SPRAWOZDANIE 2

Regresja liniowa

Krzyszczuk Michał

7 grudnia 2017, z późniejszymi zmianami 2 stycznia 2018

Ad. 1 Wczytanie danych.

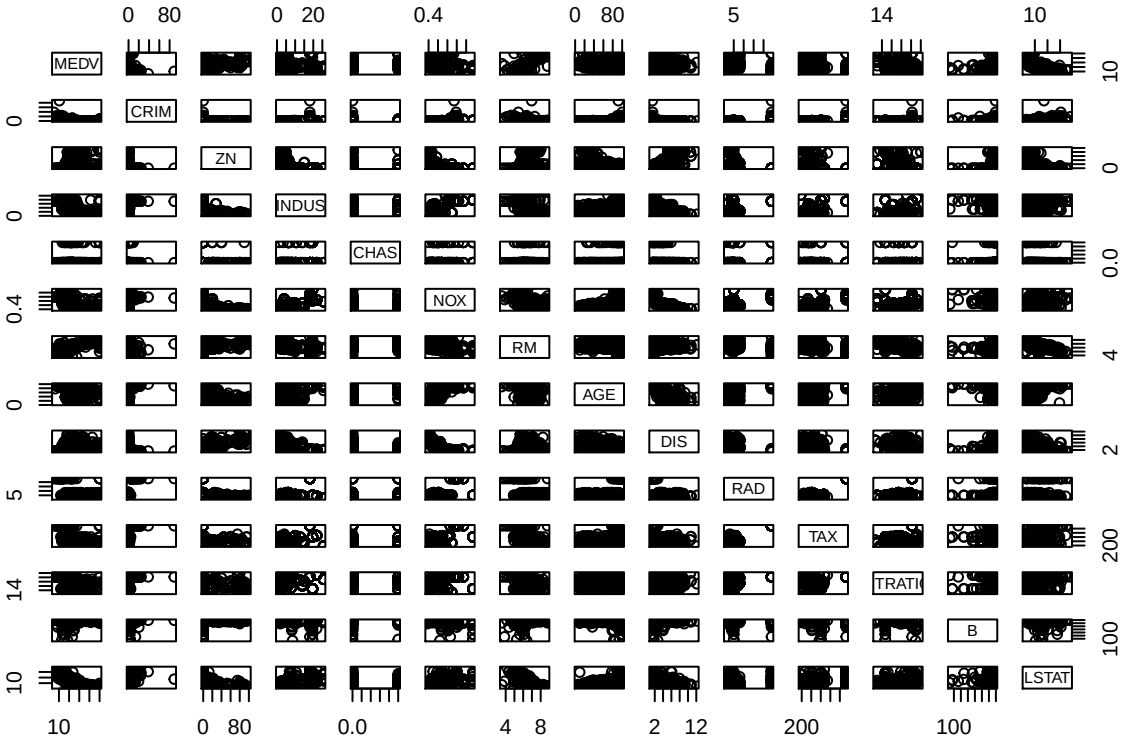
```
boston_data <- read.csv("boston.csv",header=TRUE)#boston.csv is in working directory
```

Ad.2 Podzielenie danych.

```
vector <- split(boston_data,rep(1:2,c(400,106)))
boston_data1 <- vector$`1`
boston_data2 <- vector$`2`
rm(vector)
```

Ad.3

```
pairs(boston_data1)
```



Model regresji liniowej wykorzystujący wszystkie zmienne niezależne.

```
model1 <- lm(MEDV~CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT,data=boston_data1)
summary(model1)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
##     DIS + RAD + TAX + PTRATIO + B + LSTAT, data = boston_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5636  -2.6945  -0.6151   1.6949  25.0328
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.672600   6.151703   4.661 4.34e-06 ***
## CRIM         -0.191246   0.054036  -3.539 0.000450 ***
## ZN            0.044229   0.014111   3.134 0.001854 **
## INDUS         0.055221   0.065532   0.843 0.399944
## CHAS          1.716314   0.891171   1.926 0.054850 .
## NOX         -14.995722   4.557588  -3.290 0.001093 **
## RM            4.887730   0.484947  10.079  < 2e-16 ***
## AGE           0.002609   0.014330   0.182 0.855615
## DIS          -1.294808   0.211724  -6.116 2.36e-09 ***
## RAD           0.484787   0.087347   5.550 5.31e-08 ***
## TAX          -0.015401   0.004447  -3.463 0.000594 ***
## PTRATIO      -0.808795   0.140085  -5.774 1.60e-08 ***
## B            -0.001292   0.006537  -0.198 0.843381
## LSTAT        -0.517954   0.059511  -8.704  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.808 on 386 degrees of freedom
## Multiple R-squared:  0.7339, Adjusted R-squared:  0.7249
## F-statistic: 81.87 on 13 and 386 DF,  p-value: < 2.2e-16
```

Analiza modelu
Na podstawie "t value" określającemu stosunek wartości zmiennej do średniego błedu badam wpływ zmiennych objaśniających na zmienną objaśnianą. Najmniejsze co do modułu
wartości w kolumnie "t value" generują zmienne INDUS, B, AGE oraz CHAS. W oparciu o powyższe rozważania stworzono "minimalny model regresji", usuwając powyższe zmienne objaśniające:

```
min_model <- update(model1, .~. -B -Age -CHAS -INDUS)
summary(min_model)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + LSTAT, data = boston_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2475  -2.5863  -0.5779   1.7020  24.7773
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.700774   5.615939   4.933 1.21e-06 ***
## CRIM         -0.200802   0.053895  -3.726 0.000224 ***
## ZN            0.043275   0.014111   3.067 0.002315 **
## NOX         -13.013414   4.209331  -3.092 0.002134 **
## RM            4.894335   0.478809  10.222  < 2e-16 ***
## AGE           0.003347   0.014351   0.233 0.815715
## DIS          -1.343486   0.206911  -6.493 2.57e-10 ***
## RAD           0.488035   0.083392   5.852 1.03e-08 ***
## TAX          -0.014778   0.004221  -3.501 0.000517 ***
## PTRATIO      -0.808726   0.138123  -5.855 1.01e-08 ***
## LSTAT        -0.517478   0.059210  -8.740  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.82 on 389 degrees of freedom
## Multiple R-squared:  0.7304, Adjusted R-squared:  0.7235
## F-statistic: 105.4 on 10 and 389 DF,  p-value: < 2.2e-16
```

Współczynnik R^2 spadł o:

```
summary(model1)$r.squared-summary(min_model)$r.squared
```

```
## [1] 0.003408593
```

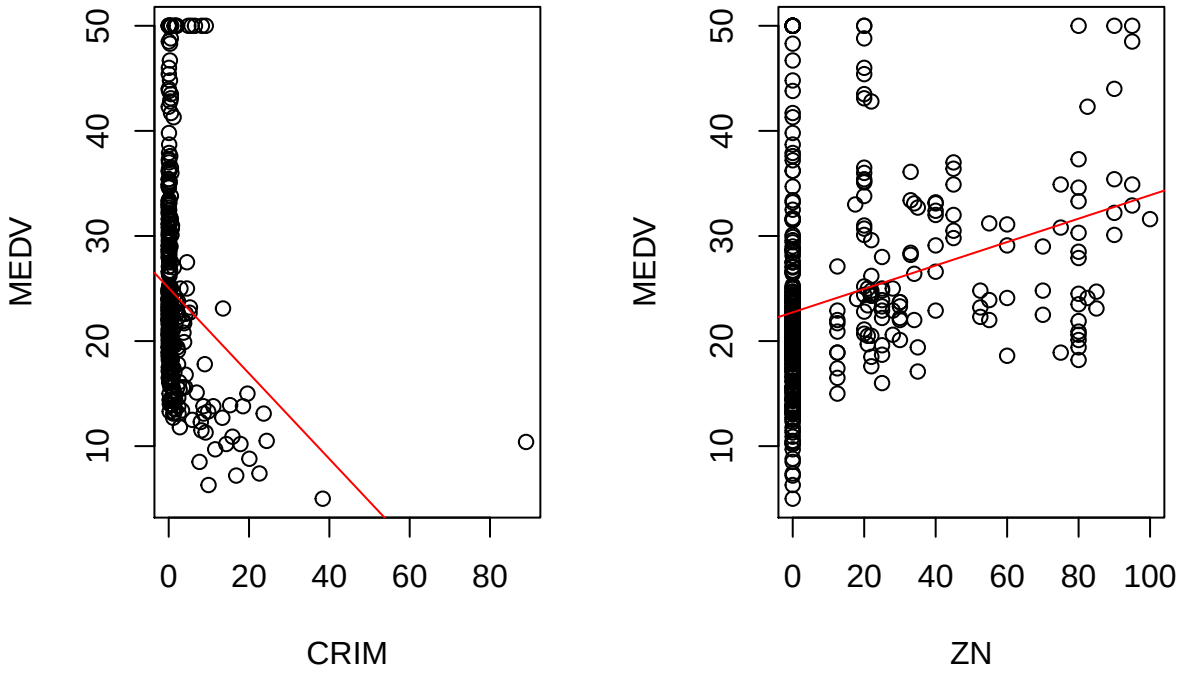
Współczynnik adjusted R^2 spadł o:

```
summary(model1)$adj.r.squared-summary(min_model)$adj.r.squared
```

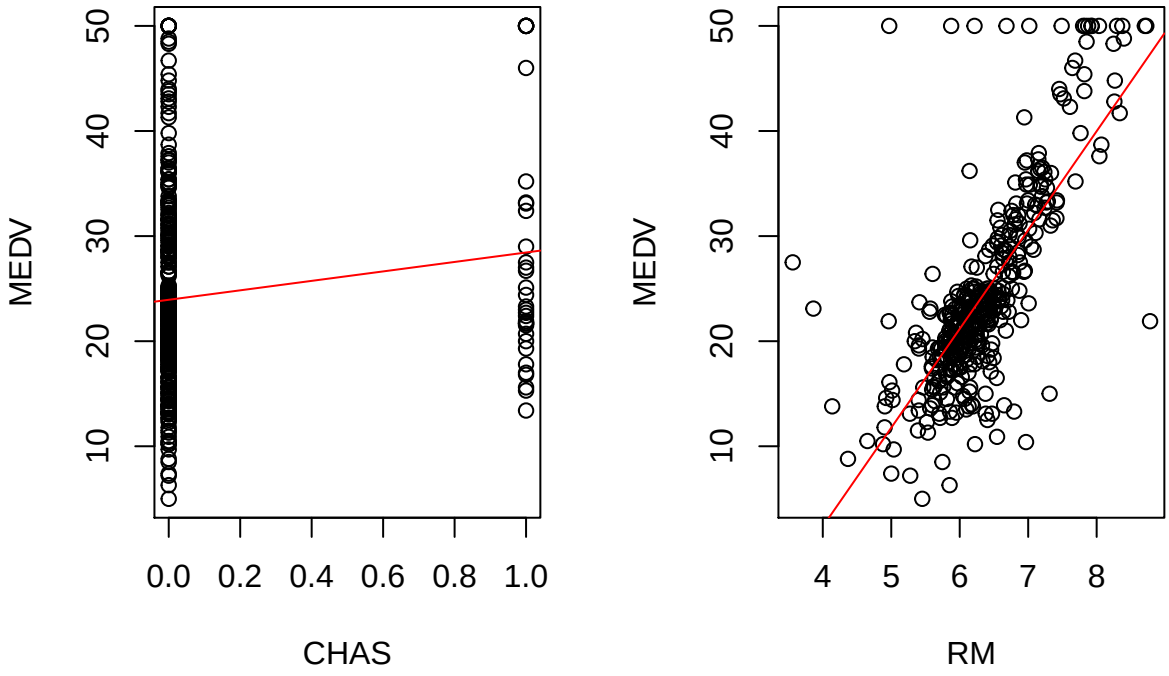
```
## [1] 0.00137452
```

Wykresy punktowe zmiennej objaśnianej dla modelu minimalnego wraz z prostą regresji.

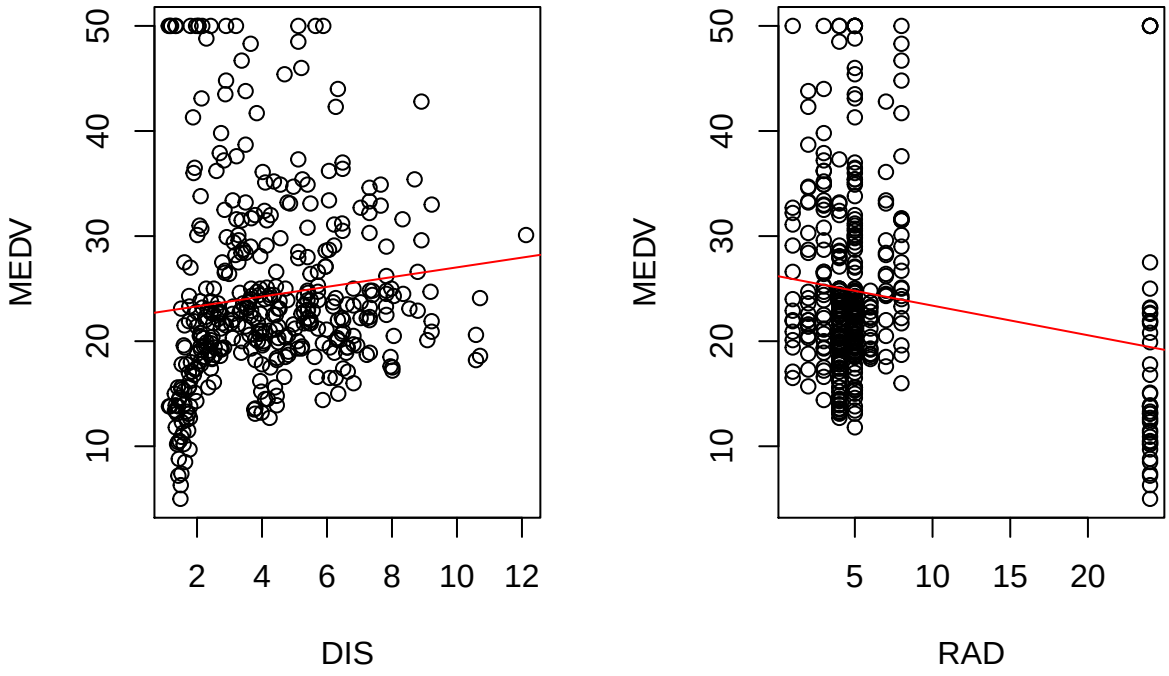
```
par(mfrow=c(1,2))
plot(boston_data1$CRIM,boston_data1$MEDV, xlab="CRIM", ylab="MEDV");
abline(lm(MEDV~CRIM, data=boston_data1),col="red")
plot(boston_data1$ZN,boston_data1$MEDV, xlab="ZN", ylab="MEDV");
abline(lm(MEDV~ZN, data=boston_data1),col="red")
```



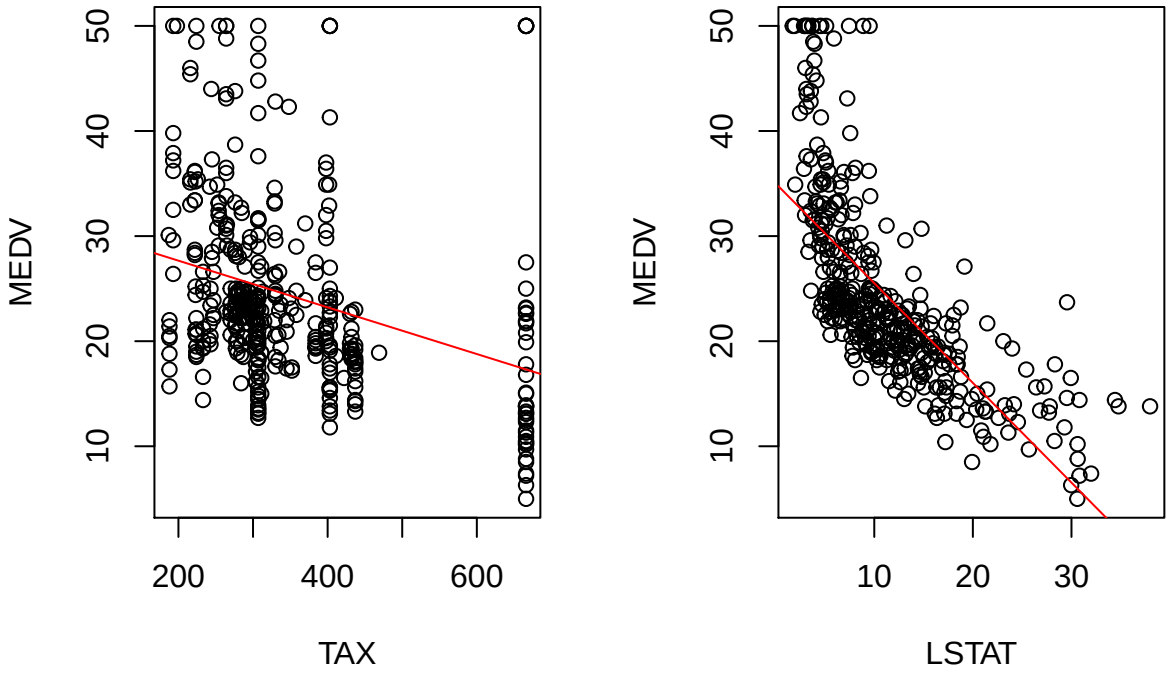
```
par(mfrow=c(1,2))
plot(boston_data1$CHAS,boston_data1$MEDV, xlab="CHAS", ylab="MEDV");
abline(lm(MEDV~CHAS, data=boston_data1),col="red")
plot(boston_data1$RM,boston_data1$MEDV, xlab="RM", ylab="MEDV");
abline(lm(MEDV~RM, data=boston_data1),col="red")
```



```r
par(mfrow=c(1,2))
plot(boston_data1$DIS,boston_data1$MEDV, xlab="DIS", ylab="MEDV");
abline(lm(MEDV~DIS, data=boston_data1),col="red")
plot(boston_data1$RAD,boston_data1$MEDV, xlab="RAD", ylab="MEDV");
abline(lm(MEDV~RAD, data=boston_data1),col="red")
```

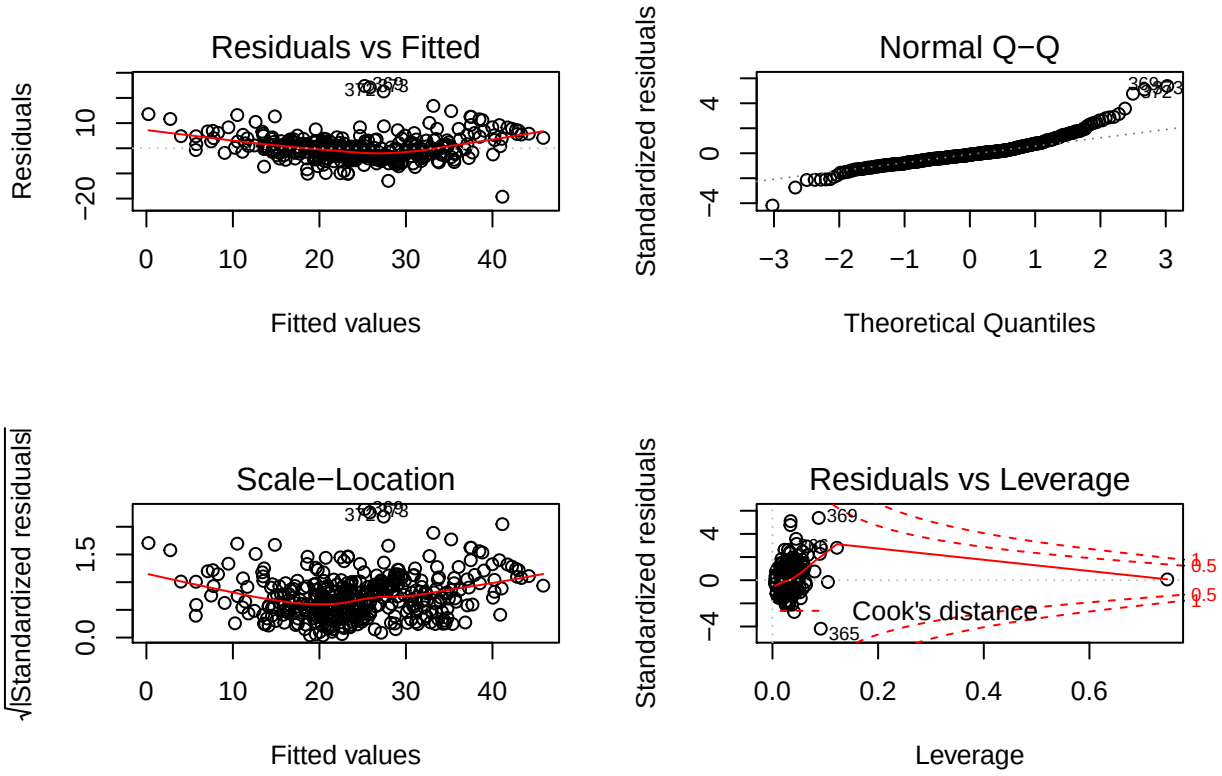


```
par(mfrow=c(1,2))
plot(boston_data1$TAX,boston_data1$MEDV, xlab="TAX", ylab="MEDV");
abline(lm(MEDV~TAX, data=boston_data1),col="red")
plot(boston_data1$LSTAT,boston_data1$MEDV, xlab="LSTAT", ylab="MEDV");
abline(lm(MEDV~LSTAT, data=boston_data1),col="red")
```



Wykres wielkości residuów od dopasowanej wartości.

```
par(mfrow=c(2,2));plot(min_model)
```



Wykres Residuals vs Fitted pozwala wnioskować, że regresja liniowa oraz dobór zmiennych objaśniających został wykonany prawidłowo ponieważ, wartości reszt są bliskie 0, a czerwona krzywa jest podobna do linii prostej o wartości zbliżonej do zera. Pozwala to stwierdzić ,że zależnośc jest liniowa. Wykres Normal Q-Q to wykres kwantylowy dla rozkładu normalnego w przedziale [-2;2] bardzo dokładnie pokrywa się z wykropkowaną linią, co oznacza ,że spełniony jes warunek normalności układu reszt. Wykres Scale-Location pozwala sprawdzić czy spełnione jest założenie o jednorodnej wariancji. Wykres Residuals vs Leverage ułatwia wykrycie obserwacji wpływowych. Na powyższym wykresie każda obserwacja jest wewnątrz obszaru ograniczonego przez przerywane czerwone linie. Gdyby jednak któraś z nich nie spełniała powyższego warunku należałoby się zastanowić nad jej usunięciem z modelu.

Ad.4 Model, a mniejszy zbiór danych.

```
prediction <- predict(min_model,boston_data2)
error <- abs(prediction-boston_data2$MEDV)
mean(error)
```

```
## [1] 5.212451
```