Zápočet

Kristína Blašková

December 2017

Q01

Zjistete, zdali data neobsahují chybějící hodnoty (NA), pokud ano tak příslušná pozorování z dat odstrante. Ověřte rozmery datového souboru a shrňte základní popisné charakteristiky všech promenných.

1. V dátovom súbore nie sú žiadne na hodnoty.

```
library(MASS)
any(is.na(Boston))
```

```
## [1] FALSE
```

2. Rozmery sú 506 riadkov a 14 stĺpcov.

```
dim(Boston)
```

```
## [1] 506  14
```

3. Základné popisné charakteristiky sú vo výstupe nižšie:

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

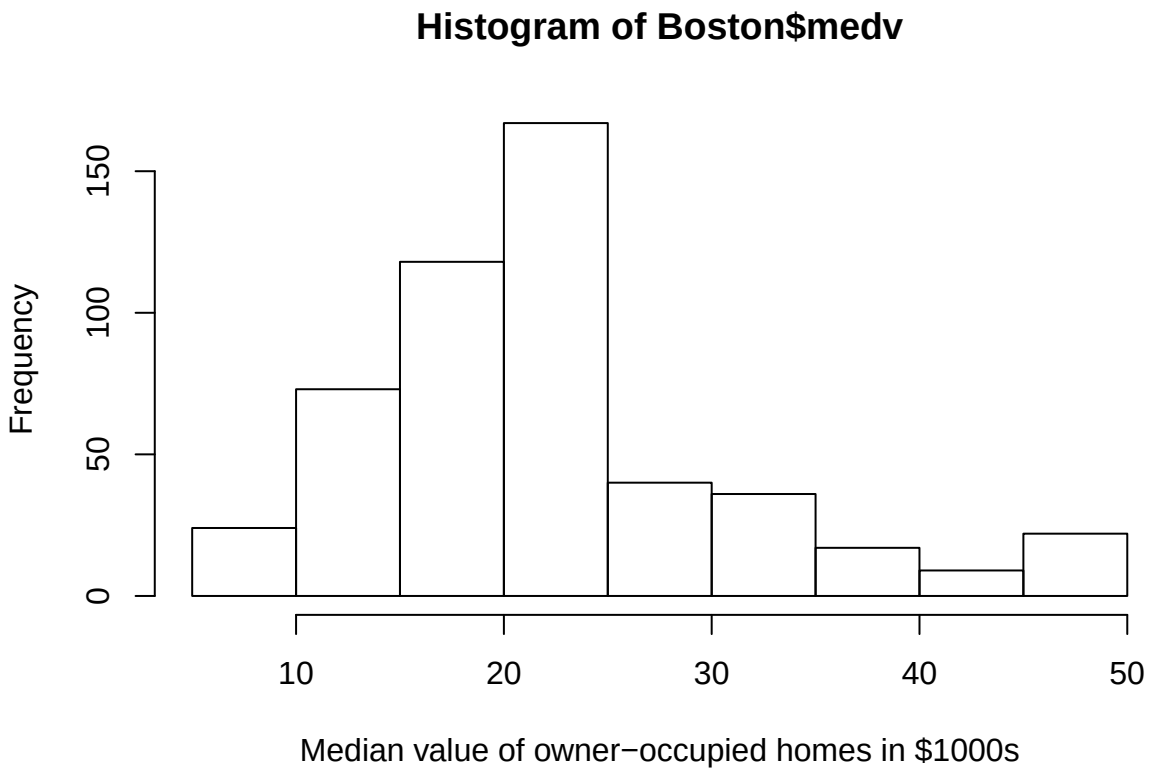
```r
str(Boston)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Q2

Vykreslete histogram a odhad hustoty pro odezvu medv.

```
medlab = "Median value of owner-occupied homes in $1000s"
 hist(Boston$medv, breaks = 15, xlab = medlab)
```

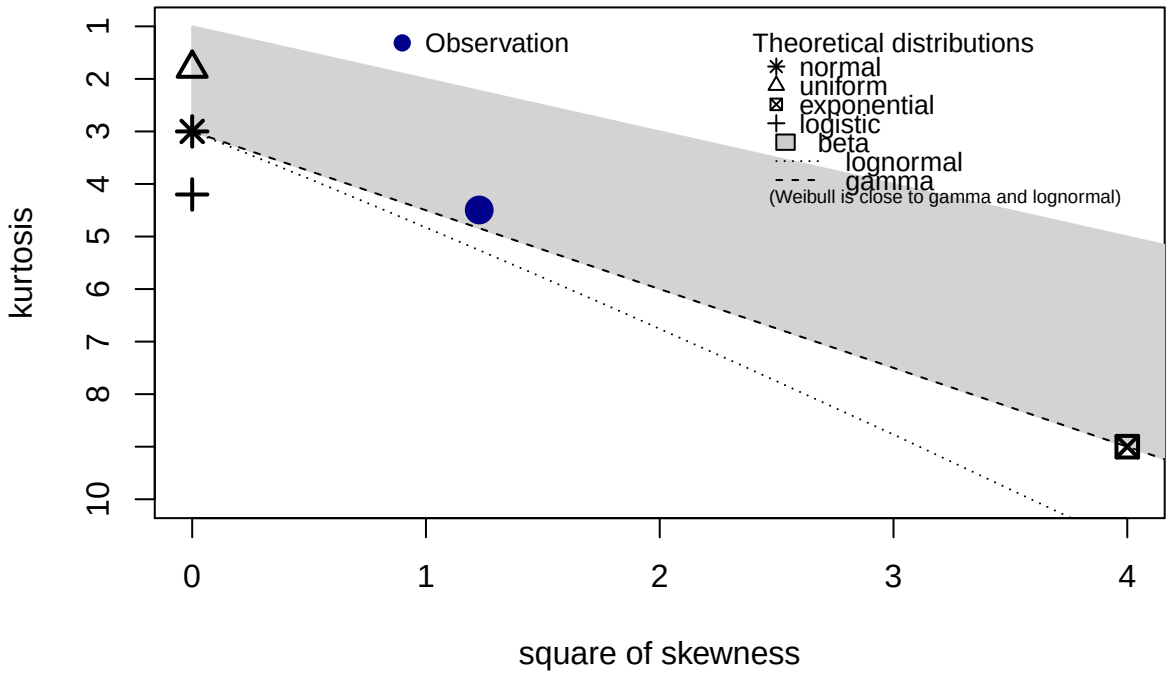


Pre odhad hustoty použijeme balíček fitdistrplus, ktorý obsahuje funkciu descdist, ktorá nám na základe špicatosti a šikmosti nájde najvhodnejšiu distribúciu pre premennú medv.

```
library(fitdistrplus)
```

```
## Loading required package: survival
```

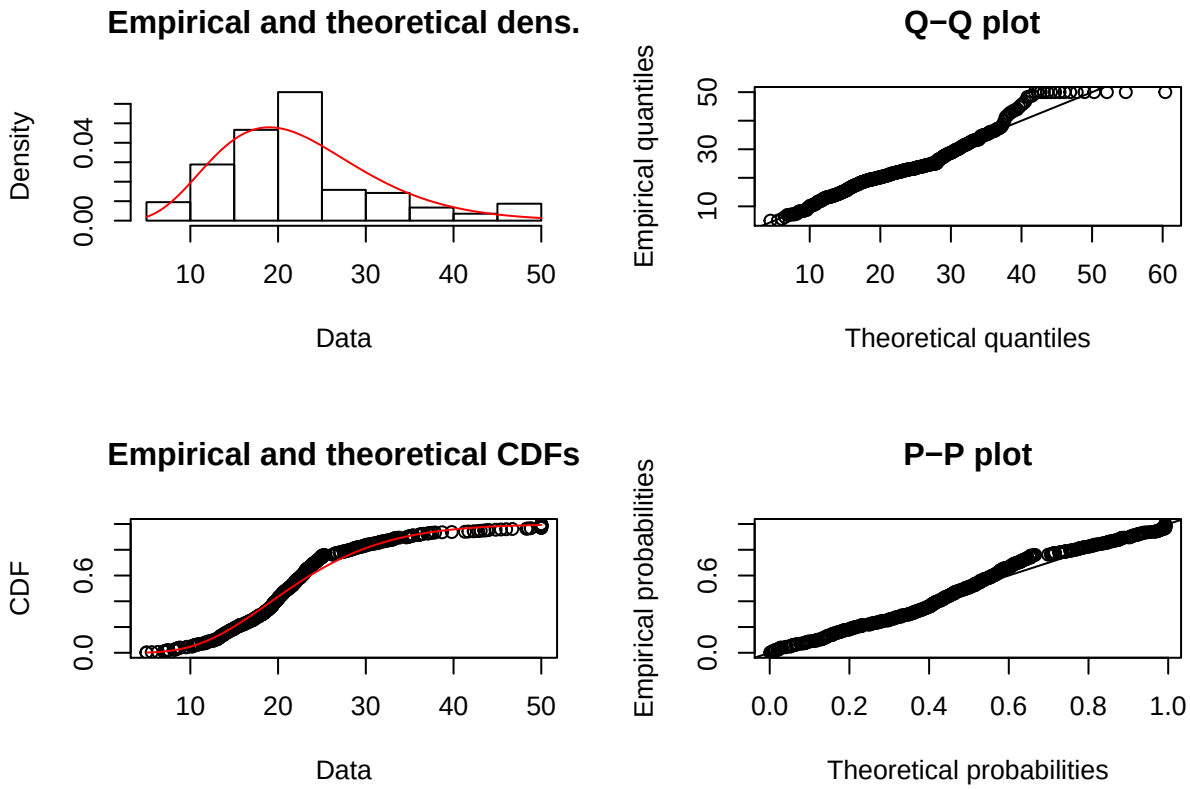
```
descdist(Boston$medv, discrete = FALSE)
```




```
## summary statistics
## ------
## min:  5   max:  50
## median:  21.2
## mean:  22.53281
## estimated sd:  9.197104
## estimated skewness:  1.108098
## estimated kurtosis:  4.495197
```

Jedno z vhodných rozdelení je aj gamma, ktoré použijeme:

```
medv <- Boston$medv
dist <- fitdistrplus::fitdist(medv, "gamma")
plot(dist)
```

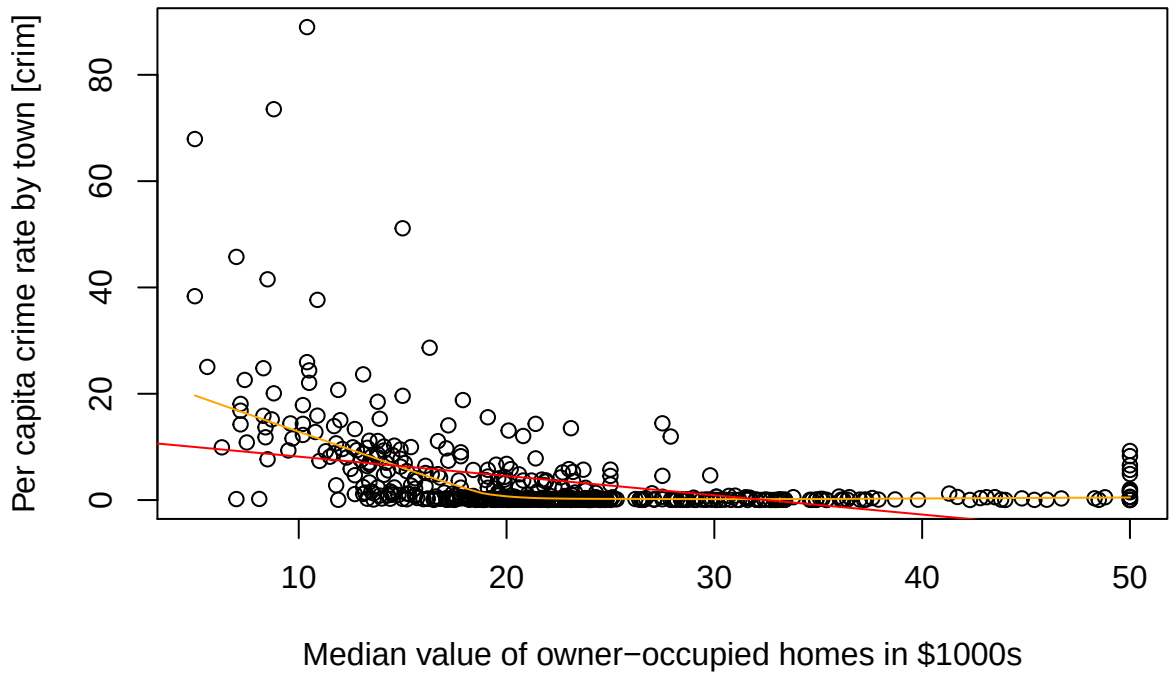


Z grafu hustoty vidíme, že pri hodnote medv 50 sa dáta najviac nezhodujú s preloženou gamma distribúciou. To nám potvrdzuje aj Q-Q plot.

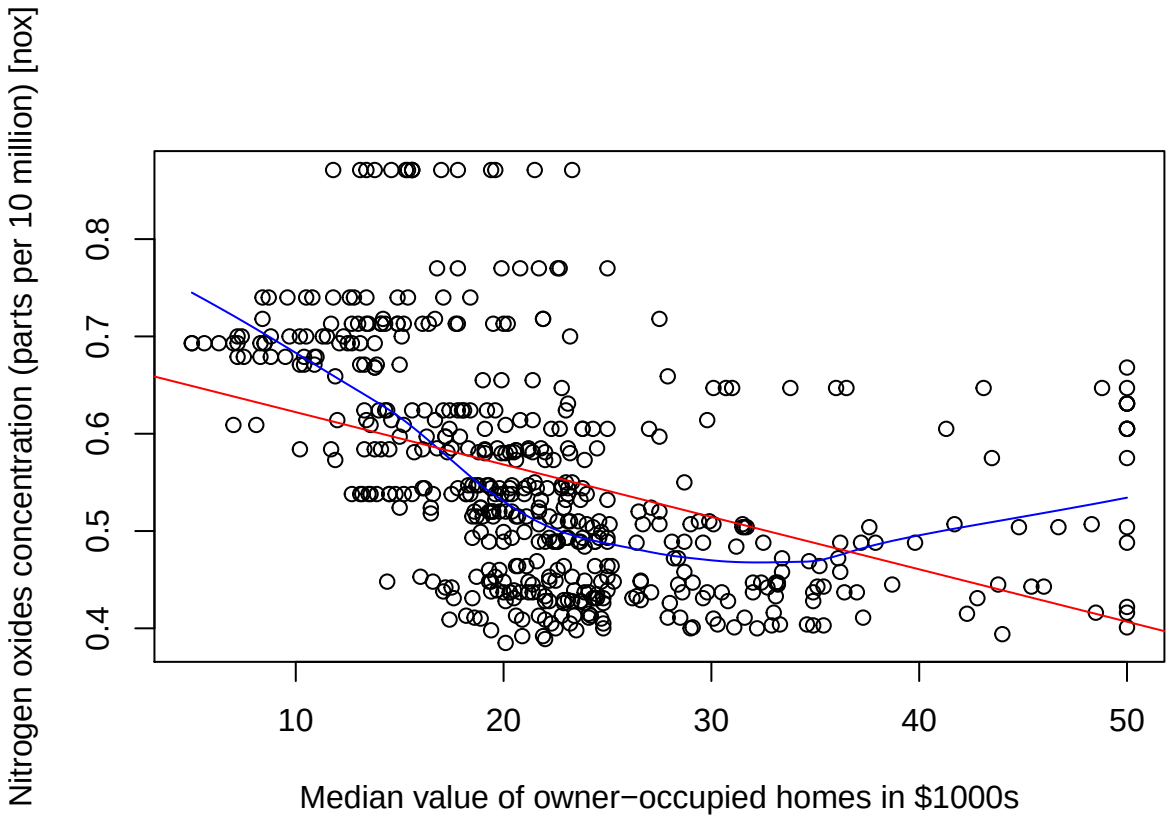
Q03

Pro promenné crim, nox, rm, lstat, ptratio, dis vykreslete scatterplot - závislost dané promenné na odezvě a proložte body jak lineárním odhadem tak vyhlazenou křivkou (lines(lowess(X,Y))).

```
library(car)

{plot(Boston$medv, Boston$crim, xlab = medlab ,
      ylab = "Per capita crime rate by town [crim]")
lines(lowess(Boston$medv,Boston$crim), col="orange")
abline(lm(Boston$crim~Boston$medv), col="red")}
```

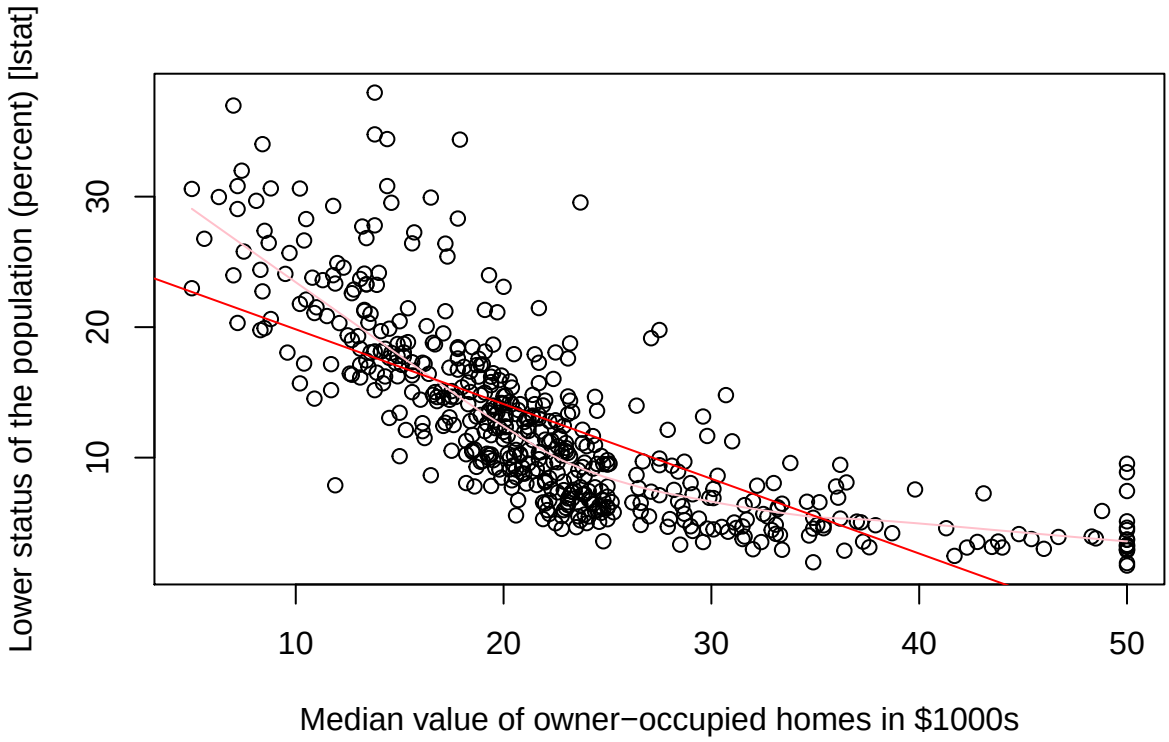



```
{plot(Boston$medv, Boston$nox, xlab = medlab ,
      ylab = "Nitrogen oxides concentration (parts per 10 million) [nox]")
lines(lowess(Boston$medv,Boston$nox), col="blue")
abline(lm(Boston$nox~Boston$medv), col="red")}
```

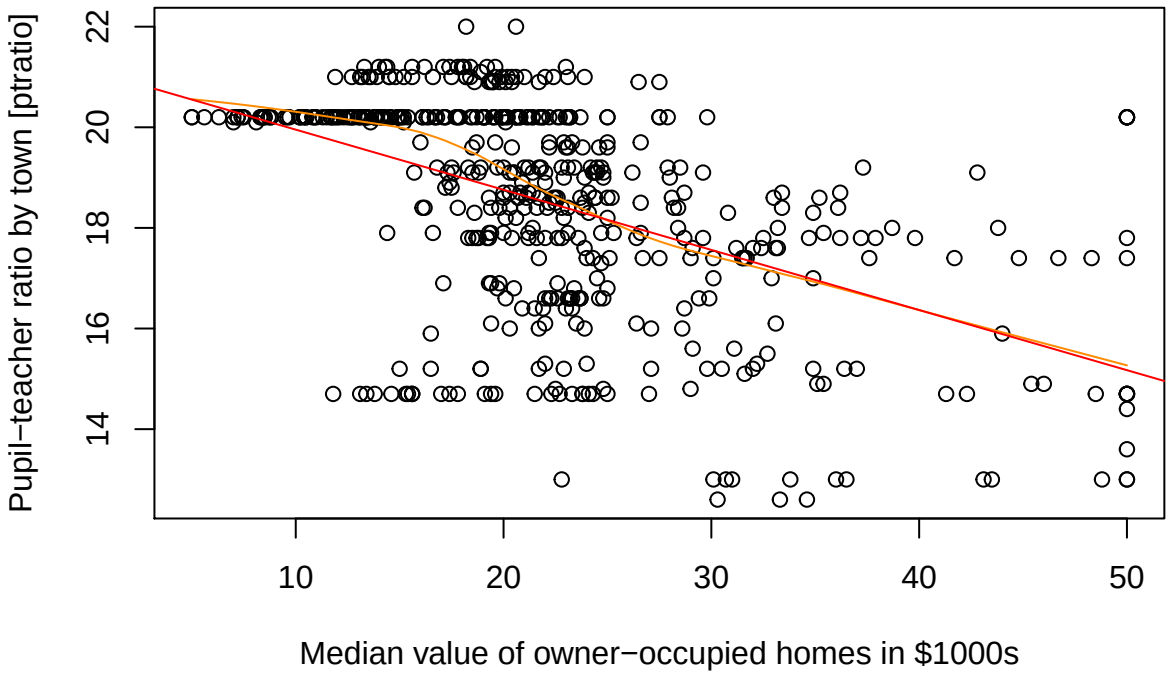



```
{plot(Boston$medv, Boston$rm, xlab = medlab ,
      ylab = "Average number of rooms per dwelling [rm]")
lines(lowess(Boston$medv,Boston$rm), col="yellow")
abline(lm(Boston$rm~Boston$medv), col="red")}
```

```
{plot(Boston$medv, Boston$lstat, xlab = medlab ,
      ylab = "Lower status of the population (percent) [lstat]")
lines(lowess(Boston$medv,Boston$lstat), col="pink")
abline(lm(Boston$lstat~Boston$medv), col="red")}
```




```
{plot(Boston$medv, Boston$ptratio, xlab = medlab ,
      ylab = "Pupil-teacher ratio by town [ptratio]")
lines(lowess(Boston$medv,Boston$ptratio), col="darkorange")
abline(lm(Boston$ptratio~Boston$medv), col="red")}
```

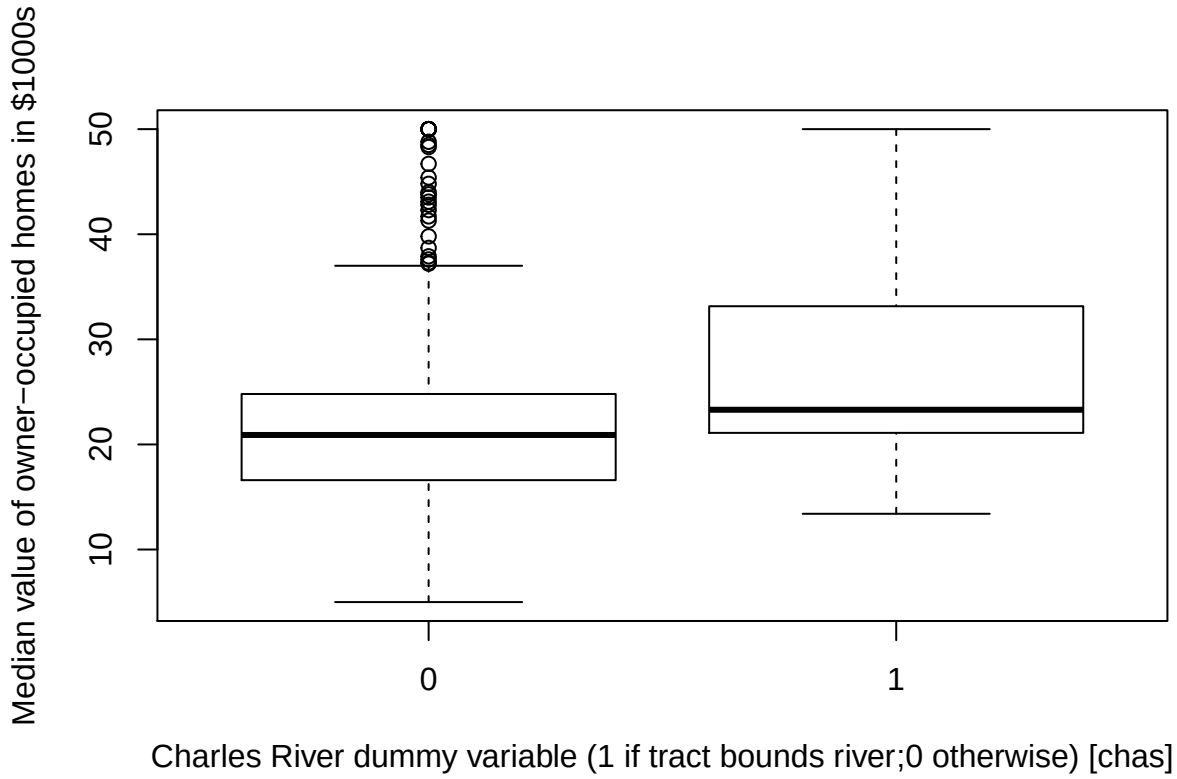



```
{plot(Boston$medv, Boston$dis, xlab = medlab ,
      ylab = "Weighted mean of distances to five Boston employment centres [dis]")
  abline(lm(Boston$dis~Boston$medv), col="red")
lines(lowess(Boston$medv,Boston$dis), col="green")}
```

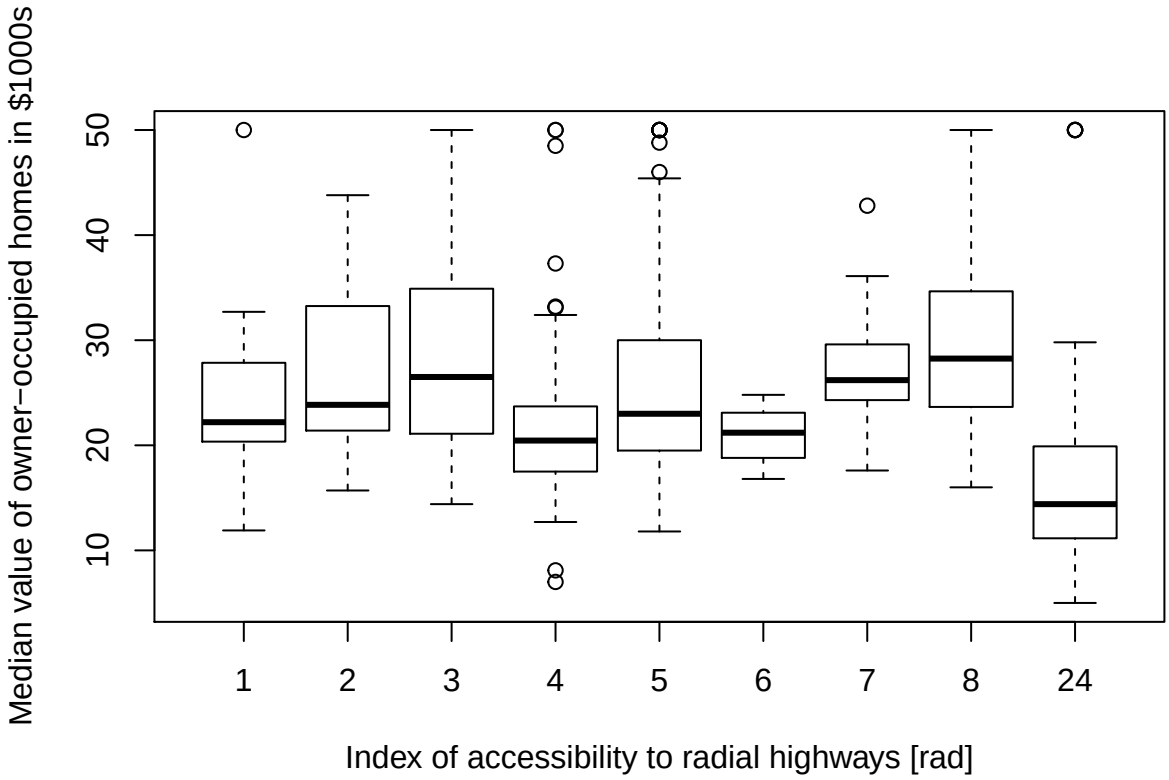
Q4

Pro proměnné chas a rad a jejich vztah k odezvě vykreslete krabicové diagramy (boxploty). Proměnnou rad transformujte tak, aby obsahovala pouze dvě úrovně (levely) a vykreslete opět krabicový diagram.

```
boxplot(medv~chas,data=Boston, ylab = medlab,
        xlab = "Charles River dummy variable (1 if tract bounds river;0 otherwise) [chas]")
```




```
boxplot(medv~rad,data=Boston, ylab = medlab,
        xlab = "Index of accessibility to radial highways [rad]")
```



Z obrázku vidíme, že máme v súčastnosti 9 levelov. Vzhľadom na význam premennej, dáva zmysel zlúčiť levely 1-8 a ponechať level 24. To spravíme tak, že pomocou ifelse vytvoríme nový stĺpec a priradíme mu hodnotu 24, ak v stĺpci rad v rovnakom riadku tiež vidíme hodnotu 24, a hodnotu 8 inak. ifelse pôsobí na vektory ako celok a preto nepotrebujeme žiaden for cyklus.

```
bostoncopy <- Boston
bostoncopy$radnew <- ifelse(bostoncopy$rad ==24 , 24, 8)
```

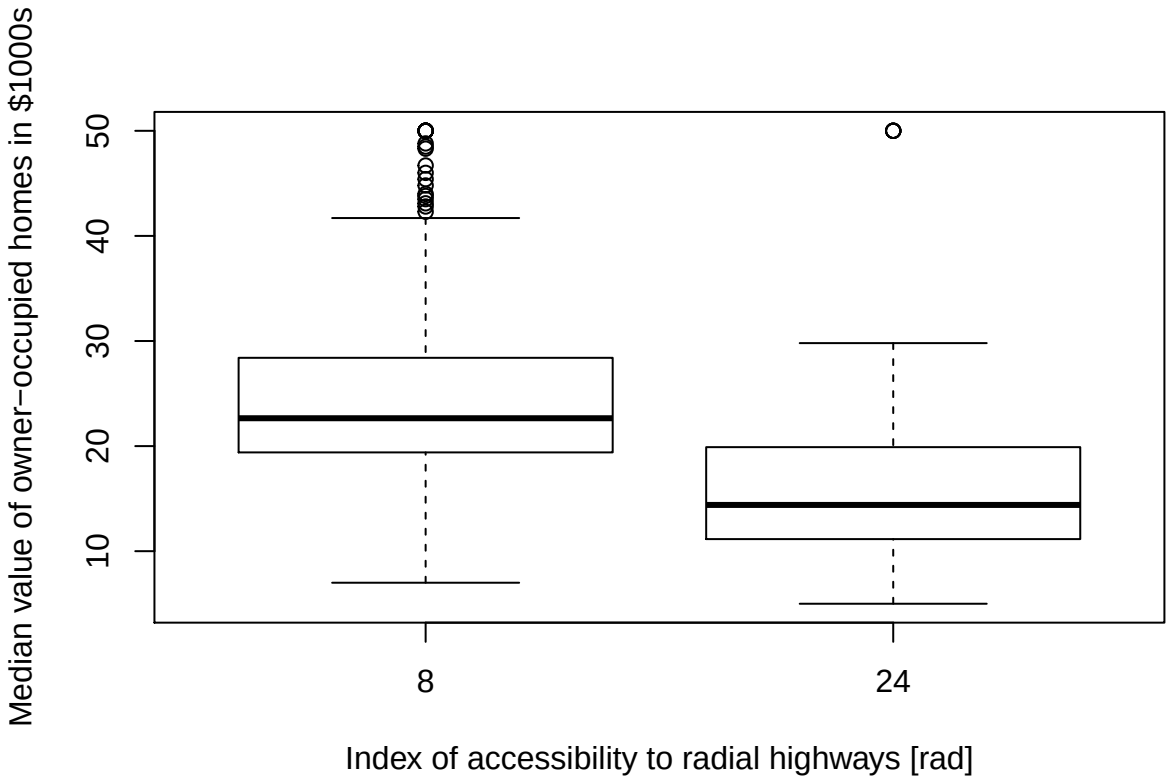
Následne môžeme pristúpiť k novému boxplotu:

```
factor(bostoncopy$radnew)
```

```
##   [1] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
##  [24] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
##  [47] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
##  [70] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
##  [93] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [116] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [139] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [162] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [185] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [208] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [231] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [254] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [277] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [300] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [323] 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## [346] 8  8  8  8  8  8  8  8  8  8  8  24 24 24 24 24 24 24 24 24 24 24 24
## [369] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
## [392] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
```

```
## [415] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
## [438] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
## [461] 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24 24
## [484] 24 24 24 24 24 8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
## Levels: 8 24
```

```
boxplot(medv~radnew,data=bostoncopy, ylab = medlab,
        xlab = "Index of accessibility to radial highways [rad]")
```



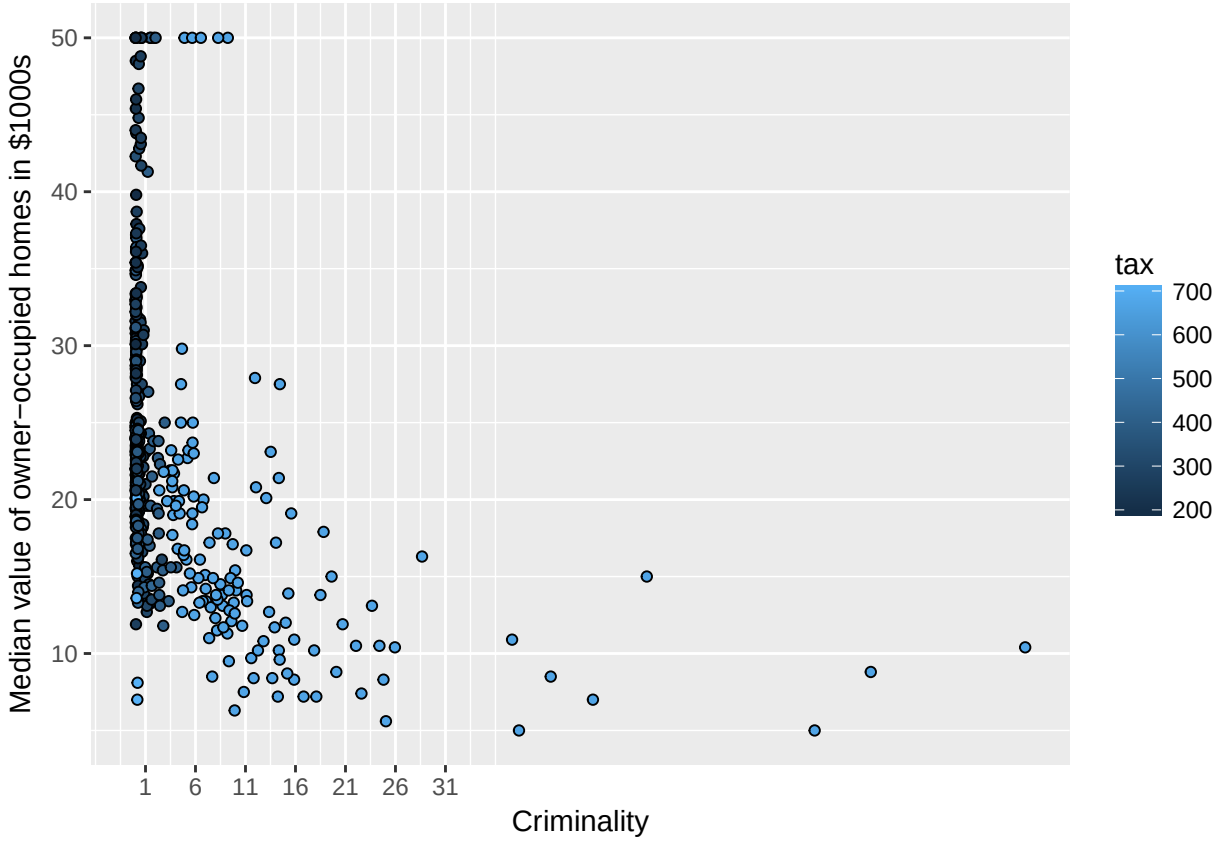
Q05

Navrhnete další zobrazení datového souboru. Proveďte ho a popište jeho účel.

Našim cieľom (podľa zadania) je preskúmať ceny nehnuteľností. Využiť môžeme aj bubble charts, vďaka ktorým si môžme všimnúť závislosti medzi 3 premennými a zároveň nám to ukáže, či nie je vhodné nejaké premenné rozfaktorovať.

Ako prvé si vykreslíme závislosť ceny domov od kriminality a pozrieme sa, aké dane platia ľudia za tieto nehnuteľnosti:

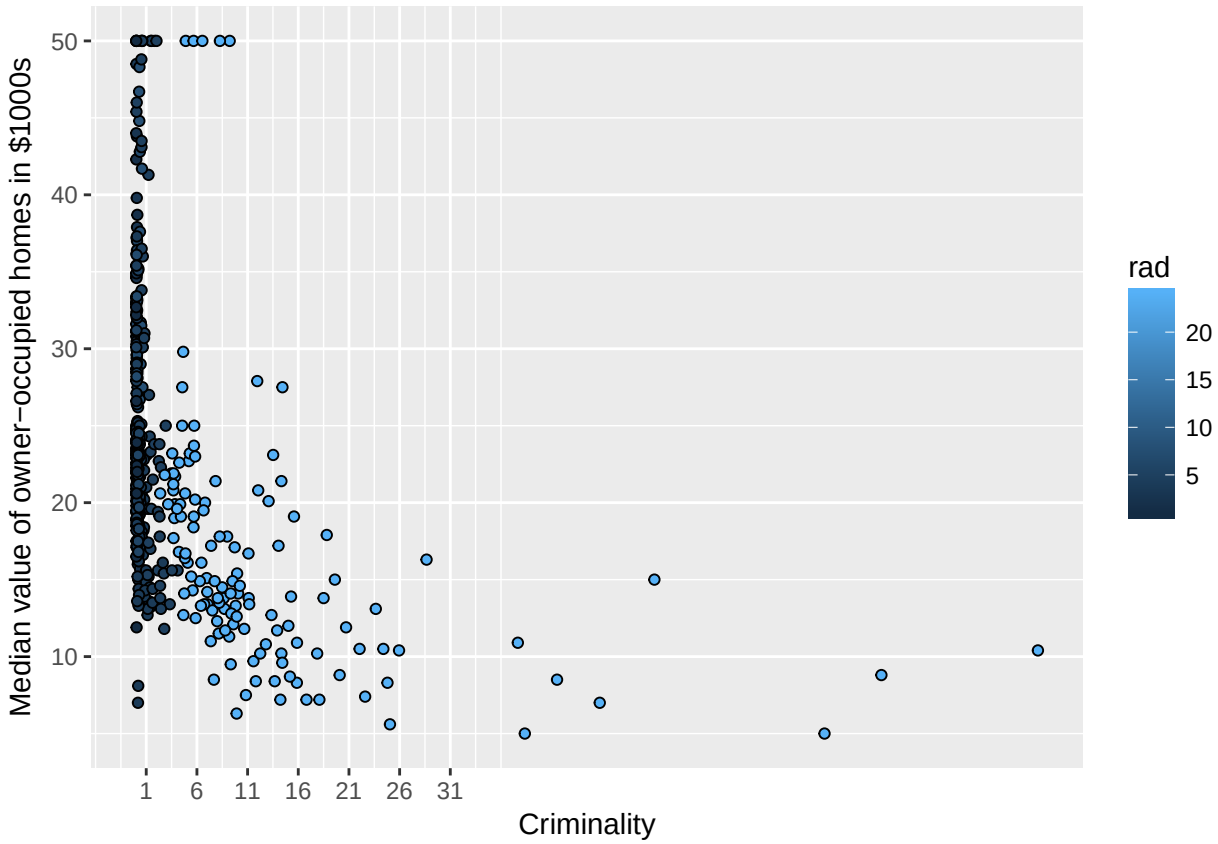
```
library(ggplot2)

ggplot(Boston, aes(x = crim, y = medv, fill = tax)) +
       geom_point(shape = 21) +
  labs(x = "Criminality", y = medlab) +
       scale_x_continuous(breaks = seq(1, 31, 5))
```



Čím je vyššia kriminalita, tým menej nehnuteľnosť stojí, ale zároveň sa platia vyššie dane. To môže byť spôsobené tým, že na miestach s vyššou kriminalitou sú zásahy zdravotníkov, policajtov, alebo hasičov oveľa potrebnejšie. Taktiež sa pravdepodobne oveľa častejšie zničí verejný majetok a tieto dane môžu slúžiť ako financovanie týchto dôsledkov.

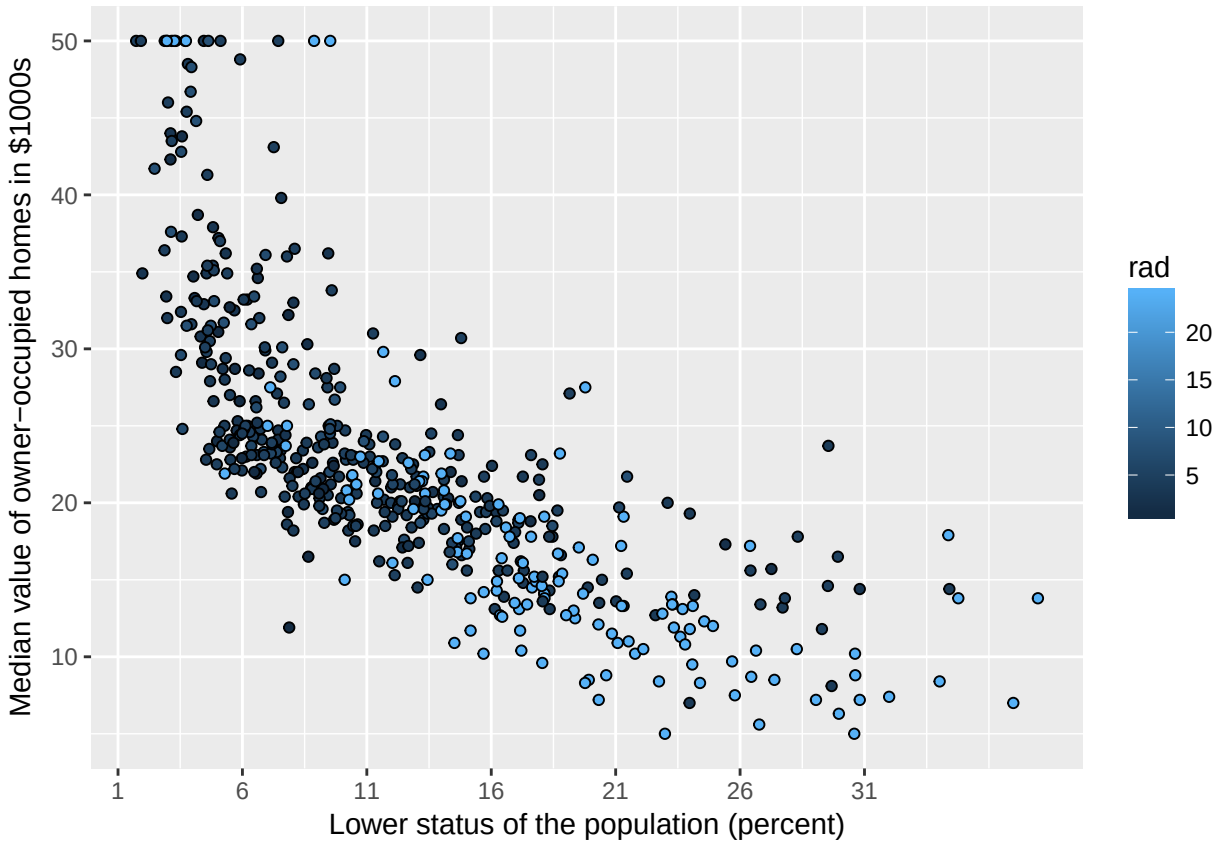
Ďalej sa pozrieme, ako súvisí cena nehnuteľností, kriminalita a vzdialenosť od obchvatov.

```
ggplot(Boston, aes(x = crim, y = medv, fill = rad)) +
       geom_point(shape = 21) +
  labs(x = "Criminality", y = medlab) +
       scale_x_continuous(breaks = seq(1, 31, 5))
```



Čím vyššia je kriminalita, tým nižšie sú ceny nehnuteľností a tým vačšia je nižšia dostupnosť na diaľnicu. Teraz skúsime vymeniť kriminalitu za počet ľudí s nižším sociálnym statusom:

```
ggplot(Boston, aes(x = lstat, y = medv, fill = rad)) +
       geom_point(shape = 21) +
  labs(x = "Lower status of the population (percent)", y = medlab) +
        scale_x_continuous(breaks = seq(1, 31, 5))
```



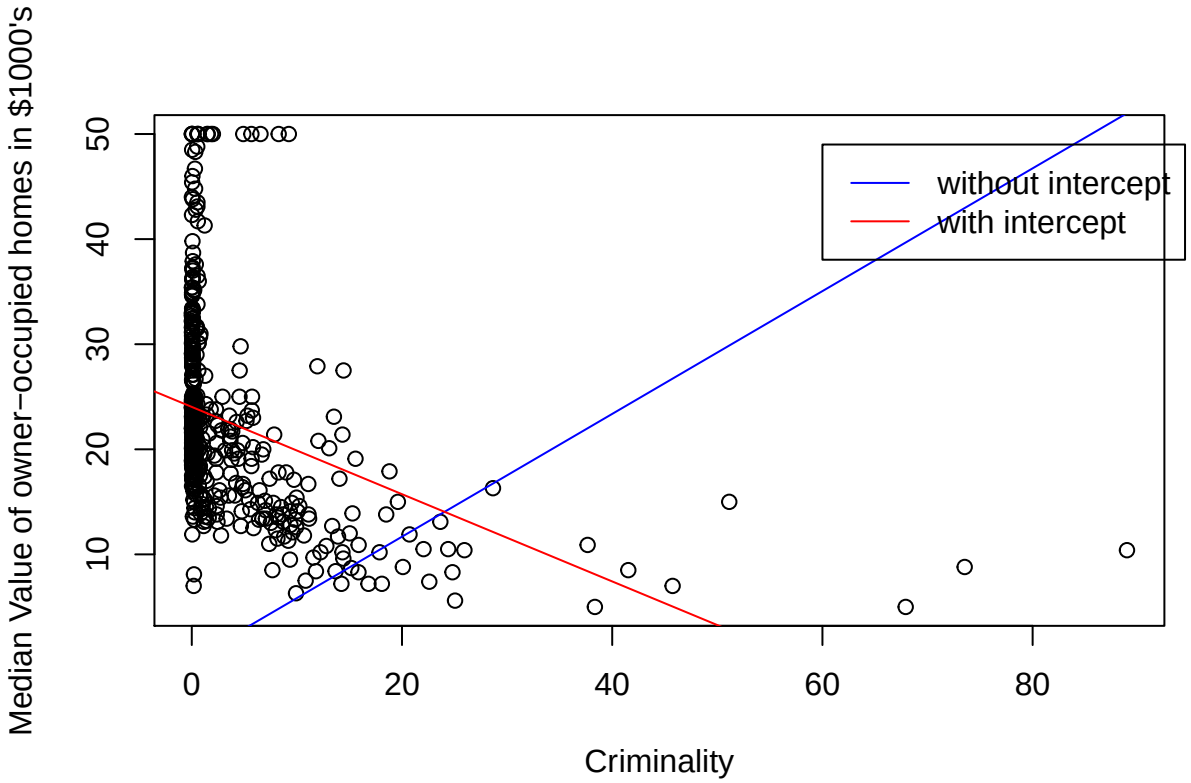
Vidíme, že väčšina ľudí s nižším sociálnym statusom to má ďaleko na diaľnicu, to ale platilo aj pre miesta s vyššou kriminalitou. Z toho vyplýva, že množina ľudí s nižším statusom sa vysoko prekrýva s množinou kriminálnikov.

Q06

Sestavte jednoduchý regresní model a na jeho základech zjistete zdali kriminalita v okolí ovlivnuje cenu nemovitostí určených k bydlení. Pokud ano, o kolik je cena nemovitostí nižší v závislosti na míře kriminality? Ověřte předpoklady pro použití lineárního modelu (validujte např. symetrii a normalitu residui) a diskutujte výstup.

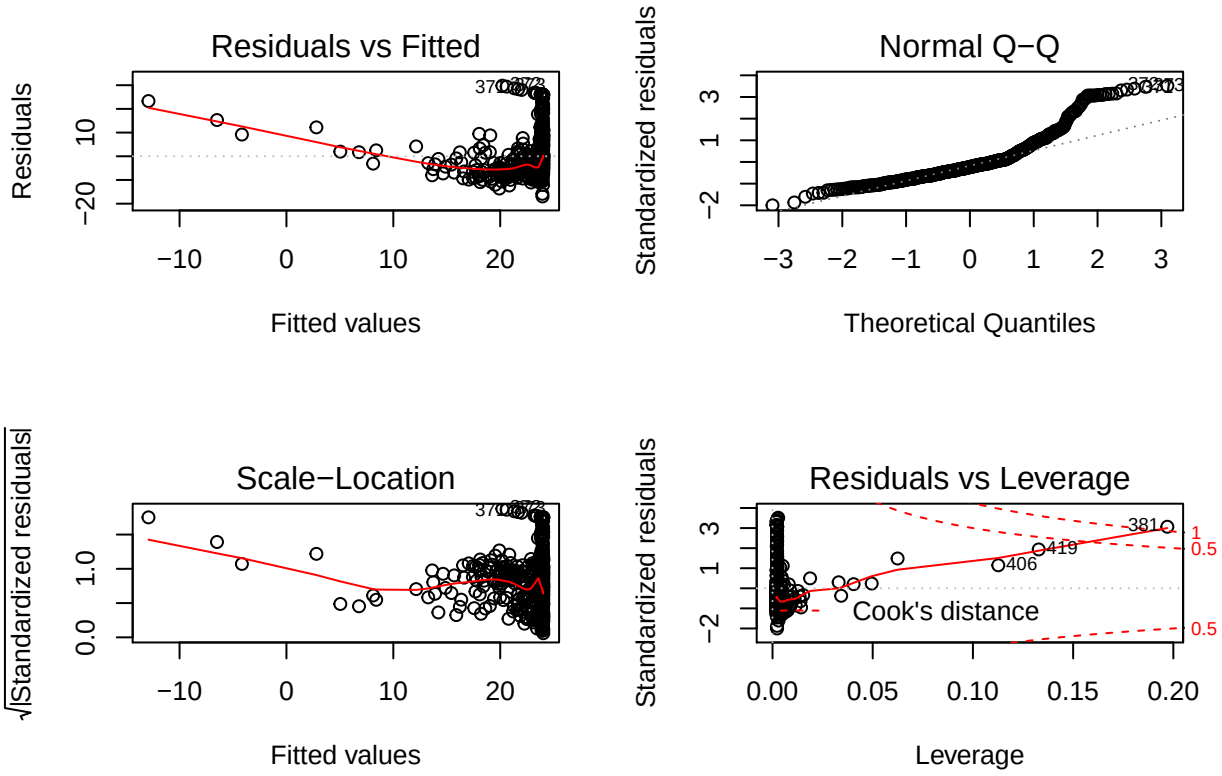
Na nasledujúcom grafe vidíme dáta vs lineárny model s a bez interceptu.

```
lm0 <- lm(medv ~ -1 + crim ,data = Boston)
lm1 <- lm(medv ~ crim ,data = Boston)
{plot(medv ~ crim,data = Boston
     ,xlab="Criminality", ylab="Median Value of owner-occupied homes in $1000's")
abline(lm0,col ="blue")
abline(lm1, col ="red")
par(xpd=TRUE)
legend(60,49,legend = c("without intercept","with intercept"),lty = c(1,1),
       col = c("blue","red"))}
```



Model bez interceptu nedáva zmysel (nepredpokladáme nulovú cenu nehnuteľností pri nulovej kriminalite). Model s interceptom je silno ovplyvnený širokým spektrom ceny nehnuteľností pri nízkej kriminalite - ak by sme tieto dáta odstránili, lineárny model by bol dobrým modelom. Pozrime sa ďalej kvantitatívne na predpoklady lin. modelu s interceptom:

```
{par(mfrow=c(2,2))
plot(lm1)}
```



V grafoch residuals vs fitted vidíme, že hodnoty na okrajoch sú asymetrické. Avšak môžeme povedať, že kriminalita a cena od seba rozhodne závisia lineárne, ak si odmyslíme hodnoty s nízkou kriminalitou. Zaujímavé sú pre nás teda až oblasti s vyššou kriminalitou, kde cena klesá. Cena klesá teda aspoň o koeficient lineárneho modelu s interceptom (deriváciu), ktorý je hodnoty:

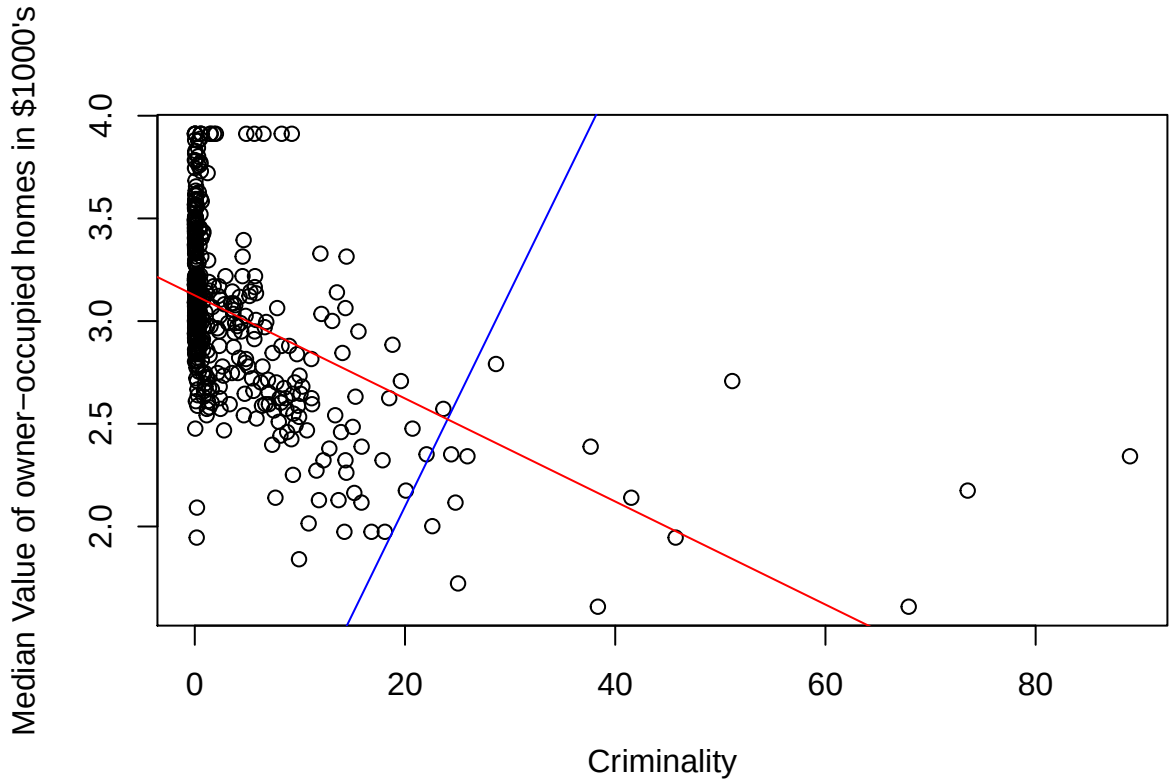
```
lm1$coefficients[2]
```

```
##       crim
## -0.4151903
```

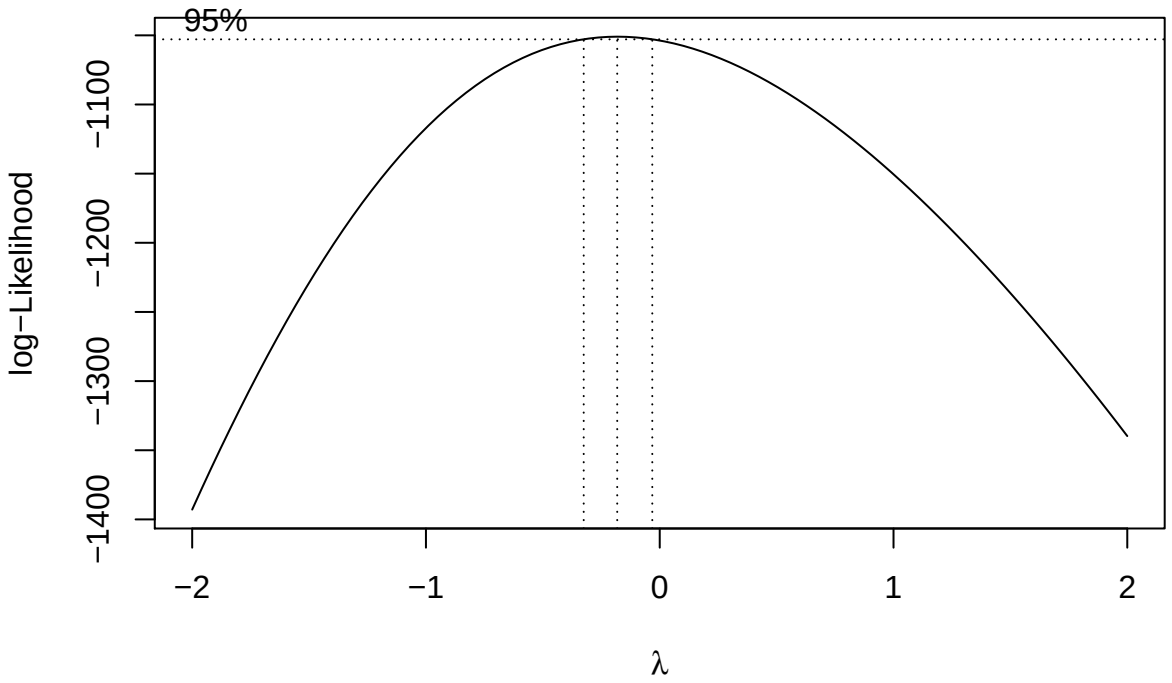
Q07

Vyzkoušejte model s logaritmickou transformací odezvy. Vykreslete optimální log-verohodnostní profil u Box-Coxovy transformace a porovnejte navrženou transformaci s provedenou logaritmickou.

```
#klasicka log transformacia
lmlog0 <- lm(log(medv) ~ -1 + crim ,data = Boston)
lmlog1 <- lm(log(medv) ~ crim ,data = Boston)
{plot(log(medv) ~ crim,data = Boston
     ,xlab="Criminality", ylab="Median Value of owner-occupied homes in $1000's")
abline(lmlog0,col ="blue")
abline(lmlog1, col ="red")
par(xpd=TRUE)
legend(2.8,-1,legend = c("without intercept","with intercept"),lty = c(1,1),
       col = c("blue","red"))}
```



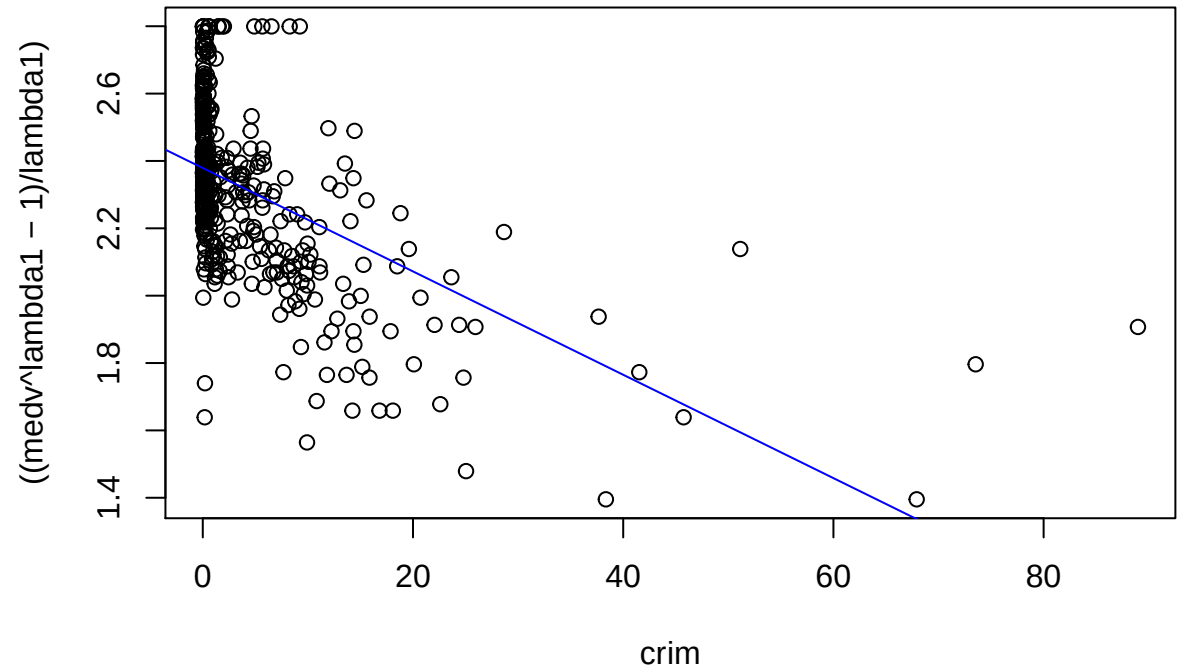

```
#box-cox
b1 = boxcox(lm(medv ~ -1 + crim ,data = Boston), lambda = seq(-2, 2, 1/10), plotit = TRUE,
            eps = 1/50, xlab = expression(lambda),
            ylab = "log-Likelihood")
```



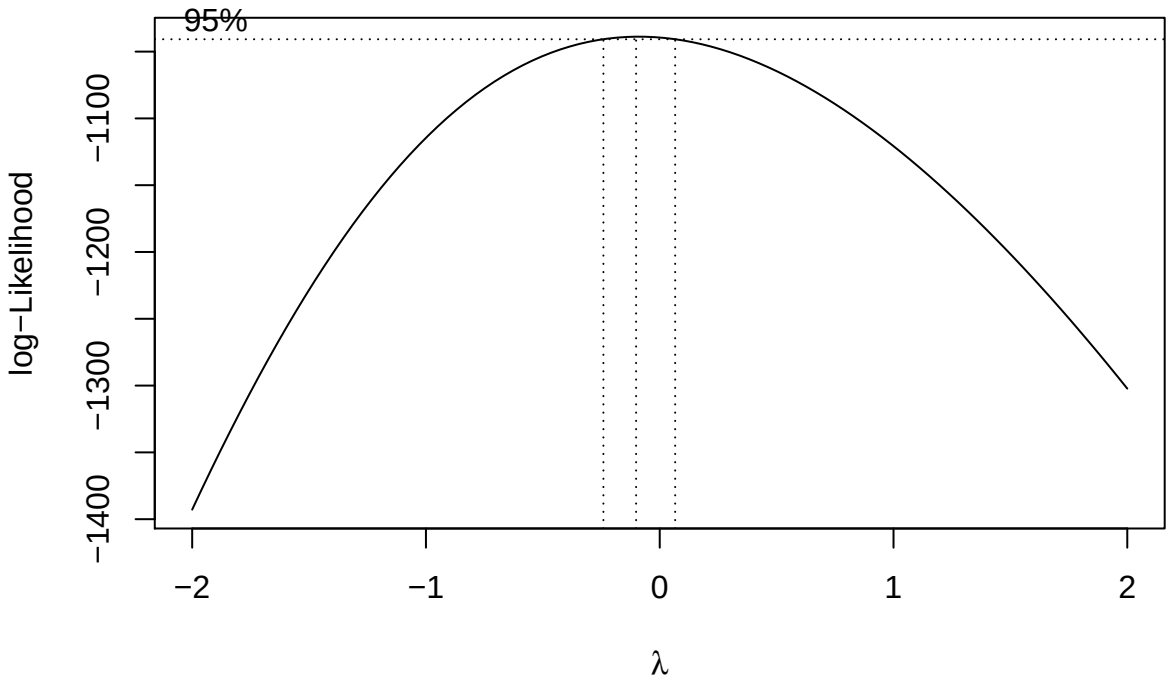

```
lambda1 = b1$x[which(b1$y==max(b1$y))]
lambda1
```

```
## [1] -0.1818182
```

```
mnew <- lm(((medv^lambda1-1)/lambda1) ~ crim, data = Boston)
{plot(((medv^lambda1-1)/lambda1) ~ crim,data = Boston)
abline(mnew,col ="blue")}
```



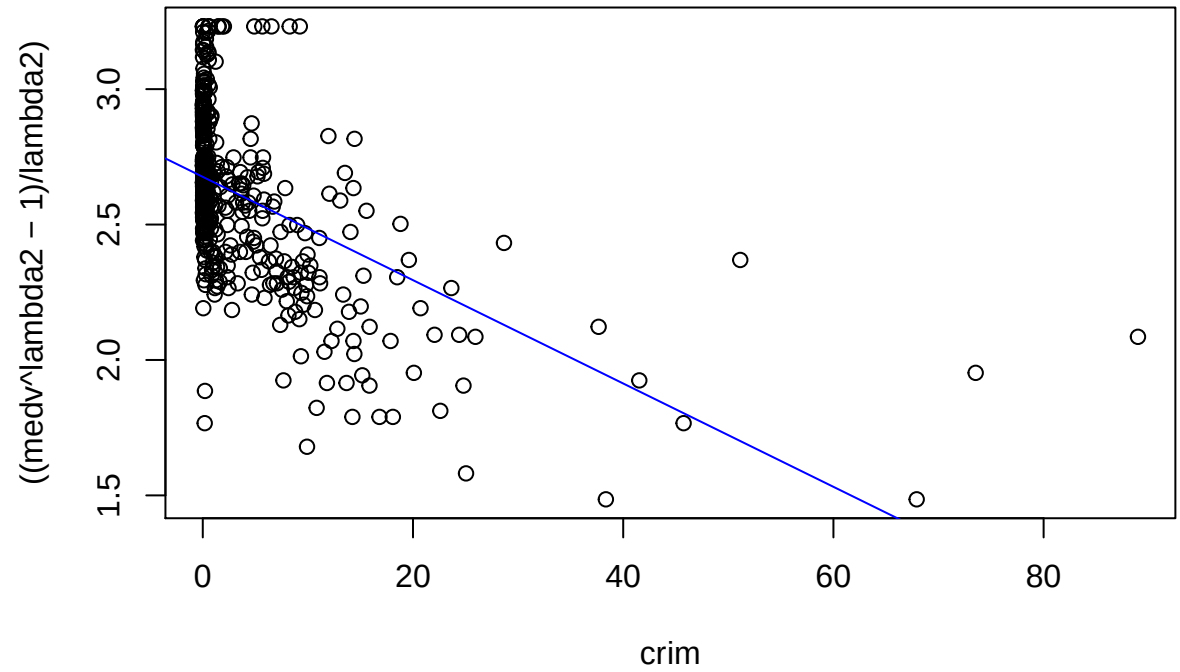

```
b2 = boxcox(lm(medv ~ crim ,data = Boston), lambda = seq(-2, 2, 1/10), plotit = TRUE,
            eps = 1/50, xlab = expression(lambda),
            ylab = "log-Likelihood")
```




```
lambda2 = b2$x[which(b2$y==max(b2$y))]
lambda2
```

```
## [1] -0.1010101
```

```
mnew <- lm(((medv^lambda2-1)/lambda2) ~ crim, data = Boston)
{plot(((medv^lambda2-1)/lambda2) ~ crim,data = Boston)
abline(mnew,col ="blue")}
```



Vidíme, že lambda z modelu bez interceptu nemá v konfidenčnom intervale nulu, ale lambda z modelu s interceptom má. Pre model s interceptom je teda dostatočne dobrá aproximácia logaritmus.

Q08

Z předchozího modelu vyctěte procentuální navýšení/pokles ceny nemovitostí při změně míry kriminality o jeden stupeň (odpověď typu: cena nemovitosti v prumeru klesne o ???% při nárustu míry kriminality o 1 jednotku).

```
beta1 <- summary(lmlog1)$coefficients[2, 1]
criminality_decrease = (1 - exp(beta1))*100
criminality_decrease
```

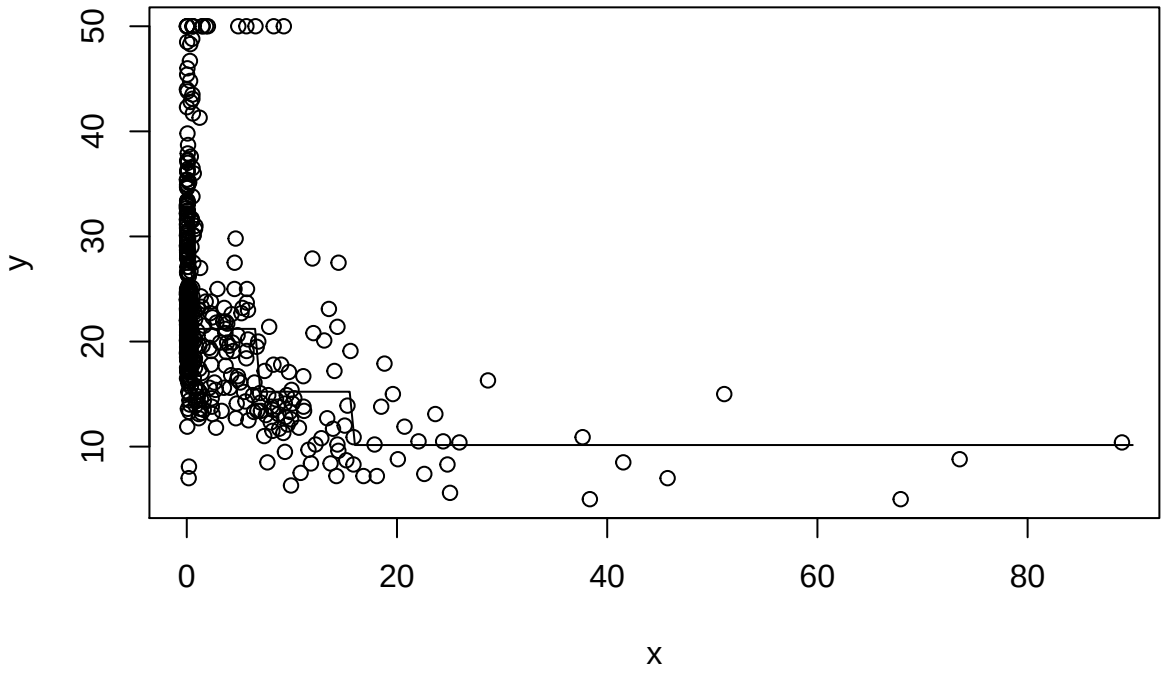
```
## [1] 2.477661
```

Cena nehnuteľnosti v priemere klesne o 2.477689% pri náraste miery kriminality o 1 jednotku.

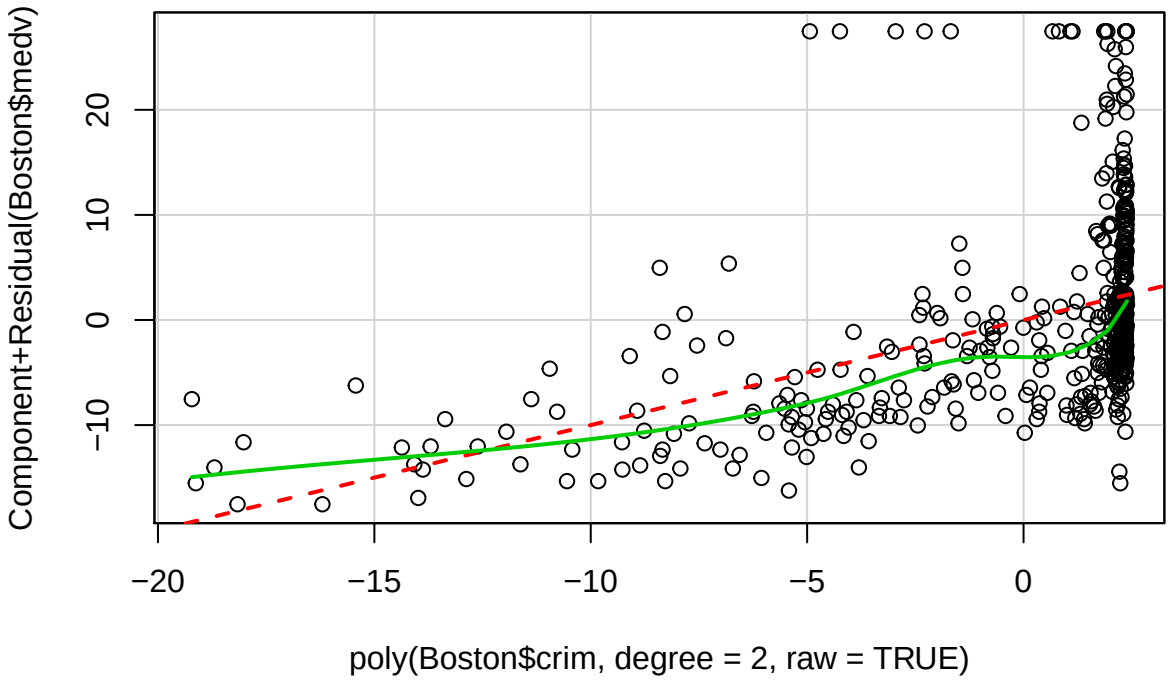
Q09

Zkuste transformovat proměnnou crim. Vyzkoušejte například po částech konstantní transformaci, lineární transformaci, splines a polynomiální transformaci(kvadratickou a kubickou). Zkuste využít informací získaných například z crPlots(model).

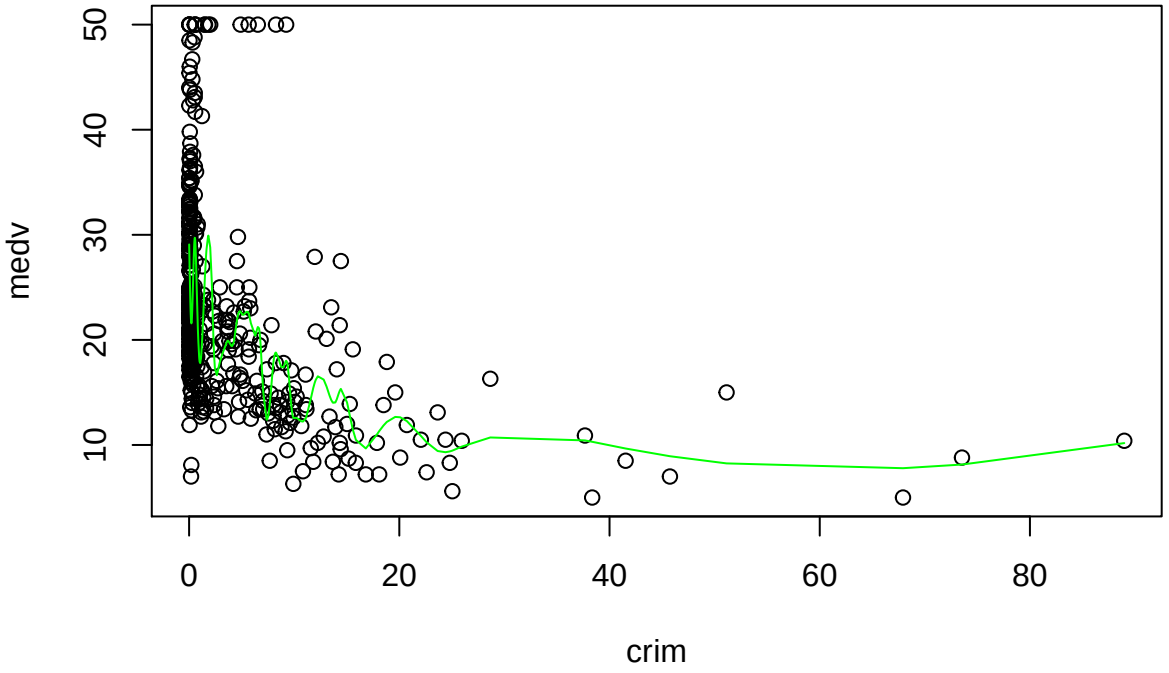
```
library(car)
library(rpart)
#piecewise constant regression
#tree <- rpart(medv ~ crim, data=Boston)
{ x <- Boston$crim
  y <- Boston$medv
  df <- data.frame(x=x,y=y)
  tree <-rpart(y~x,data=df)
plot_tree <- function(tree, x, y) {
  s <- seq(0, 90, by=.5)
  plot(x, y)
  lines(s, predict(tree, data.frame(x=s)))
}
plot_tree(tree, x, y)}
```



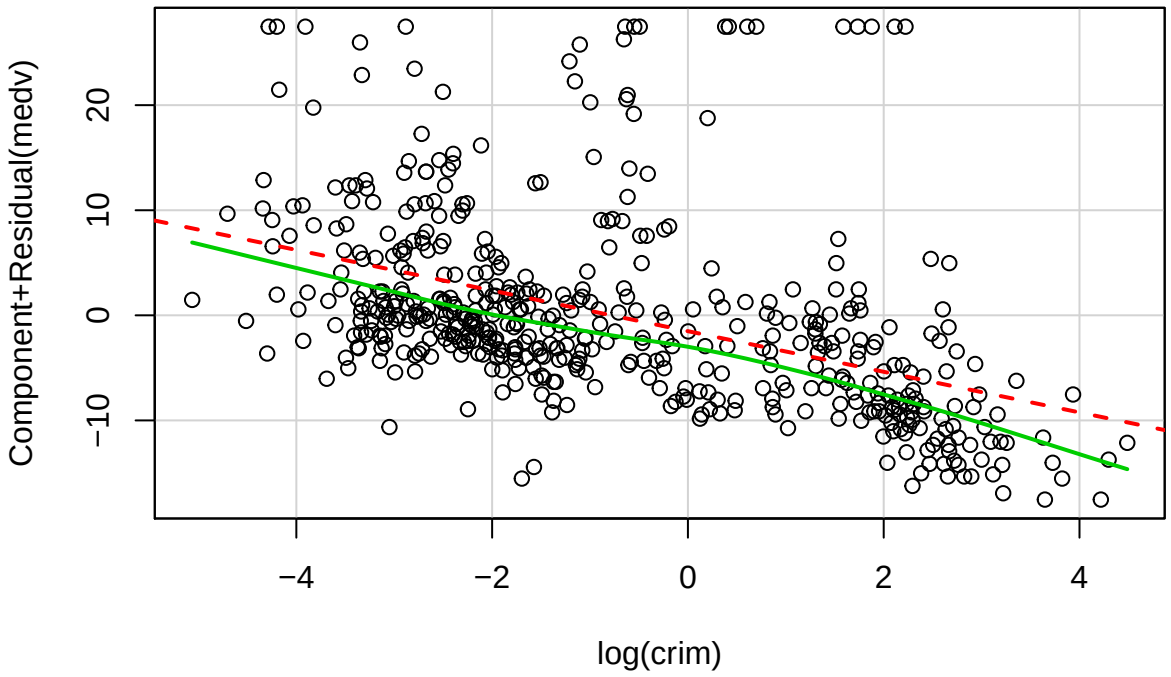

```
#quadratic transformation
polymodel=lm(Boston$medv ~ poly(Boston$crim, degree = 2, raw = TRUE))
crPlots(polymodel)
```




```
#smooth spline
{spln = with(Boston, smooth.spline(crim, medv))
plot(medv~crim, data=Boston)
lines(spln, col="green")}
```




```
#log transformation of crim
mojmodel1 <- lm( medv ~ log(crim) , data = Boston)
crPlots(mojmodel1)
```

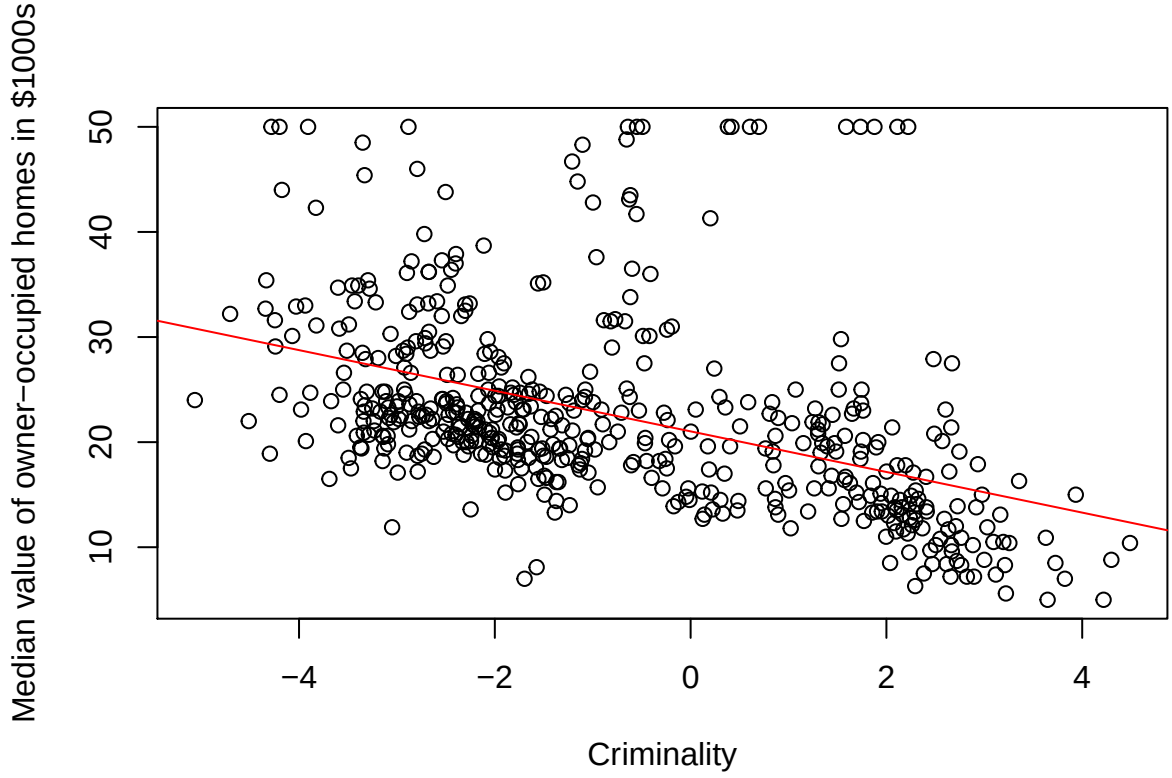

Component+Residual(medv)
20
10
0
−10
−4
−2
0
2
4
log(crim)

Q10

Vykreslete scatterplot predikovaných cen nemovitostí na základe vybraného modelu, proložte skrze data odhadnutou regresní přímku a vykreslete efekty pomocí plot(allEffects(model)). Validujte výsledný model pomocí příslušných testu na rezidua a pomocí příslušných obrázku (QQplot, residua vs. fitted, atd.)

Použijeme model s logaritmickou transformáciou kriminality:

```
{plot(log(Boston$crim), Boston$medv, ylab = medlab , xlab = "Criminality")
abline(mojmodel1, col="red")}
```



Vykreslíme efekty:

```
library(effects)
```

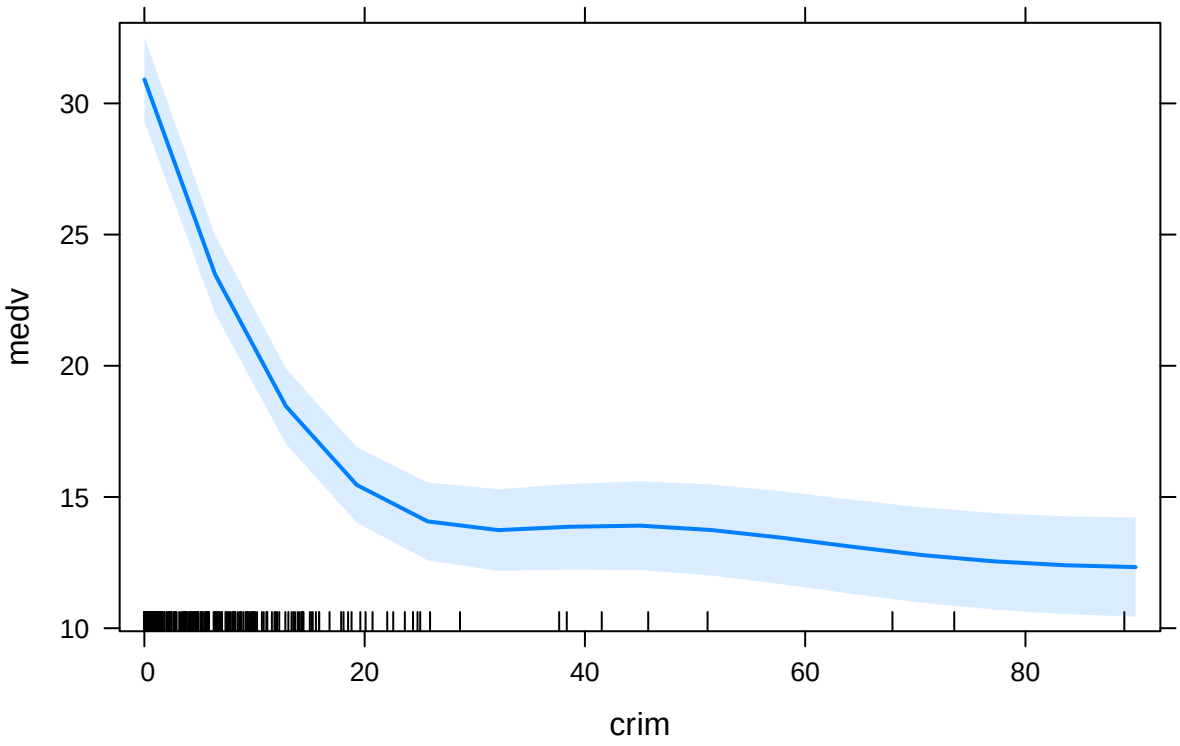
```
## Loading required package: carData
```

```
##
## Attaching package: 'carData'
```

```
## The following objects are masked from 'package:car':
##
##     Guyer, UN, Vocab
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
plot(allEffects(mojmodel1))
```



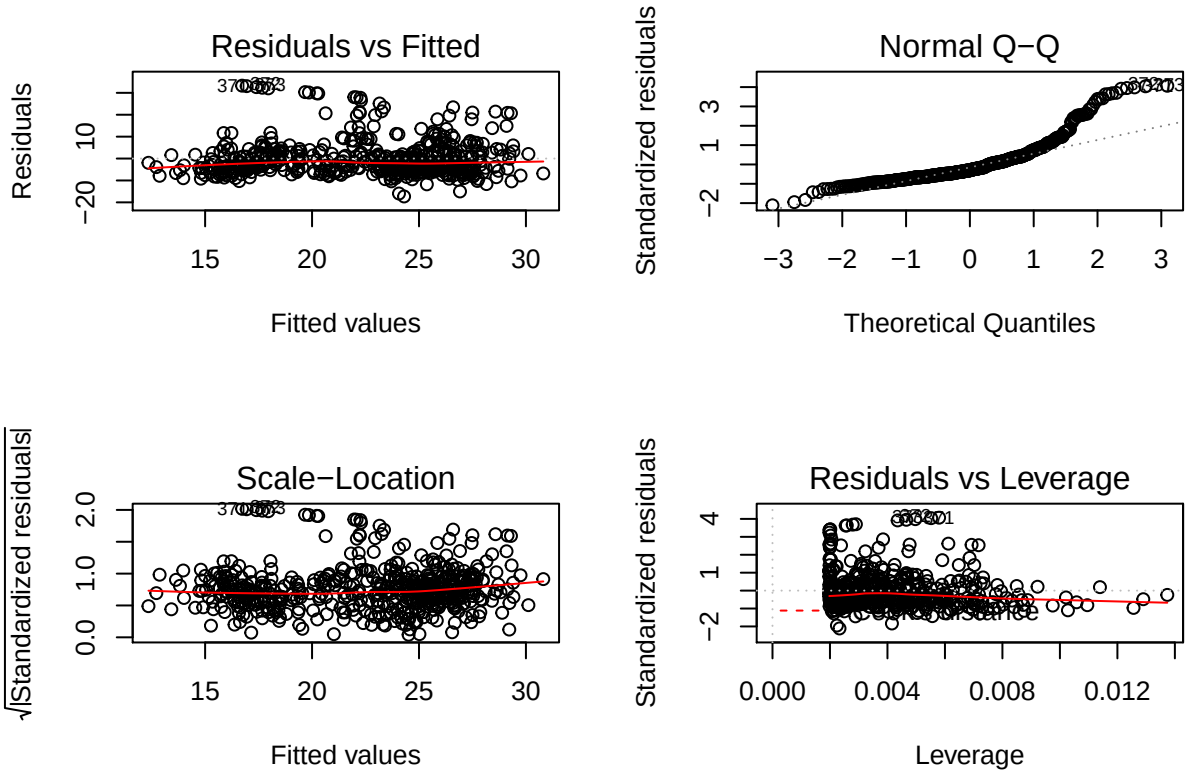
Validujeme model pomocou príslušných testov, ktoré v sebe zahŕňajú funkcie summary a plot:

```
summary(mojmodel1)
```

```
##
## Call:
## lm(formula = medv ~ log(crim), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.303  -5.159  -2.427   2.666  33.271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.0246     0.3877   54.23   <2e-16 ***
## log(crim)    -1.9325     0.1688  -11.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.201 on 504 degrees of freedom
## Multiple R-squared:  0.2064, Adjusted R-squared:  0.2048
## F-statistic: 131.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

Kedže p hodnota je veľmi malá, vzťah medzi premennými je signifikantný. Minimum a maximum reziduí nie sú symetrické, no viac nám povie graf:

```
{par(mfrow=c(2,2))
plot(mojmodel1)}
```



Residuá sú symetrické až na zopár outlierov. Rozptyl je homogénny. Na poslednom grafe vidíme, že Cookova vzdialenosť sa ani nedostala do škály na grafe, t.j. nemáme dáta, ktoré keby sme vynechali, tak by zásadne ovplyvnili našu regresnú priamku.

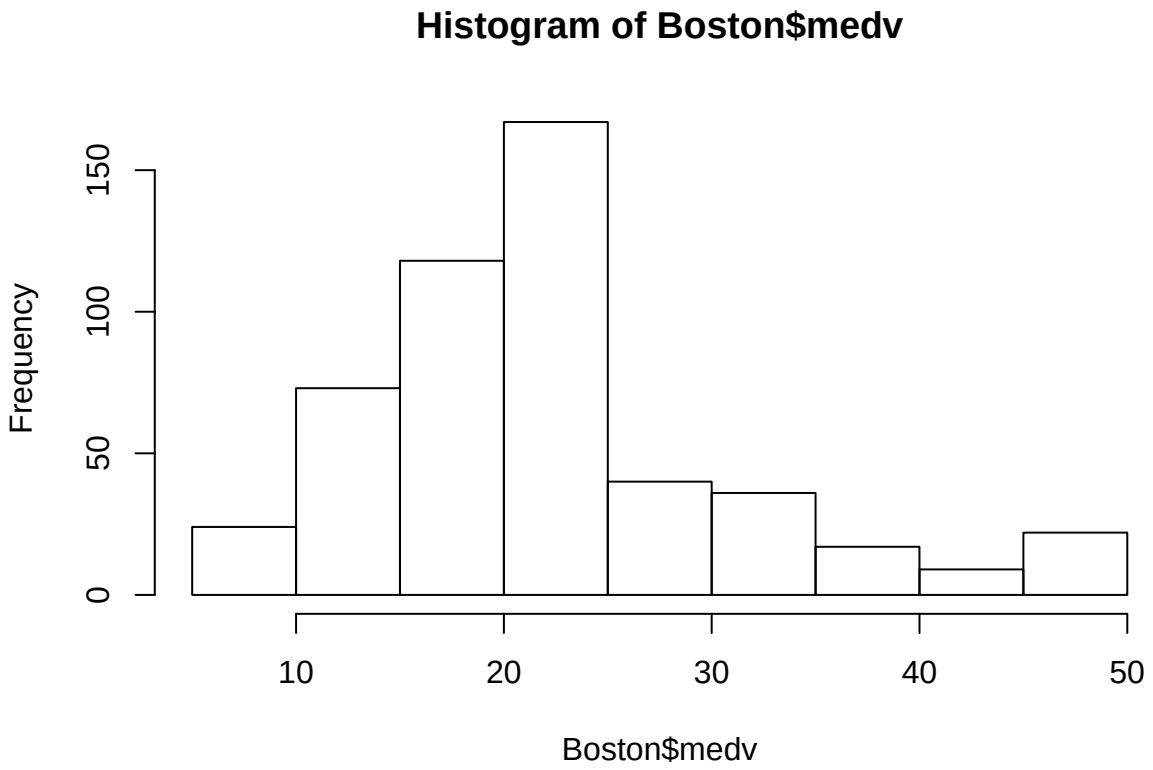
Q11

Medián ceny nemovitostí je spojitá promenná, vypište tabulku četností jednotlivých hodnot. Diskutujte zdali nekteré hodnoty nejsou zpusobeny zaokrouhlením, useknutím a podobne. Měření která považujete z tohoto pohledu za neduveryhodná, případne za outliery odstraňte.

```
table(Boston$medv)
```

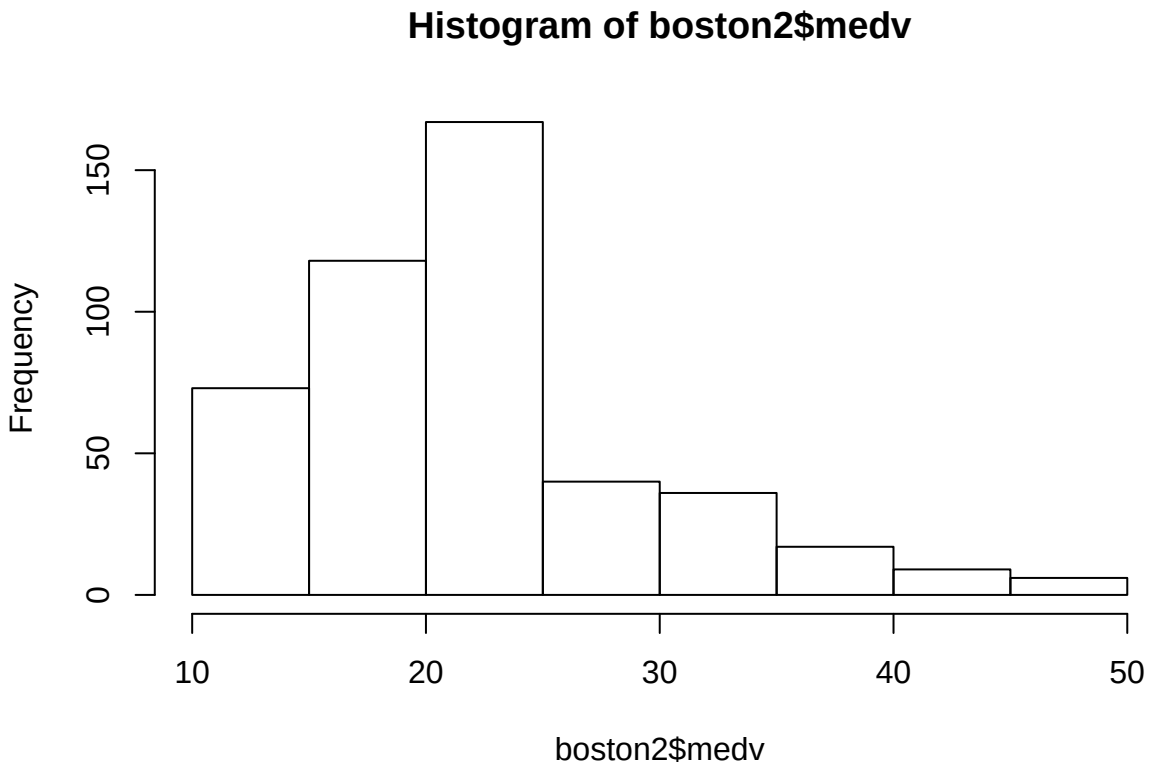
```
##
##    5  5.6  6.3    7  7.2  7.4  7.5  8.1  8.3  8.4  8.5  8.7  8.8  9.5  9.6
##    2    1    1    2    3    1    1    1    2    2    2    1    2    1    1
##  9.7 10.2 10.4 10.5 10.8 10.9   11 11.3 11.5 11.7 11.8 11.9   12 12.1 12.3
##    1    3    2    2    1    2    1    1    1    2    2    2    1    1    1
## 12.5 12.6 12.7 12.8   13 13.1 13.2 13.3 13.4 13.5 13.6 13.8 13.9   14 14.1
##    1    1    3    1    1    4    1    3    4    2    2    5    2    1    3
## 14.2 14.3 14.4 14.5 14.6 14.8 14.9   15 15.1 15.2 15.3 15.4 15.6 15.7   16
##    1    2    2    3    2    1    3    3    1    3    1    2    5    1    1
## 16.1 16.2 16.3 16.4 16.5 16.6 16.7 16.8   17 17.1 17.2 17.3 17.4 17.5 17.6
##    3    2    1    1    2    2    2    2    1    3    3    1    3    3    1
## 17.7 17.8 17.9   18 18.1 18.2 18.3 18.4 18.5 18.6 18.7 18.8 18.9   19 19.1
##    1    5    1    1    1    3    2    3    4    2    3    2    4    2    4
## 19.2 19.3 19.4 19.5 19.6 19.7 19.8 19.9   20 20.1 20.2 20.3 20.4 20.5 20.6
##    2    5    6    4    5    2    3    4    5    5    2    4    4    3    6
## 20.7 20.8 20.9   21 21.1 21.2 21.4 21.5 21.6 21.7 21.8 21.9   22 22.1 22.2
##    2    3    2    3    2    5    5    2    2    7    2    3    7    1    5
## 22.3 22.4 22.5 22.6 22.7 22.8 22.9   23 23.1 23.2 23.3 23.4 23.5 23.6 23.7
##    2    2    3    5    2    4    4    4    7    4    4    2    1    2    4
## 23.8 23.9   24 24.1 24.2 24.3 24.4 24.5 24.6 24.7 24.8   25 25.1 25.2 25.3
##    4    5    2    3    1    3    4    3    2    3    4    8    1    1    1
## 26.2 26.4 26.5 26.6 26.7   27 27.1 27.5 27.9   28 28.1 28.2 28.4 28.5 28.6
##    1    2    1    3    1    1    2    4    2    1    1    1    2    1    1
## 28.7   29 29.1 29.4 29.6 29.8 29.9 30.1 30.3 30.5 30.7 30.8   31 31.1 31.2
##    3    2    2    1    2    2    1    3    1    1    1    1    1    1    1
## 31.5 31.6 31.7   32 32.2 32.4 32.5 32.7 32.9   33 33.1 33.2 33.3 33.4 33.8
##    2    2    1    2    1    1    1    1    1    1    2    2    1    2    1
## 34.6 34.7 34.9 35.1 35.2 35.4   36 36.1 36.2 36.4 36.5   37 37.2 37.3 37.6
##    1    1    3    1    1    2    1    1    2    1    1    1    1    1    1
## 37.9 38.7 39.8 41.3 41.7 42.3 42.8 43.1 43.5 43.8   44 44.8 45.4   46 46.7
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 48.3 48.5 48.8   50
##    1    1    1   16
```

```
hist(Boston$medv, breaks = 15)
```



Nehnuteľnosti s hodnotou <10 boli zaokrúhlené na celé čísla. Stalo sa tak pravdepodobne pre nízky počet týchto hodnôt (vid histogram). Nízku početnosť môžme pozorovať aj pre nehnuteľnosti s hodnotou medzi 35 a 45. Nárast pre hodnotu 50 by som popisovala useknutiu dát po túto hodnotu. Odstránime preto dáta, ktoré boli zaokrúhľované inak ako zvyšok (všetko pod 10) a dáta, ktoré sa zaokruhlili na 50.

```
boston2 <- Boston[Boston$medv != 50,]
boston2 <- boston2[(boston2$medv >= 10),]
hist(boston2$medv)
```


Histogram of boston2$medv
Frequency
0
50
100
150
10
20
30
40
50
boston2$medv

Q12

Zkonstruujte lineární model s logaritmicky transformovanou odezvou medv a všemi nezávislými promennými, které máte k dispozici. Na základě kritérií jako jsou AIC, BIC, R2, F, atd. Vyberte nejvhodnejší model. Ten validujte a okomentujte jeho výběr.

Pomocou funkcie step vyskúšame lineárne modely so všetkými vysvetlujúcimi premennými. Kedže sme nešpecifikovali smer, posledný model je ten, ktorý našiel najnižšiu hodnotu BIC resp AIC a teda je podľa tohto kritéria najlepší. Aplikujeme AIC a BIC kritérium, výstup týchto funkcií nám zároveň vypočíta R^2 a F.

Hľadajme najlepší AIC model:

```
library(stats)
model= lm(log(medv) ~(.), data=boston2)
aicmodel = step(model)
```

```
## Start:  AIC=-1739.62
## log(medv) ~ (crim + zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat)
##
##           Df Sum of Sq    RSS     AIC
## - indus    1   0.00001 10.496 -1741.6
## - chas     1   0.03288 10.529 -1740.2
## <none>                 10.496 -1739.6
## - zn       1   0.12187 10.618 -1736.2
## - age      1   0.12534 10.621 -1736.1
## - black    1   0.40332 10.899 -1724.0
## - tax      1   0.45109 10.947 -1722.0
## - nox      1   0.55911 11.055 -1717.4
## - rad      1   0.58116 11.077 -1716.5
## - crim     1   0.69438 11.191 -1711.8
## - dis      1   1.13552 11.632 -1693.8
## - ptratio  1   1.52123 12.017 -1678.5
## - rm       1   1.56058 12.057 -1677.0
## - lstat    1   2.22140 12.717 -1652.2
##
## Step:  AIC=-1741.62
## log(medv) ~ crim + zn + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat
##
##           Df Sum of Sq    RSS     AIC
## - chas     1   0.03296 10.529 -1742.2
## <none>                 10.496 -1741.6
## - zn       1   0.12306 10.619 -1738.2
## - age      1   0.12543 10.621 -1738.1
## - black    1   0.40409 10.900 -1726.0
## - tax      1   0.52628 11.022 -1720.8
## - nox      1   0.60921 11.105 -1717.3
## - rad      1   0.61219 11.108 -1717.2
## - crim     1   0.69447 11.191 -1713.8
## - dis      1   1.18883 11.685 -1693.6
## - ptratio  1   1.55745 12.053 -1679.2
## - rm       1   1.57448 12.071 -1678.5
## - lstat    1   2.25319 12.749 -1653.0
##
```

```
## Step:  AIC=-1742.16
## log(medv) ~ crim + zn + nox + rm + age + dis + rad + tax + ptratio +
##     black + lstat
##
##           Df Sum of Sq    RSS     AIC
## <none>                 10.529 -1742.2
## - zn       1   0.12219 10.651 -1738.8
## - age      1   0.12342 10.652 -1738.7
## - black    1   0.41275 10.942 -1726.2
## - tax      1   0.56745 11.097 -1719.7
## - nox      1   0.58939 11.118 -1718.8
## - rad      1   0.64602 11.175 -1716.4
## - crim     1   0.71355 11.243 -1713.6
## - dis      1   1.19843 11.727 -1693.9
## - rm       1   1.58916 12.118 -1678.7
## - ptratio  1   1.60228 12.131 -1678.2
## - lstat    1   2.24512 12.774 -1654.1
```

Najlepší BIC model:

```
bicmodel = step(model, criterion = "BIC", k = log(dim(boston2)[1]))
```

```
## Start:  AIC=-1681.61
## log(medv) ~ (crim + zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat)
##
##           Df Sum of Sq    RSS     AIC
## - indus    1   0.00001 10.496 -1687.8
## - chas     1   0.03288 10.529 -1686.3
## - zn       1   0.12187 10.618 -1682.4
## - age      1   0.12534 10.621 -1682.2
## <none>                 10.496 -1681.6
## - black    1   0.40332 10.899 -1670.2
## - tax      1   0.45109 10.947 -1668.1
## - nox      1   0.55911 11.055 -1663.6
## - rad      1   0.58116 11.077 -1662.6
## - crim     1   0.69438 11.191 -1657.9
## - dis      1   1.13552 11.632 -1639.9
## - ptratio  1   1.52123 12.017 -1624.7
## - rm       1   1.56058 12.057 -1623.2
## - lstat    1   2.22140 12.717 -1598.3
##
## Step:  AIC=-1687.75
## log(medv) ~ crim + zn + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat
##
##           Df Sum of Sq    RSS     AIC
## - chas     1   0.03296 10.529 -1692.4
## - zn       1   0.12306 10.619 -1688.5
## - age      1   0.12543 10.621 -1688.4
## <none>                 10.496 -1687.8
## - black    1   0.40409 10.900 -1676.3
## - tax      1   0.52628 11.022 -1671.1
## - nox      1   0.60921 11.105 -1667.6
## - rad      1   0.61219 11.108 -1667.5
```

```
## - crim     1   0.69447 11.191 -1664.0
## - dis      1   1.18883 11.685 -1643.9
## - ptratio  1   1.55745 12.053 -1629.4
## - rm       1   1.57448 12.071 -1628.8
## - lstat    1   2.25319 12.749 -1603.3
##
## Step:  AIC=-1692.43
## log(medv) ~ crim + zn + nox + rm + age + dis + rad + tax + ptratio +
##     black + lstat
##
##           Df Sum of Sq    RSS     AIC
## - zn       1   0.12219 10.651 -1693.2
## - age      1   0.12342 10.652 -1693.2
## <none>                 10.529 -1692.4
## - black    1   0.41275 10.942 -1680.7
## - tax      1   0.56745 11.097 -1674.1
## - nox      1   0.58939 11.118 -1673.2
## - rad      1   0.64602 11.175 -1670.8
## - crim     1   0.71355 11.243 -1668.0
## - dis      1   1.19843 11.727 -1648.3
## - rm       1   1.58916 12.118 -1633.1
## - ptratio  1   1.60228 12.131 -1632.6
## - lstat    1   2.24512 12.774 -1608.5
##
## Step:  AIC=-1693.2
## log(medv) ~ crim + nox + rm + age + dis + rad + tax + ptratio +
##     black + lstat
##
##           Df Sum of Sq    RSS     AIC
## <none>                 10.651 -1693.2
## - age      1   0.15085 10.802 -1692.8
## - black    1   0.41493 11.066 -1681.5
## - tax      1   0.48388 11.135 -1678.6
## - rad      1   0.59147 11.243 -1674.2
## - nox      1   0.62206 11.273 -1672.9
## - crim     1   0.66872 11.320 -1671.0
## - dis      1   1.10410 11.755 -1653.4
## - rm       1   1.77827 12.430 -1627.4
## - ptratio  1   2.11227 12.764 -1615.0
## - lstat    1   2.21759 12.869 -1611.2
```

Pozrime sa na ich F a R^2 hodnoty: AIC:

```
summary(aicmodel)
```

```
##
## Call:
## lm(formula = log(medv) ~ crim + zn + nox + rm + age + dis + rad +
##     tax + ptratio + black + lstat, data = boston2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57637 -0.09043 -0.01310  0.08468  0.70551
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.714e+00  1.690e-01  21.972  < 2e-16 ***
## crim        -7.893e-03  1.423e-03  -5.547 4.95e-08 ***
## zn           1.046e-03  4.557e-04   2.295   0.0222 *
## nox         -6.032e-01  1.197e-01  -5.041 6.69e-07 ***
## rm           1.234e-01  1.491e-02   8.278 1.41e-15 ***
## age         -1.013e-03  4.393e-04  -2.307   0.0215 *
## dis         -4.592e-02  6.388e-03  -7.189 2.71e-12 ***
## rad          1.174e-02  2.224e-03   5.278 2.03e-07 ***
## tax         -5.795e-04  1.172e-04  -4.947 1.07e-06 ***
## ptratio     -3.505e-02  4.216e-03  -8.312 1.10e-15 ***
## black        4.108e-04  9.737e-05   4.219 2.97e-05 ***
## lstat       -1.817e-02  1.846e-03  -9.839  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1523 on 454 degrees of freedom
## Multiple R-squared:  0.7757, Adjusted R-squared:  0.7702
## F-statistic: 142.7 on 11 and 454 DF,  p-value: < 2.2e-16
```

BIC:

```
summary(bicmodel)
```

```
##
## Call:
## lm(formula = log(medv) ~ crim + nox + rm + age + dis + rad +
##     tax + ptratio + black + lstat, data = boston2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56443 -0.09435 -0.01610  0.08715  0.71453
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.718e+00  1.698e-01  21.895  < 2e-16 ***
## crim        -7.613e-03  1.424e-03  -5.345 1.43e-07 ***
## nox         -6.187e-01  1.200e-01  -5.155 3.79e-07 ***
## rm           1.289e-01  1.479e-02   8.716  < 2e-16 ***
## age         -1.115e-03  4.391e-04  -2.538   0.0115 *
## dis         -3.925e-02  5.715e-03  -6.868 2.15e-11 ***
## rad          1.116e-02  2.220e-03   5.027 7.19e-07 ***
## tax         -5.233e-04  1.151e-04  -4.546 7.00e-06 ***
## ptratio     -3.814e-02  4.015e-03  -9.499  < 2e-16 ***
## black        4.119e-04  9.783e-05   4.210 3.08e-05 ***
## lstat       -1.805e-02  1.854e-03  -9.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.153 on 455 degrees of freedom
## Multiple R-squared:  0.7731, Adjusted R-squared:  0.7681
## F-statistic:   155 on 10 and 455 DF,  p-value: < 2.2e-16
```

Za najlepší považujem aicmodel s najnižšou hodnotou AIC, ktorý ma aj najnižšiu hodnotu F a vhodnú hodnotu R-squared.

Q13

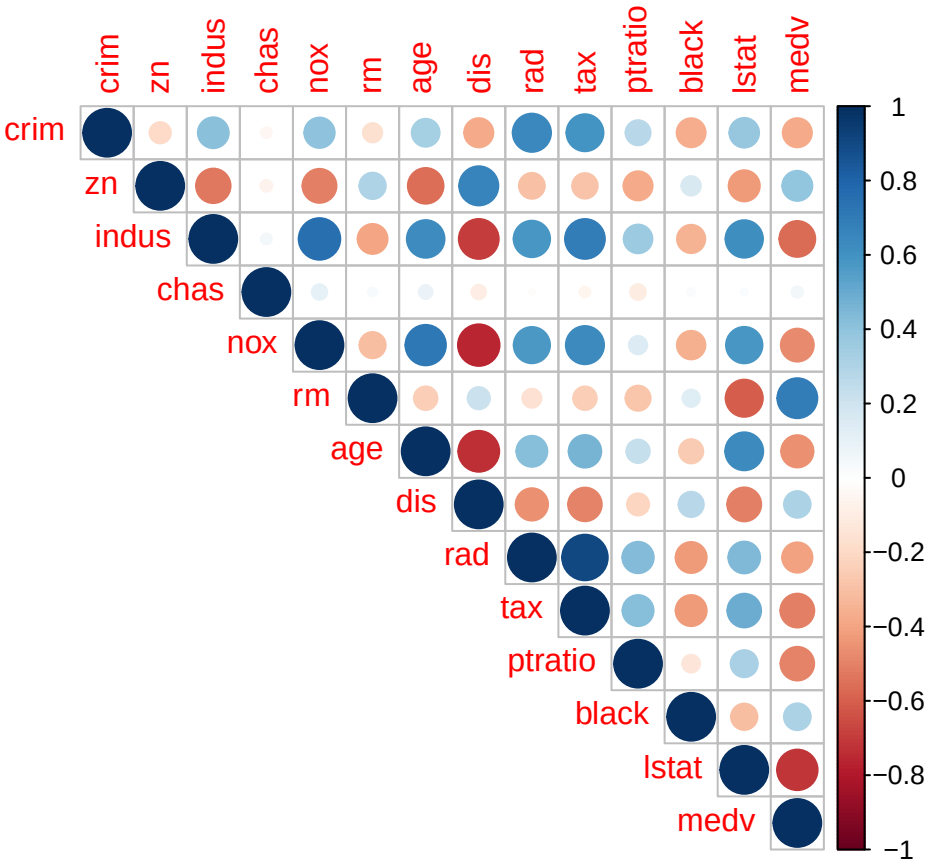
Zkoumejte případnou multikolinearitu. Spoctěte korelace mezi jednotlivými promennými, porovnejte s vaším výběrem a pomocí VIF a dalších nástroju validujte váš výber.

Predchádzajúci výber odstránil premenné age, chas, indus. Ako vidíme na diagrame nižšie, chas nekoreluje so žiadnou premennou a teda nie je dôvod zaraďovať ho do nášho modelu. indus a age taktiež vykazujú vo svojich riadkoch a stĺpcoch pomerne silné zastúpenie veľkých kruhov so sýtymi farbami - silnú koreláciu - takže tiež dáva zmysel ich vyradiť.

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(car)
M<-cor(boston2)
corrplot(M, type = "upper")
```



Hodnota VIF poukazuje na to, že by bolo dobré odstrániť aj premenné rad a tax:

```
vif(aicmodel)
```

```
##     crim       zn      nox       rm      age      dis      rad      tax
## 1.792771 2.256803 3.805033 1.840900 3.073851 3.627013 6.835998 7.149156
##  ptratio    black    lstat
## 1.615625 1.281989 2.855086
```

```
finalmodel = lm(log(medv) ~ crim + zn + nox + rm + dis + ptratio +
    black + lstat, data=boston2)
```

Q14

Pokud ve vašem výsledném modelu máte zahrnutou kriminalitu (promennou crim) porovnejte jak se zmenil vliv kriminality na medián ceny nemovitostí. Jaké je snížení prumerné ceny nemovitostí při vzrustu kriminality o jednu jednotku? Pokud crim v modelu nemáte tak ji pro tuto otázku do modelu přiřad'te.

Pôvodná hodnota koeficientu pri crim bola -0.4151903 a teraz je -0.006326841. Cena klesá o dva rády pomalšie v porovnaní s prvým modelom.

```
summary(finalmodel)$coefficients[2, 1]
```

```
## [1] -0.006326841
```

Q15

Prezentujte váš výsledný model pro predikci medv, diskutujte výsledné parametry R2 a sigma tohoto modelu. Validujte model (jak graficky, tak pomocí příslušných testu hypotéz).

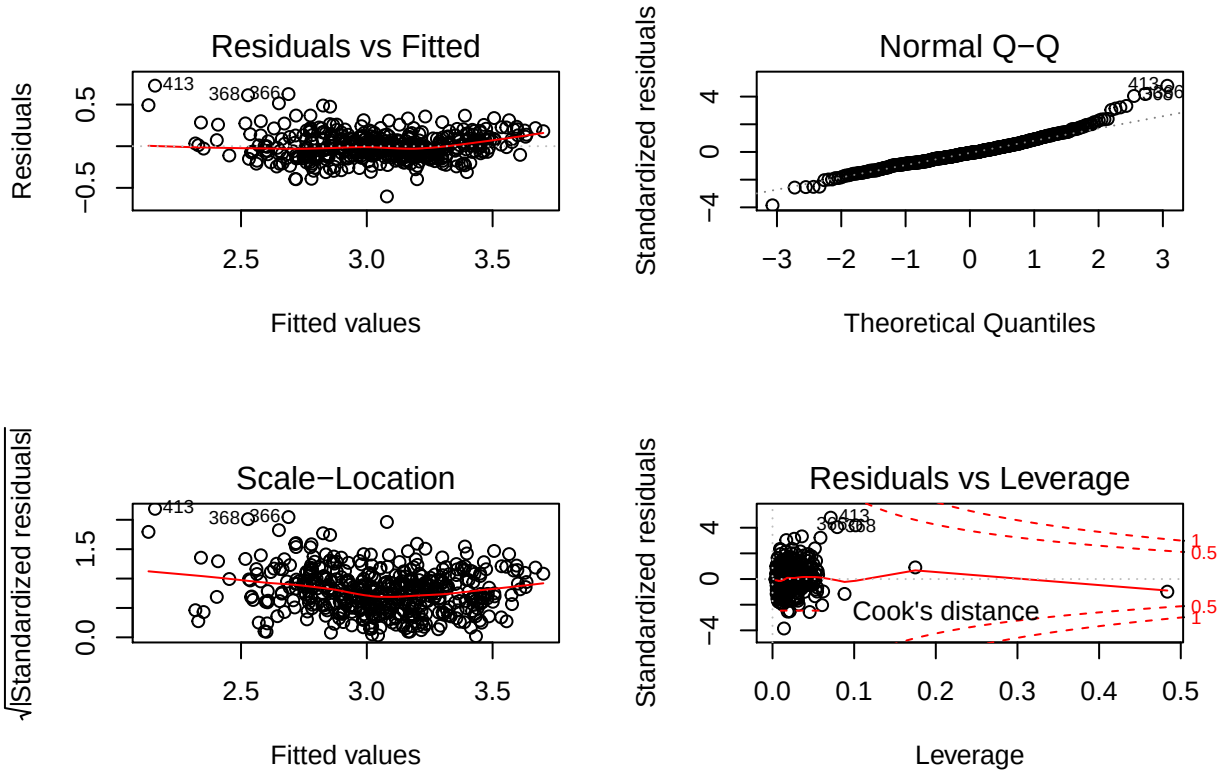
Môj výsledný model nezahŕňa premenné age, chas, indus, rad a tax.

```
summary(finalmodel)
```

```
##
## Call:
## lm(formula = log(medv) ~ crim + zn + nox + rm + dis + ptratio +
##     black + lstat, data = boston2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60333 -0.10342 -0.01108  0.08212  0.72662
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5441765  0.1641444  21.592  < 2e-16 ***
## crim        -0.0063268  0.0013098  -4.830 1.86e-06 ***
## zn           0.0007837  0.0004571   1.714 0.087120 .
## nox         -0.7164246  0.1068686  -6.704 6.00e-11 ***
## rm           0.1297517  0.0147269   8.811  < 2e-16 ***
## dis         -0.0388818  0.0063158  -6.156 1.63e-09 ***
## ptratio     -0.0349812  0.0039365  -8.886  < 2e-16 ***
## black        0.0003852  0.0000992   3.883 0.000118 ***
## lstat       -0.0198957  0.0017474 -11.386  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1577 on 457 degrees of freedom
## Multiple R-squared:  0.7578, Adjusted R-squared:  0.7536
## F-statistic: 178.7 on 8 and 457 DF,  p-value: < 2.2e-16
```

Maximá reziduí su relatívne symetrické. Premenná zn (proportion of residential land zoned for lots over 25,000 sq.ft.) má veľkú hodnotu t, ak by sme ju vvšak odstránili, zvyšným testom by to nepomohlo (step funkcia nám predsa v úlohe 12 našla najvýhodnejšiu kombináciu premenných).

```
{par(mfrow=c(2,2))
plot(finalmodel)}
```



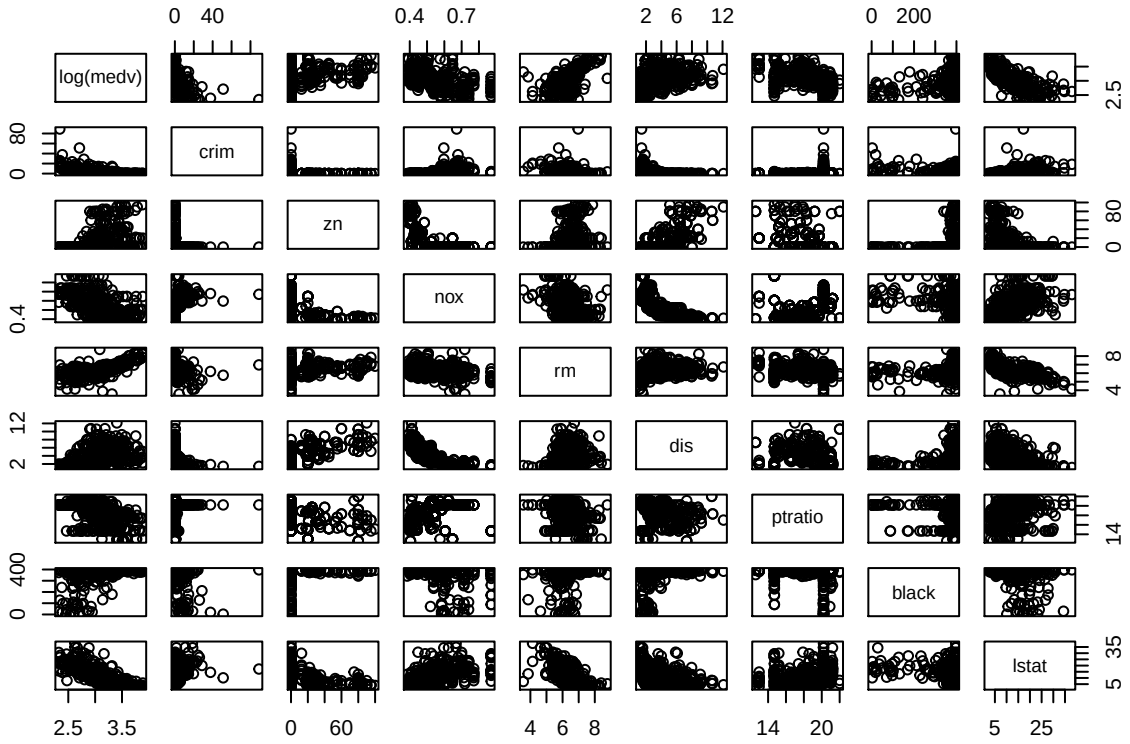
Reziduá sú symetrické až na hodnoty 413, 368, 366. Tieto hodnoty figurujú aj v QQ-plote, ale nepovažujem ich za až tak signifikantné. Aj v poslednom grafe vidíme, že ani jedna hodnota nie je vzdialená o viac ako Cookovu vzdialenosť.

Q16

Diskutujte jak by šlo případne zlepšit predikci, jaké transformace jednotlivých promenných by mohli pomoci. Převedli byste nekteré spojité proměnné na diskrétní (na faktory)? Jaké další kroky byste při analýze navrhli?

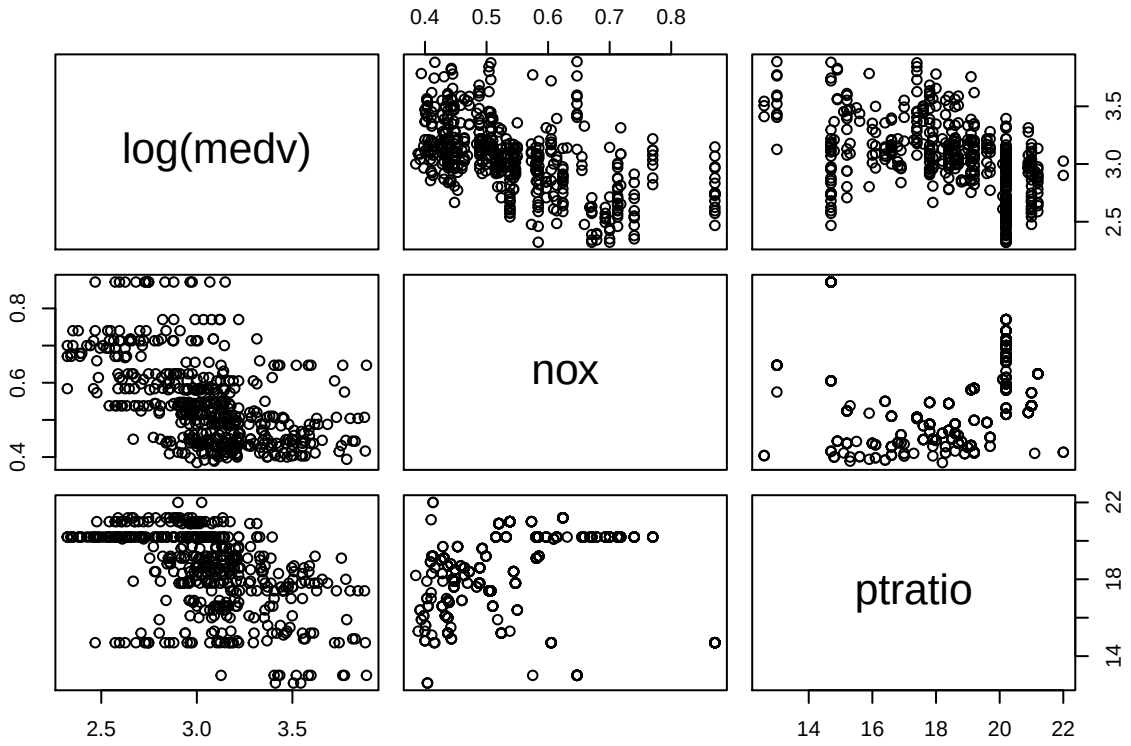
Odstránila by som dáta pre kriminalitu nižšiu ako maximum z distribučnej funkcie kriminality. V oblasti s nízkou kriminalitou sa totižto prirodzene nachádzaju drahé aj lacné nehnuteľnosti (staré, nové, luxusné, schátrané), sú pre to pre našu analýzu závislosti zbytočné. Ďalšie nápady nám môže ponúknuť funkcia pairs:

```
pairs(log(medv) ~ crim + zn + nox + rm + dis + ptratio +
    black + lstat, data=boston2)
```



Bolo by vhodné vyskúšať, či by pomohla faktorizácia premenných nox a ptratio:

```
pairs(log(medv) ~  nox + ptratio, data=boston2)
```


log(medv)
nox
ptratio
0.4
0.5
0.6
0.7
0.8
2.5
3.0
3.5
0.4
0.6
0.8
14
16
18
20
22

Q17

Myslíte, že pokud bychom cílene dokázali potlačit kriminalitu v daném městě, vedlo by to ke zvýšení cen nemovitostí určených k bydlení v dané lokalite?

Áno. Kriminalita a cena sú od seba určite závislé. Nemyslím si však, že cenový rozdiel by bol z krátkeho časového hľadiska signifikantný, pretože je ťažké zmeniť názor verejnosti na nejakú oblasť na základe dát.