim),weight=1/age,data=Boston[-c(215,254)
modelBoston.4b<-lm(log(medv)~lstat+dis+age+ptratio+nox+indus,weight=1/age,data=Boston[-c(215,254),])
anova(modelBoston.4b,modelBoston.3b)
```

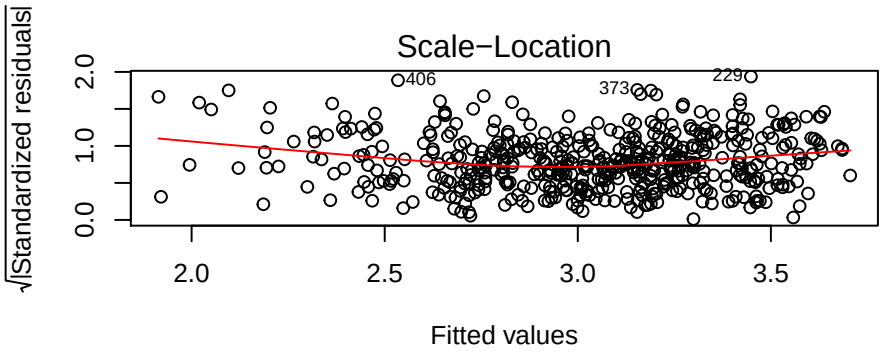
```
## Analysis of Variance Table
##
## Model 1: log(medv) ~ lstat + dis + age + ptratio + nox + indus
## Model 2: log(medv) ~ lstat + dis + age + ptratio + indus + nox + tax +
##     log(crim)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    497 0.33734
## 2    495 0.33571  2 0.0016328 1.2038 0.3009
```

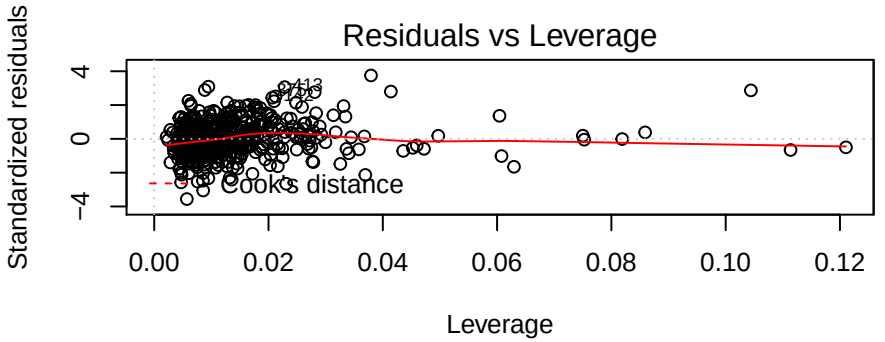
Primer

```
par(mfrow=c(2,2))
plot(modelBoston.4b)
```









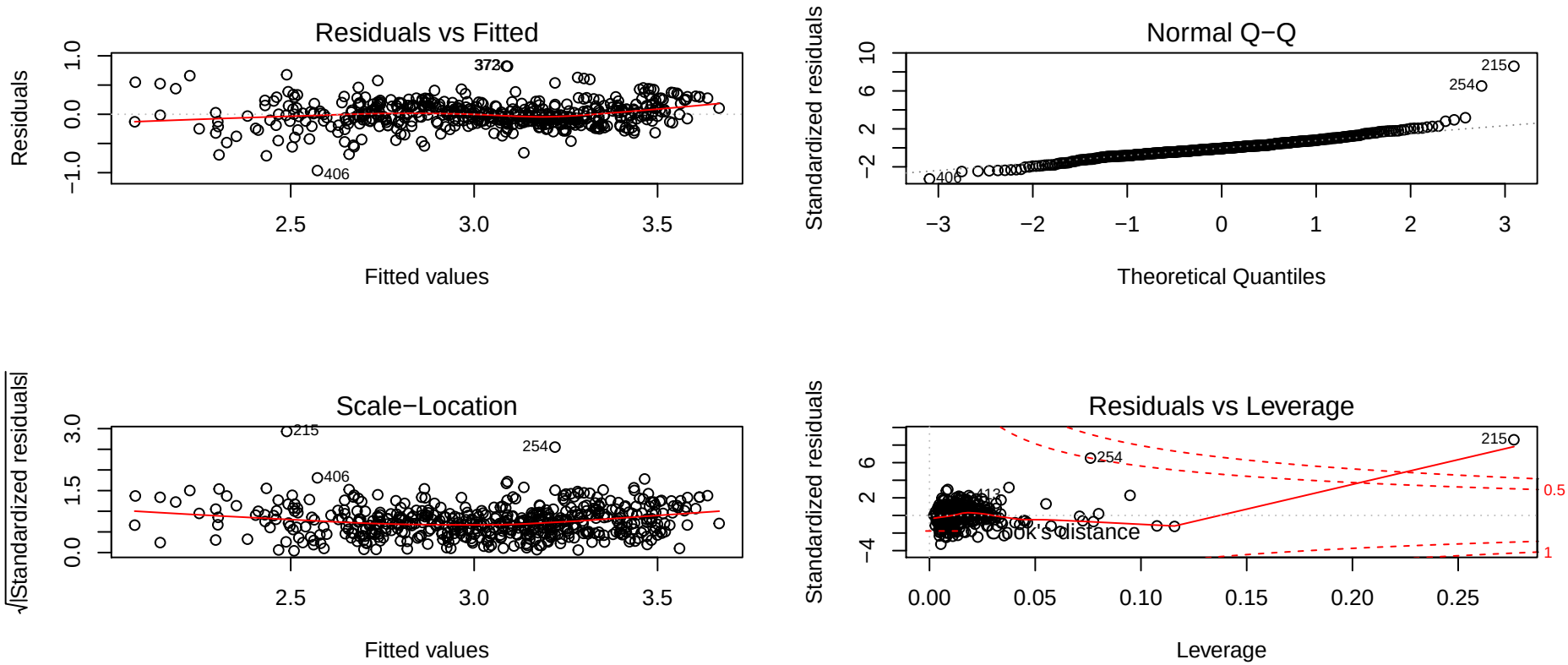
Primer

```
modelBoston.4bA<-lm(log(medv)~lstat+dis+age+ptratio+nox+indus,weight=1/age)
summary(modelBoston.4bA)
```

```
##
## Call:
## lm(formula = log(medv) ~ lstat + dis + age + ptratio + nox +
##     indus, weights = 1/age)
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.096324 -0.016290 -0.001886  0.014845  0.216095
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.784e+00  1.287e-01  37.168  < 2e-16 ***
## lstat       -3.487e-02  2.008e-03 -17.366  < 2e-16 ***
## dis         -4.204e-02  6.820e-03  -6.165 1.46e-09 ***
## age          1.326e-05  4.712e-04   0.028  0.97757
## ptratio     -4.064e-02  5.104e-03  -7.964 1.14e-14 ***
## nox         -5.701e-01  1.796e-01  -3.174  0.00160 **
## indus       -7.478e-03  2.637e-03  -2.835  0.00476 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02953 on 499 degrees of freedom
## Multiple R-squared:  0.6719, Adjusted R-squared:  0.668
## F-statistic: 170.3 on 6 and 499 DF,  p-value: < 2.2e-16
```

Primer

```
par(mfrow=c(2,2))
plot(modelBoston.4bA)
```



Primer

```
modelBoston.4bb<-lm(log(medv)~lstat+dis+ptratio+nox+indus,weight=1/age)
summary(modelBoston.4bb)
```

```
##
## Call:
## lm(formula = log(medv) ~ lstat + dis + ptratio + nox + indus,
##     weights = 1/age)
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.096324 -0.016274 -0.001847  0.014866  0.215837
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.783543   0.127909  37.398  < 2e-16 ***
## lstat       -0.034851   0.001929 -18.065  < 2e-16 ***
## dis         -0.042095   0.006547  -6.429 3.00e-10 ***
## ptratio     -0.040639   0.005096  -7.974 1.05e-14 ***
## nox         -0.567845   0.161201  -3.523 0.000467 ***
## indus       -0.007490   0.002603  -2.878 0.004180 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0295 on 500 degrees of freedom
## Multiple R-squared:  0.6719, Adjusted R-squared:  0.6686
## F-statistic: 204.8 on 5 and 500 DF,  p-value: < 2.2e-16
```

Primer

```
par(mfrow=c(2,2))
 plot(fitted.values(modelBoston.4bA),medv)
 plot(fitted.values(modelBoston.4bb),medv)
plot(fitted.values(modelBoston.4b),medv[-c(215,254)])
```