Projekt Ekonometria - analiza danych w R,
Monika Topolska, Kamil Bartocha
WMS 2021/2022

Przygotowanie danych - opis danych.

Dane znajdują się w bibliotece mlbench i zawierają dane mieszkań w Bostonie *BostonHousing.* Dla każdej z 506 obserwacji dane zawierają 14 zmiennych:
CRIM - wskaźnik przestępczości na mieszkańca według miast.
ZN - część gruntów mieszkalnych przeznaczonych na działki o powierzchni powyżej 25 000 stóp kwadratowych.
INDUS - odsetek akrów działalności niehandlowej na miasto.
CHAS - Zmienna fikcyjna Charles River.
(1, jeśli droga ogranicza rzekę; 0 w przeciwnym razie).
NOX - stężenie tlenków azotu (ilość na 10 mln).
RM - średnia liczba pokoi na mieszkanie.
AGE - odsetek mieszkań własnościowych wybudowanych przed 1940.
DIS - odległości do pięciu bostońskich centrów zatrudnienia.
RAD - wskaźnik dostępności do autostrad.
TAX - pełnowartościowa stawka podatku od nieruchomości za 10 000$.
PTRATIO - stosunek uczniów do nauczycieli.
B - 1000(Bk - 0.63)^2 gdzie Bk to odsetek czarnoskórych według miasta.
LSTAT - % niższego statusu ludności
MEDV - Średnia wartość domów zajmowanych przez właścicieli wyrażona jako 1000$
Dane zostały zebrane przez US Census Service dotyczących mieszkalnictwa w rejonie Boston Mass.
https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html
Uwaga.

Zmienna nr 14 wydaje się być ocenzurowana na 50 (co odpowiada średniej cenie 50 000 USD); Cenzurę sugeruje fakt, że najwyższa mediana ceny, wynosząca dokładnie 50 000 USD, została odnotowana w 16 przypadkach, podczas gdy 15 przypadków ma ceny od 40 000 do 50 000 USD, z cenami zaokrąglonymi do najbliższej setki. Harrison i Rubinfeld nie wspominają o żadnej cenzurze.

Przygotowanie danych - cel projektu.

Zmienna zależną jest MEDV czyli średnia wartość domów. Celem projektu jest zbadanie, które zmienne wpływają na cenę oraz sprawdzenie w jakim stopniu wpływają one na zmienną zależną.

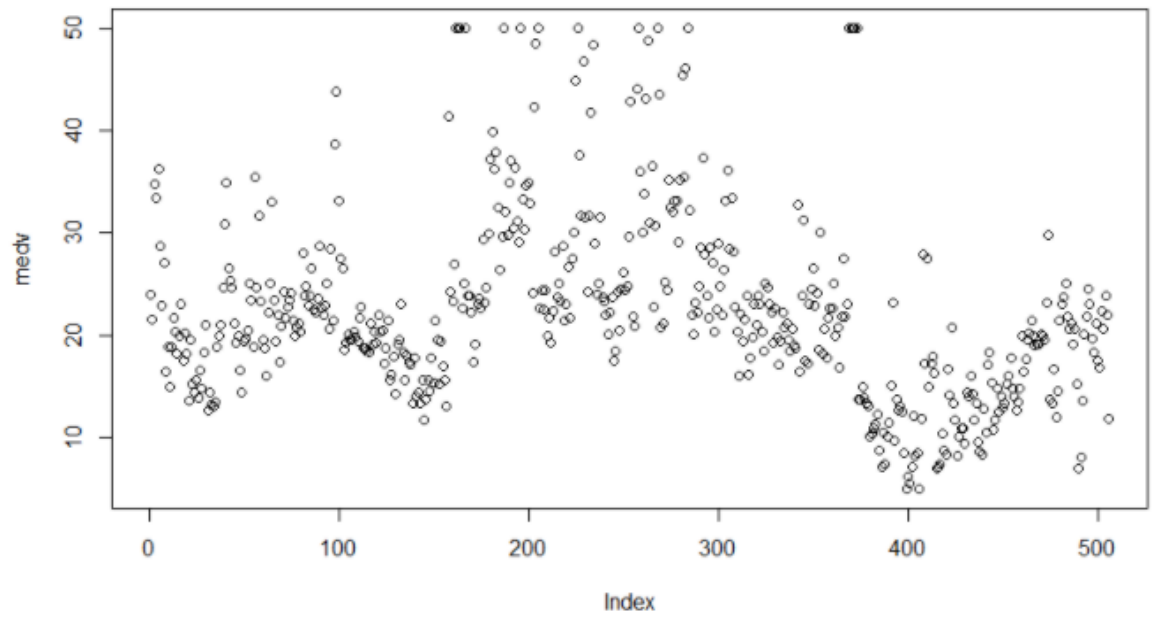
Przygotowanie danych - przegląd danych i wykresy zależności.

Przegląd danych:

Na początku przyjrzymy się podstawowym statystykom. Jak widzimy poniżej, nie ma w danych obserwacji brakujących (N/A). Wydaje się, że musimy się bardziej przyjrzeć zmiennej chas jako, że jest ona dyskretna i przyjmuje jedynie wartości 0 i 1.

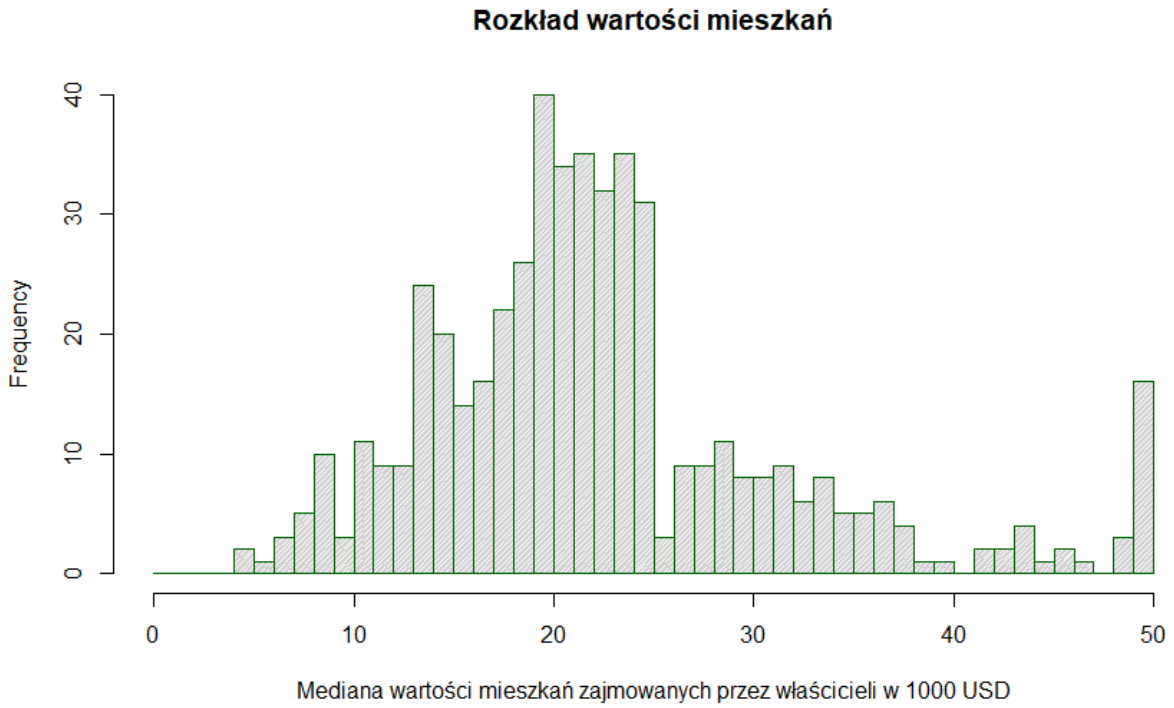
```
      crim                zn             indus        chas          nox
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471   Min.   :0.3850
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.4490
 Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.5380
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240
 Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710
       rm              age              dis              rad              tax
 Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0
 1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0
 Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000   Median :330.0
 Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549   Mean   :408.2
 3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0
 Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0
    ptratio            b              lstat            medv
 Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00
 1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02
 Median :19.05   Median :391.44   Median :11.36   Median :21.20
 Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53
 3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00
 Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00
```

Wykres zmiennej zależnej medv:

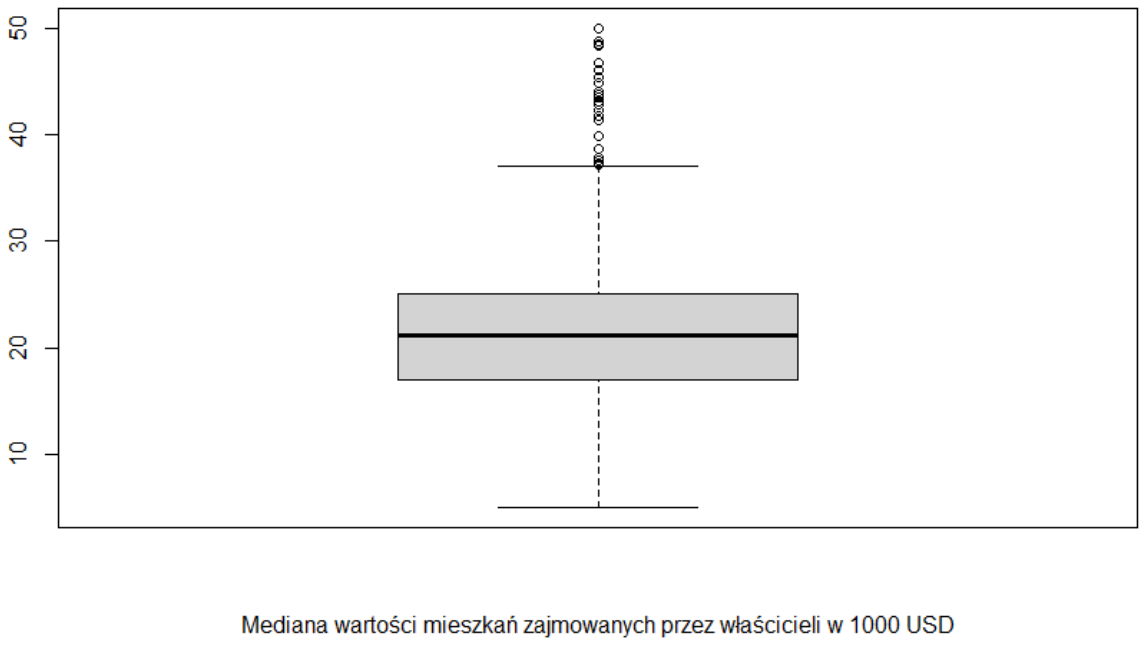


Na wykresie widocznych jest kilka obserwacji odstających o wysokich wartościach. Są to wyżej wspomniane wartości, które wydają się ocenzurowane (zakres wartości zmiennej zależnej urywa się na dokładnie 50 tys USD- takich obserwacji jest 16 na 506 wszystkich).

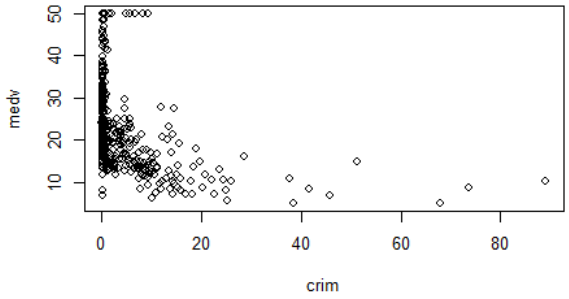
Na kolejnym wykresie przedstawiającym histogram zmiennej zależnej również obserwacje te rzucają się w oczy. Poza tym rozkład obserwacji wydaje się dość symetryczny.

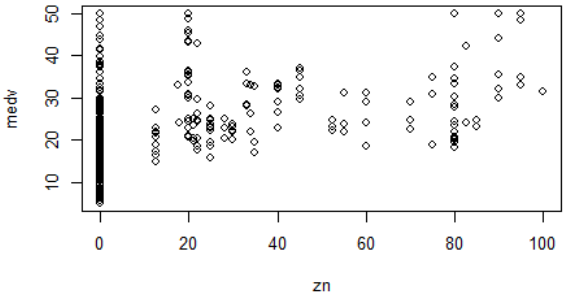


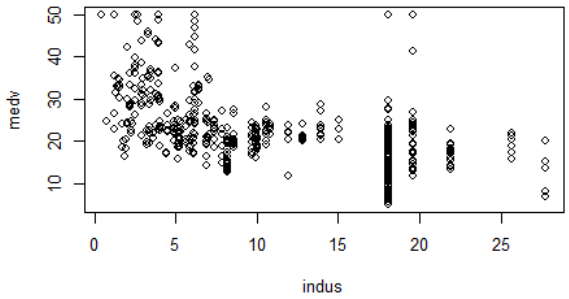
Potwierdzimy nasze przypuszczenia wykresem pudełkowym:

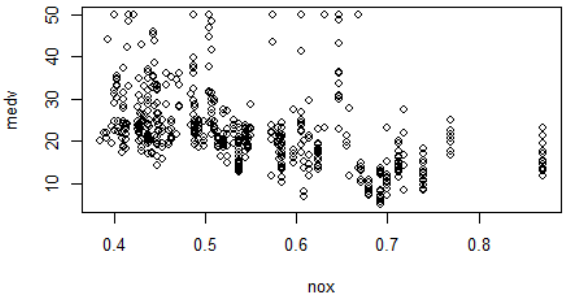


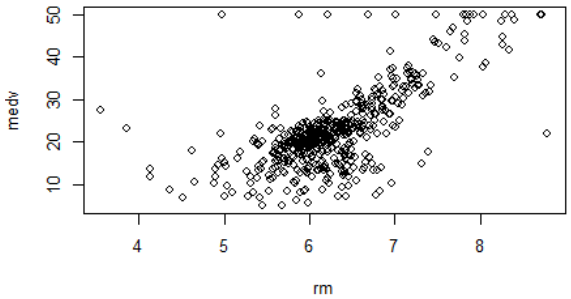
Następnie przedstawimy kolejne wykresy zależności zmiennej Y (medv) od zmiennych niezależnych:

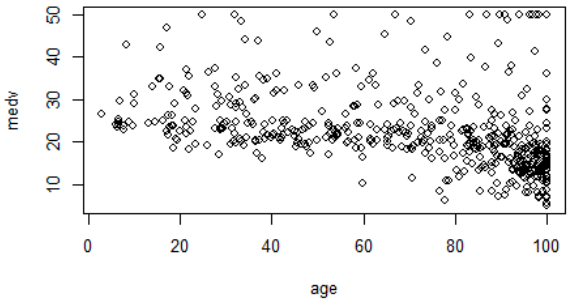


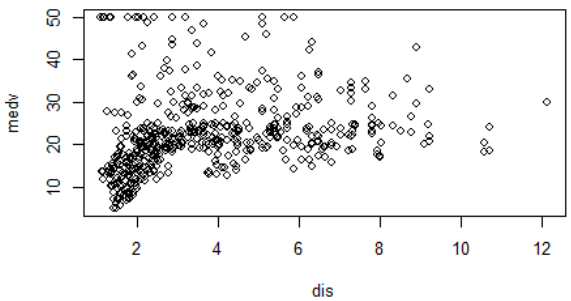


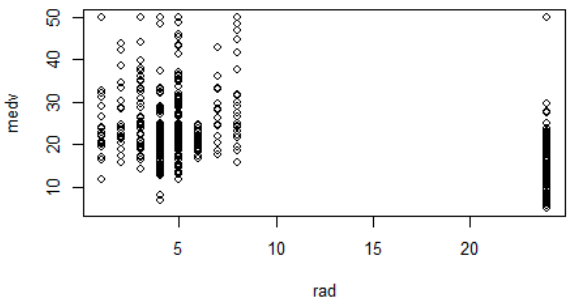


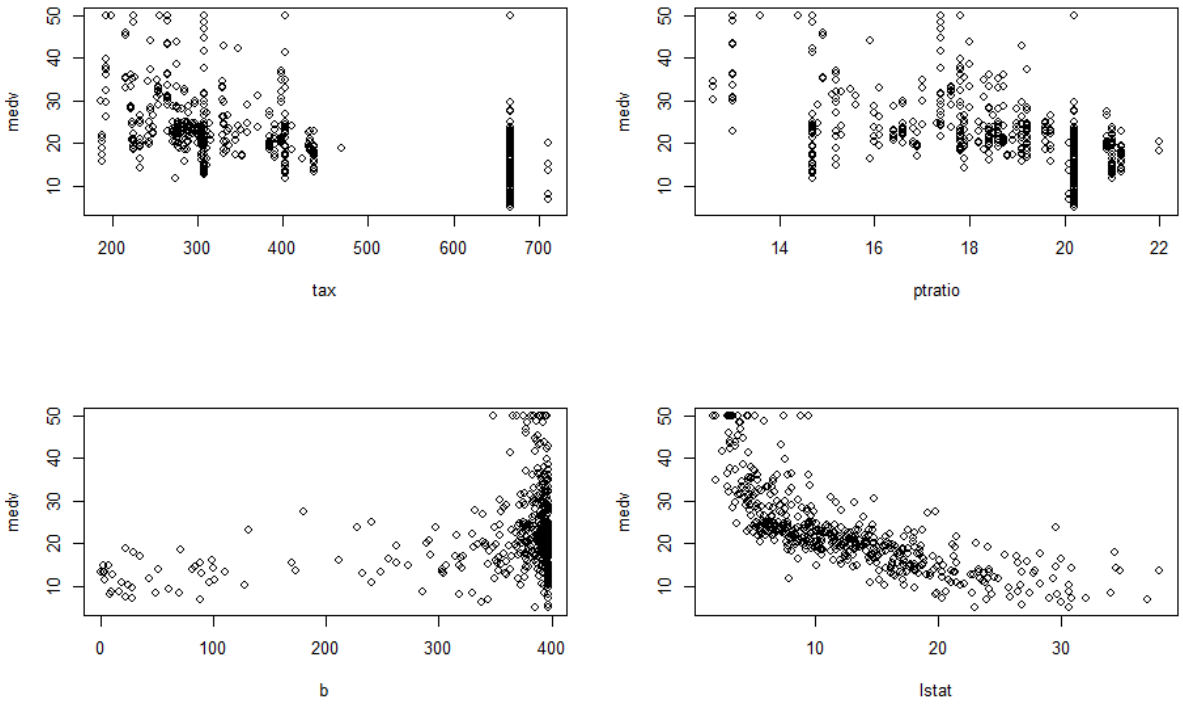






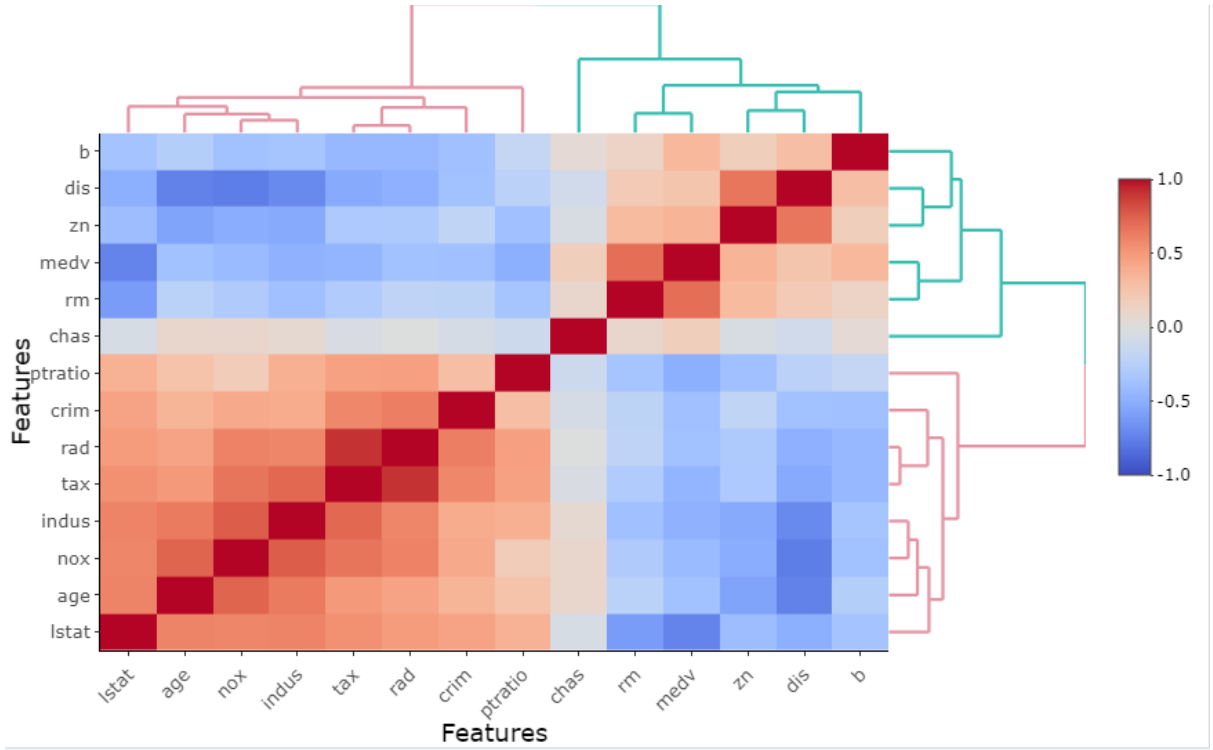






Zmienna Y na tych wykresach wydaje się dość losowo rozrzucona, aczkolwiek warto zauważyć zależność na wykresie medv~lstat (widzimy wyraźny łuk) oraz medv~rm (mocna wzrostowa zależność).

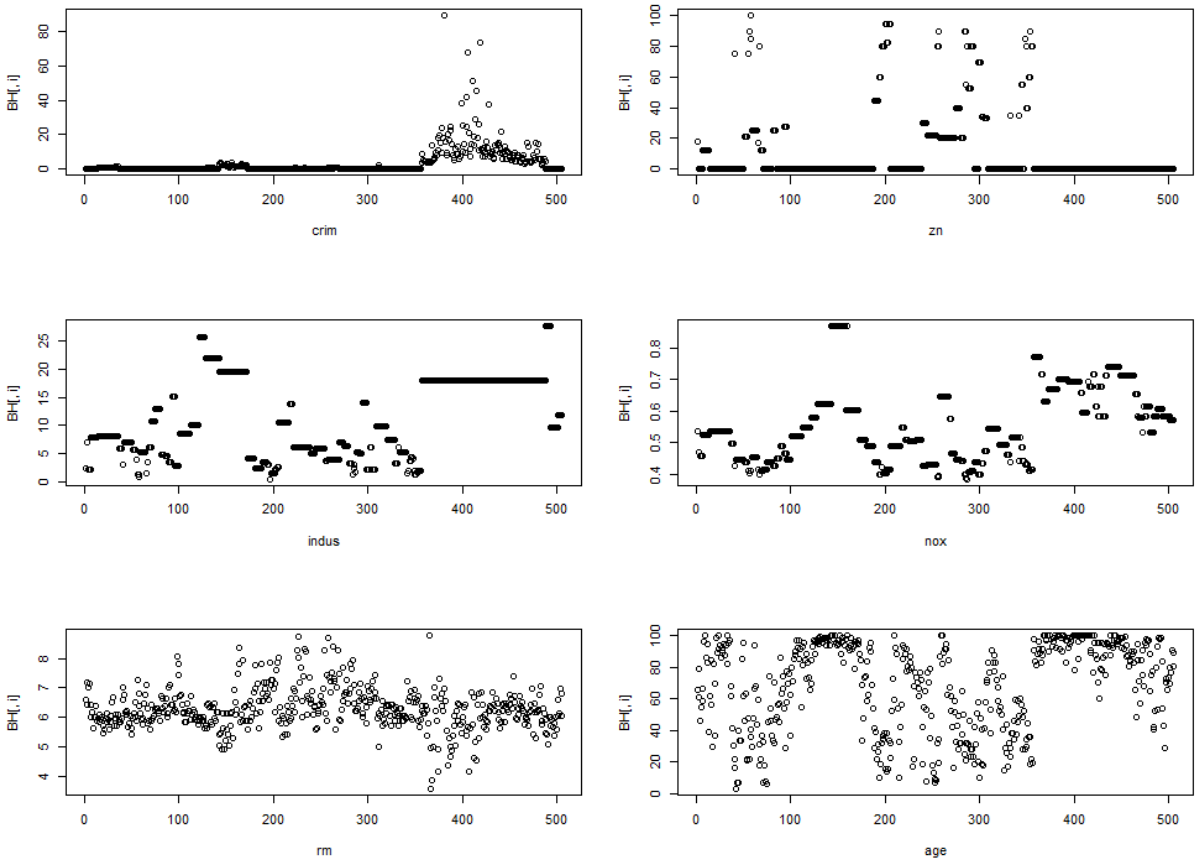
Aby potwierdzić nasze przypuszczenia dotyczące zależności tworzymy obrazową macierz korelacji:



Jak widzimy najsilniejszą korelację ze zmienną zależną ma zmienna rm (0.695) oraz lstat (-0.738). Możemy przypuszczać, że te dwie zmienne będą miały największy wpływ na średnią wartość mieszkań.

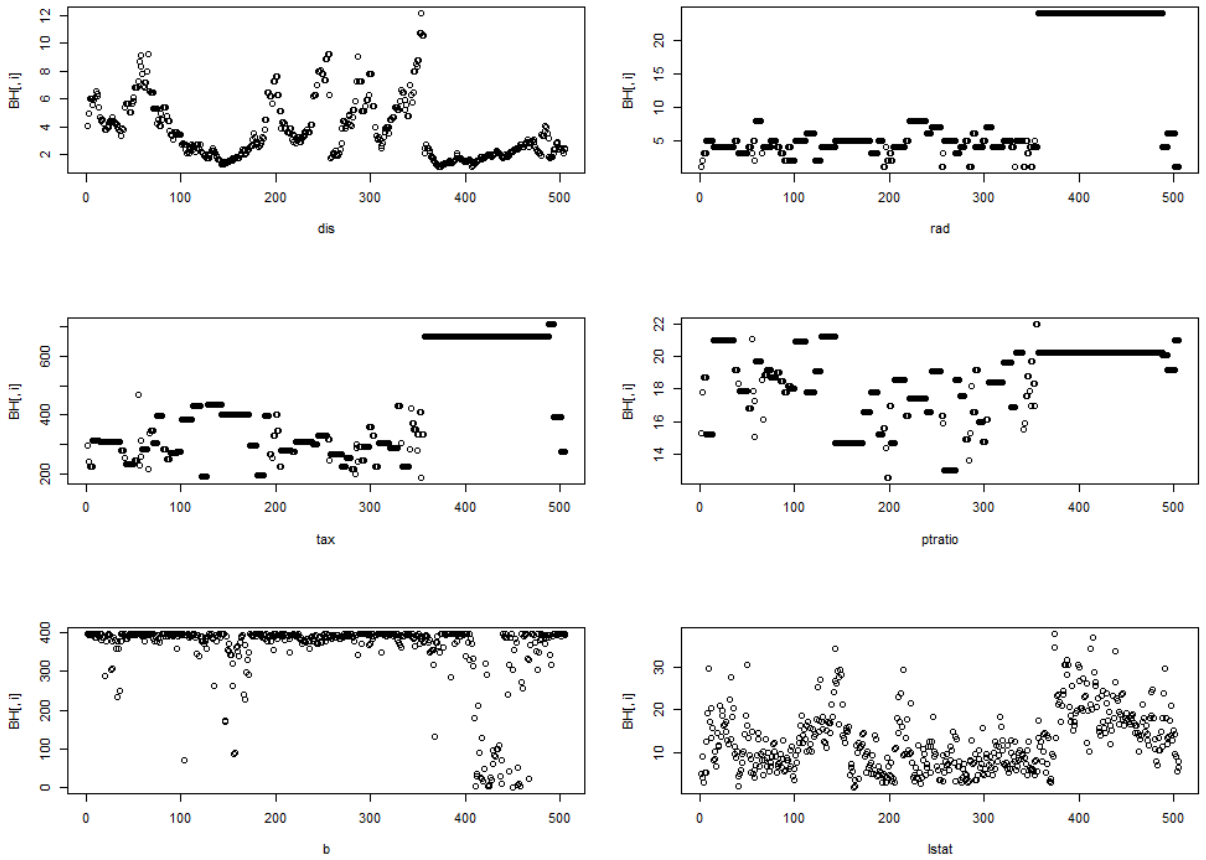
Tak jak wspomnieliśmy na początku, problematyczną jest zmienna chas jako, że jest dyskretna - postanawiamy nie uwzględniać jej w dalszej analizie. Podobnie, mocno podejrzanymi wydają się obserwacje zmiennej zależnej medv o wartości równej 50 tysięcy. Jest wysoce nieprawdopodobne by były to obserwacje naturalne - wartość ta mocno odstaje, co również było widać na wszystkich powyższych wykresach. Postanawiamy zatem usunąć te obserwacje, jako, że nie jest to duża część ogółu obserwacji, a mogłyby spowodować duże odchylenie linii regresji.

Przedstawimy teraz poszczególne wykresy kolejnych zmiennych, by przyjrzeć się rozkładowi tych obserwacji:

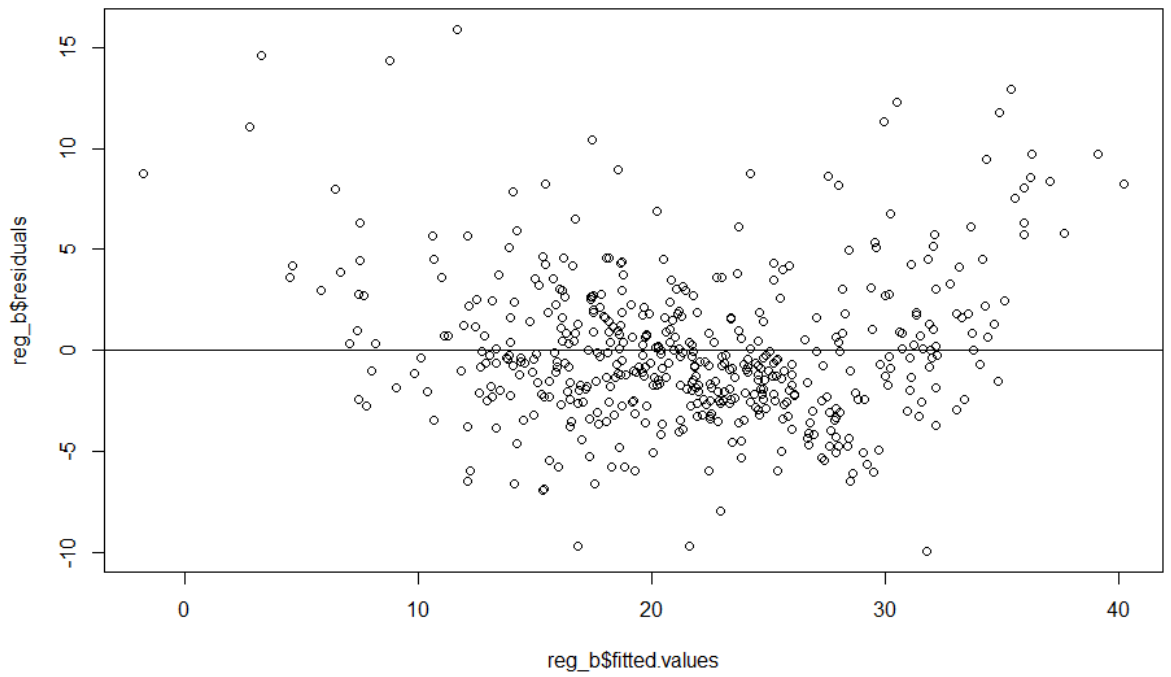


Jak widzimy powyżej zmienne crim - wskaźnik przestępczości, zn - udział działek mieszkalnych przeznaczonych dla działek powyżej 25tys stóp kwadratowych , indus - odsetek akrów działalności nie detalicznej na miasto i nox - stężenie tlenku azotu, przyjmują wartości mocno skokowo. To na pewno odbijać się będzie na naszej analizie. Jedynie zmienna rm - średnia liczba pokoi na mieszkanie i age - odsetek jednostek zamieszkanych przez właściciela wybudowanych przed 1940 r, przyjmują wartości dość losowo rozrzucone.

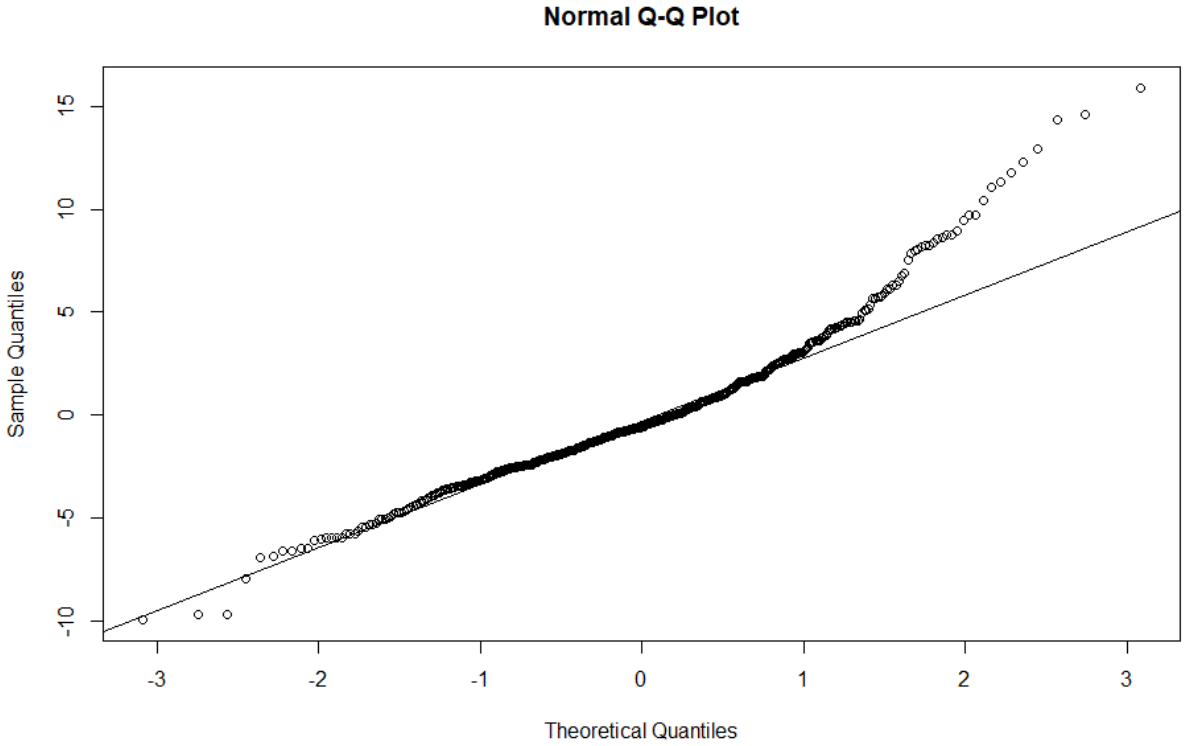
Kolejne wykresy zmiennych:



Tutaj również rad - wskaźnik dostępności do autostrad, tax – pełnowartościowa stawka podatku od nieruchomości na 10 000 USD i ptratio - stosunek uczniów do nauczycieli według miasta, przyjmują wartości skokowo. Pozostałe zmienne są ułożone bardziej losowo.
Następnie przechodzimy do analizy modelu regresji liniowej
Wykres residuów od estymowanej zmiennej medv w pełnym modelu:

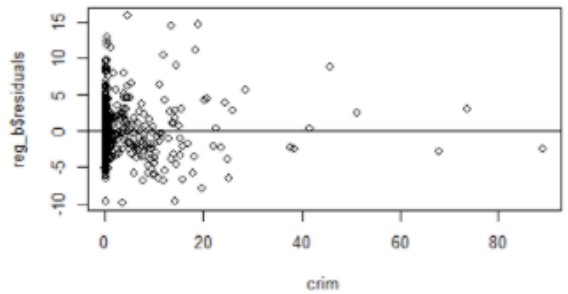


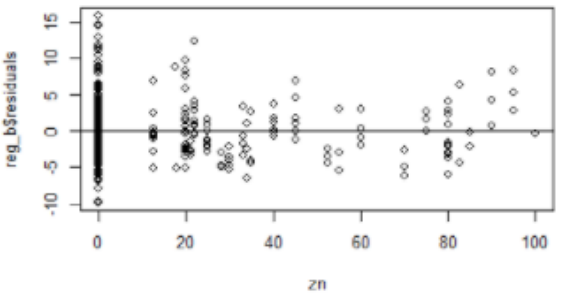
Jak widzimy, możliwe jest zauważenie pewnej zależności układającej się w łuk, prawdopodobnie odpowiednia transformacja poprawi rozmieszczenie obserwacji.
Wykres qqnorm reszt:

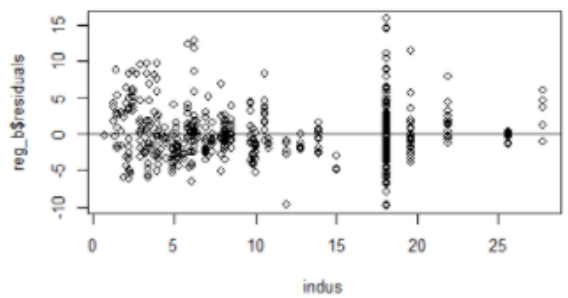


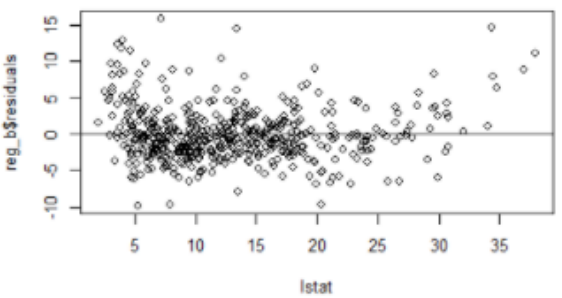
Wykres kwantyli residuów wyraźnie odstaje od linii, szczególnie w końcowych wartościach - to również sugeruje nam potrzebę transformacji, gdyż ich rozkład odbiega od rozkładu normalnego.

Następnie prezentujemy wykresy reszt od poszczególnych zmiennych niezależnych:

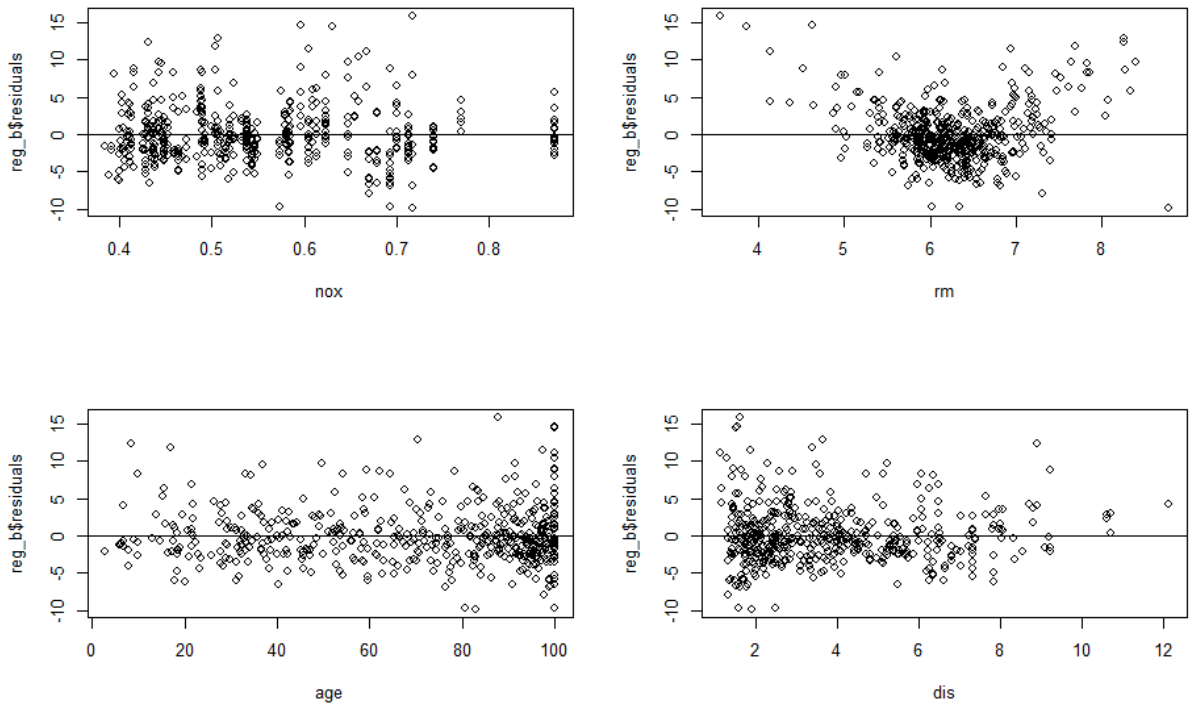




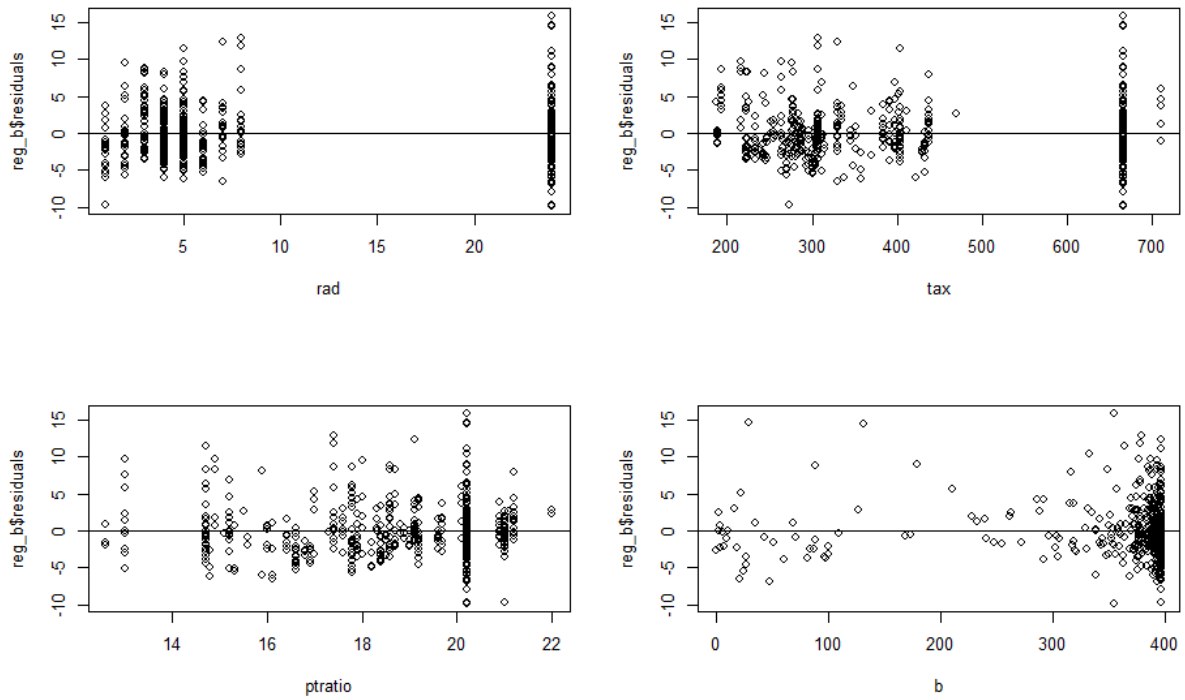




Wartości reszt przedstawiają się raczej dość losowo. Jedynie przy zmiennej lstat obserwujemy może pewną zależność.



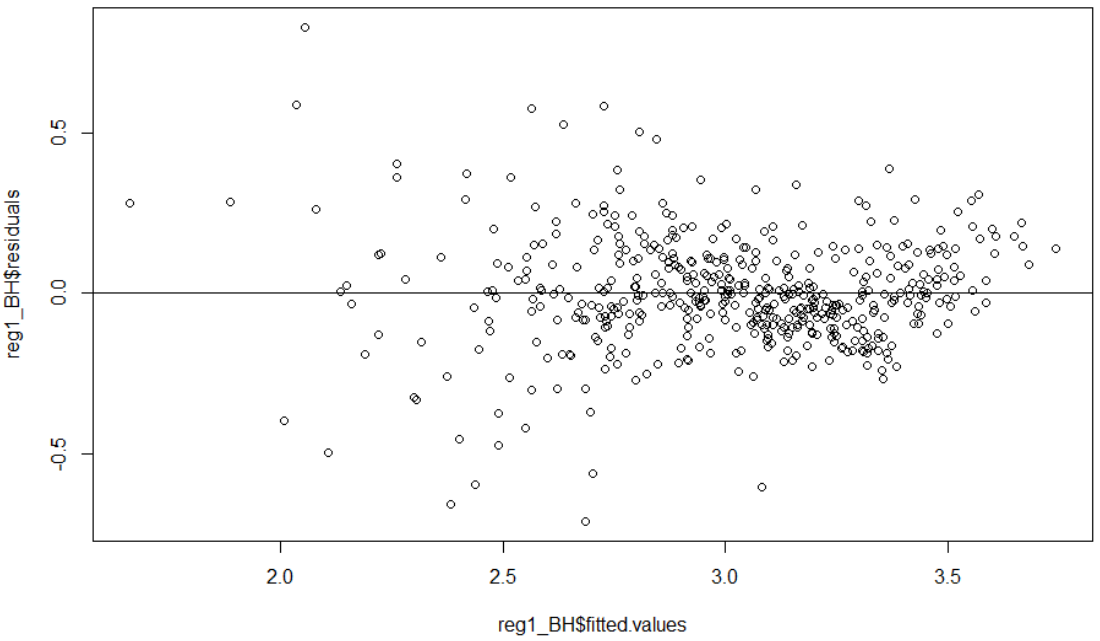
Także tutaj jedynie przy zmiennej rm możemy zaobserwować dość wyraźny kształt paraboli - sugeruje nam to potrzebę transformacji zmiennej.

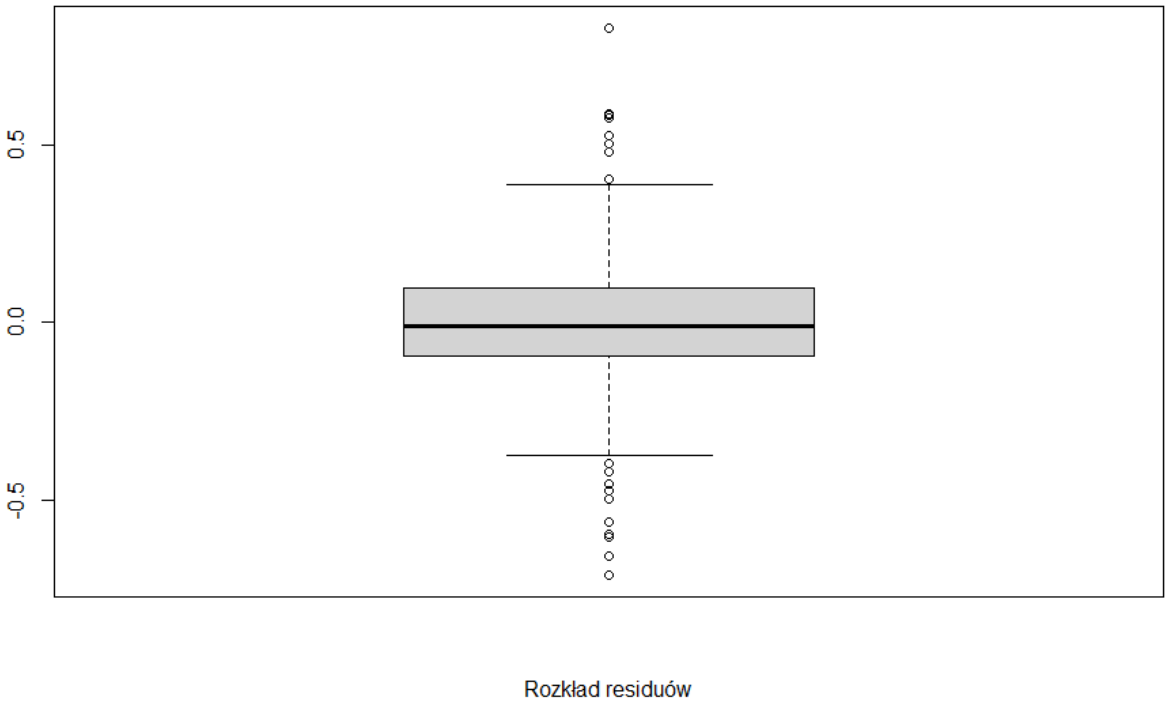


Rozkład obserwacji jest symetryczny ale widać obserwacje odstające, zwłaszcza powyżej średniej wartości.

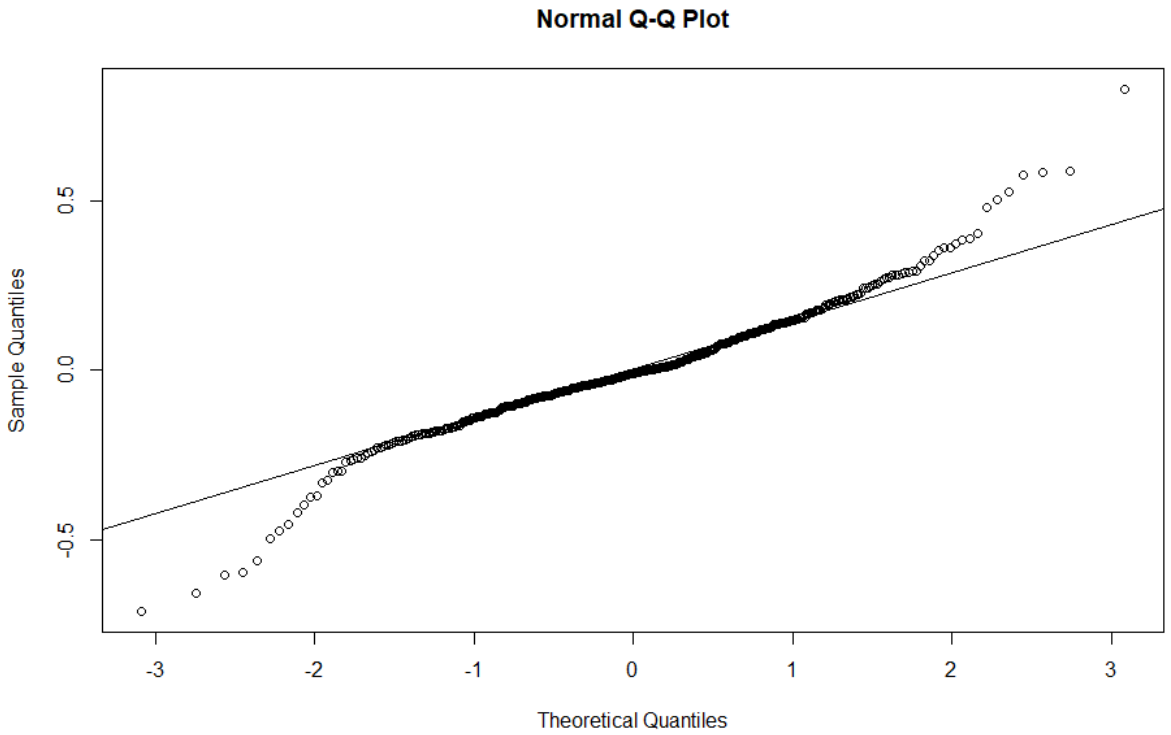
Transformacje zmiennych

Początkowo rozpoczynamy od transformacji logarytmicznej Y i przeprowadzamy ponowną regresję z tą zmienną. Na poniższym wykresie reszt od estymowanego Y widać, że po transformacji dane nie przybierają już parabolicznego kształtu. Jednak możemy podejrzewać nie stałą wariancję, ponieważ jak widzimy obserwacje po lewej stronie są o wiele bardziej rozrzucone niż te po prawej. Próbujemy kolejnych transformacji.

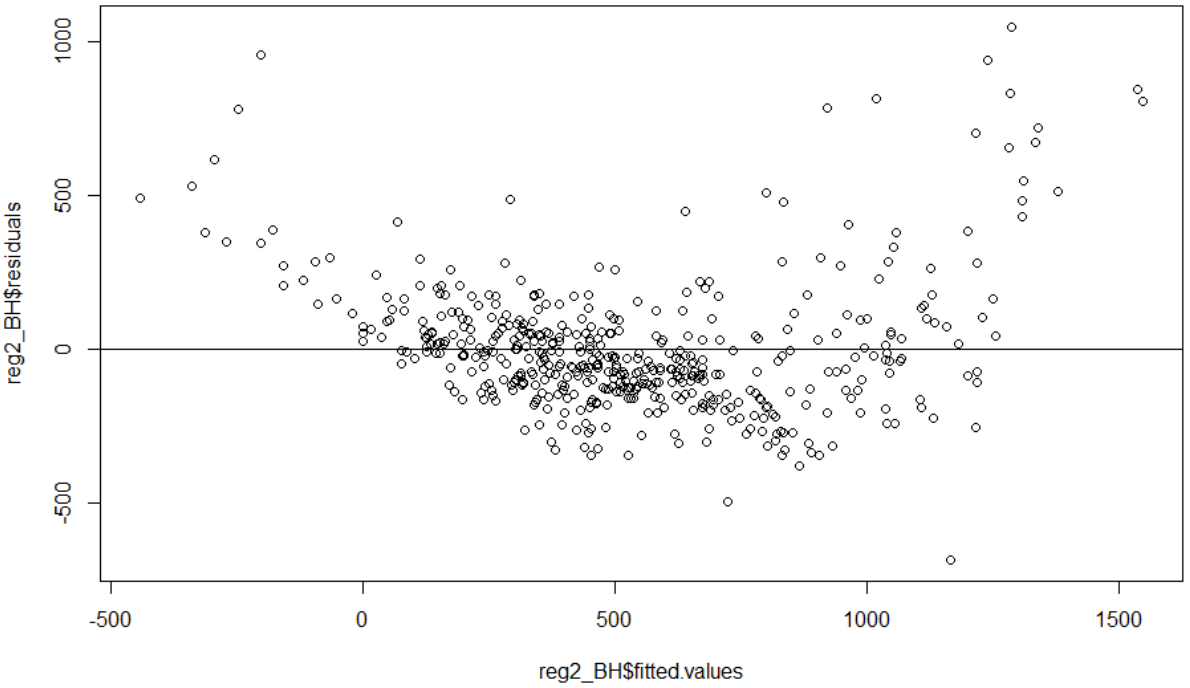


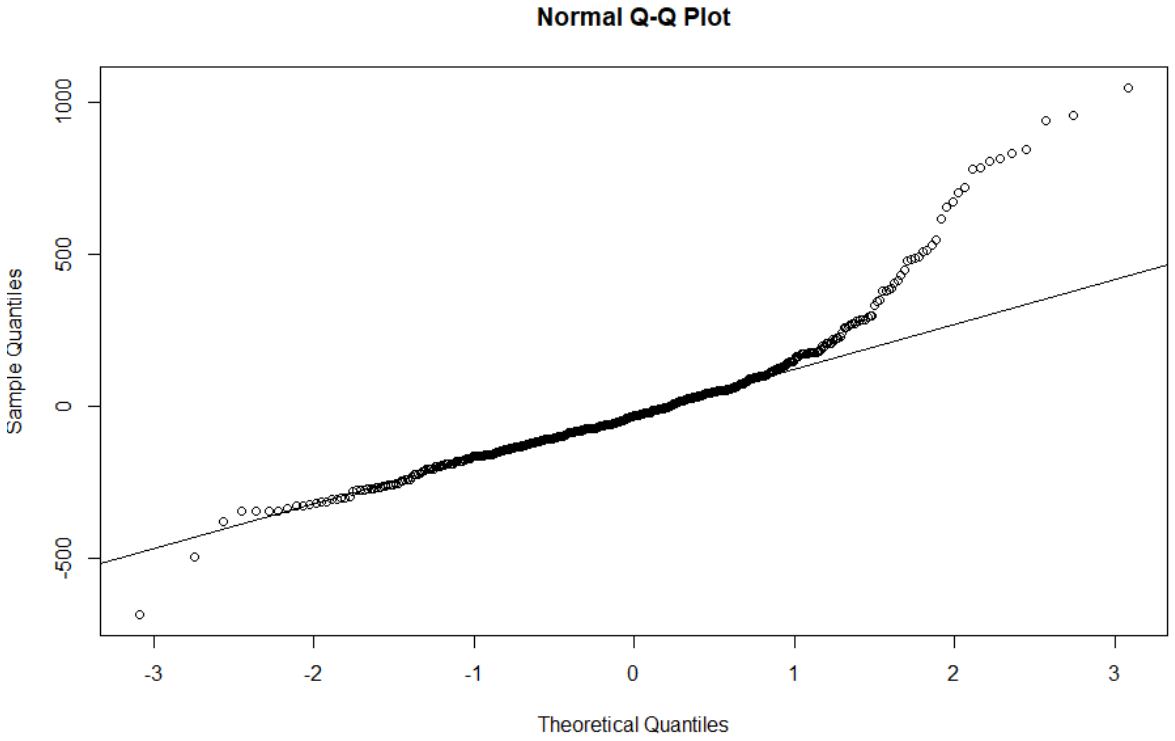


Wykres pudełkowy residuów również się zmienił - nadal mamy kilkanaście outlierów, lecz są one rozłożone symetrycznie, równomiernie są to outliery lewo- i prawostronne.

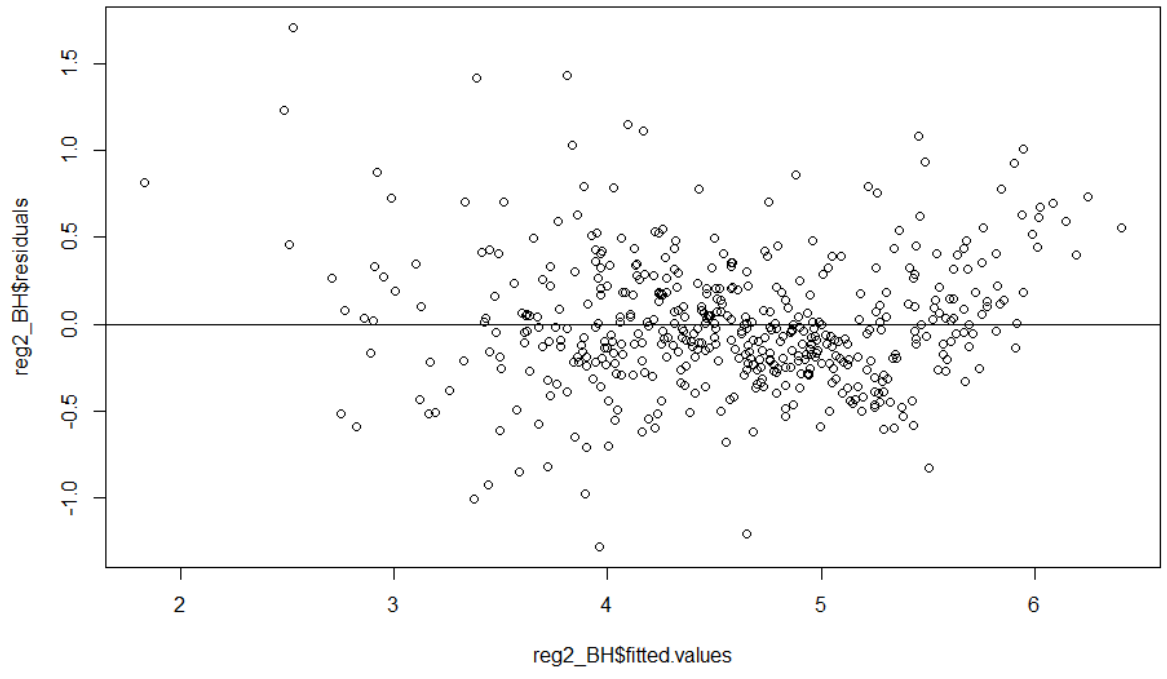


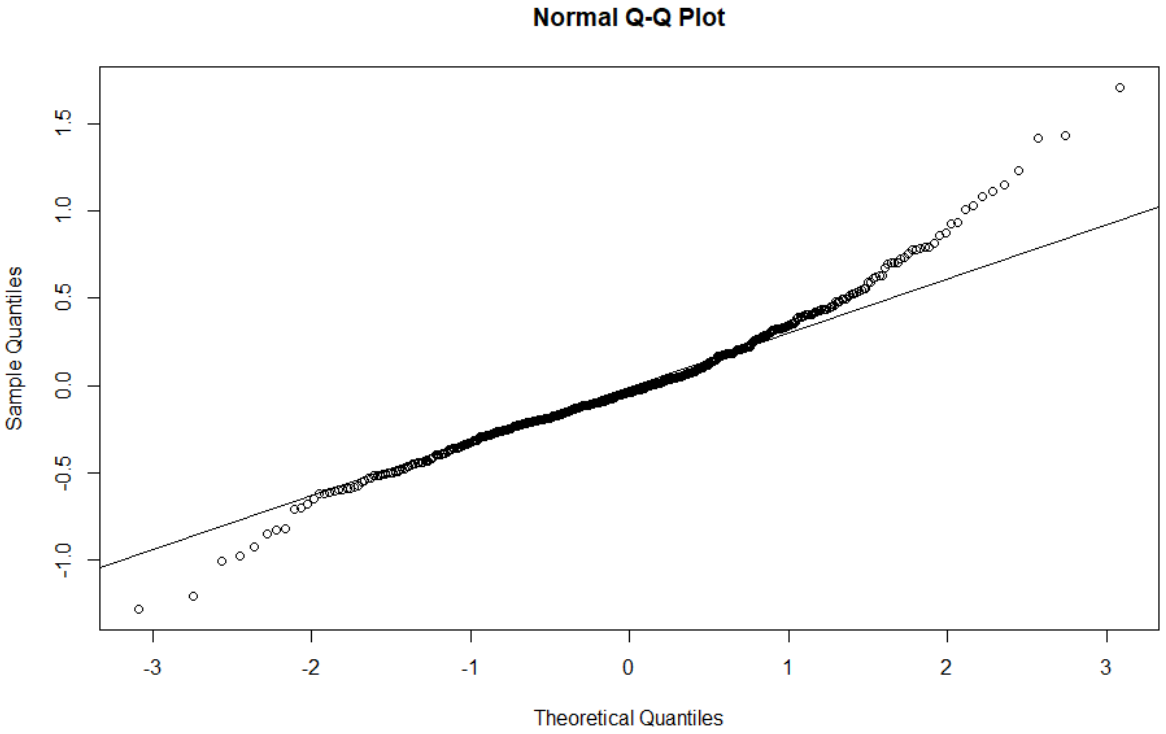
Wykres qqnorm również istotnie się zmienił, jego oba końce odchodzą teraz od linii qqline.
Kolejna transformacja - potęgowanie - przyniosła tylko gorsze efekty





Kolejna transformacja - pierwiastkowanie zmiennej już zlogarytmowanej nie przynosi zadowalających rezultatów, kolejno próbujemy pierwiastkować zmienną zależną oryginalną medv. Efektem są następujące wykresy reszt:





Naszym wnioskiem z kolejnych transformacji jest to, że pierwiastkowanie przynosi najlepsze rezultaty. Choć na pewno residua nie mają jak widzimy rozkładu normalnego, to wydają się one rozłożone niezależnie od wartości estymowanego Y.

Wybór zmiennych modelu:

Startując z modelu pełnego, używając metody automatycznego wyboru zmiennych “backward” otrzymujemy jako najlepszy model z usuniętą zmienną indus - odsetek akrów działalności niehandlowej na miasto. Parametry tego modelu są jak niżej:

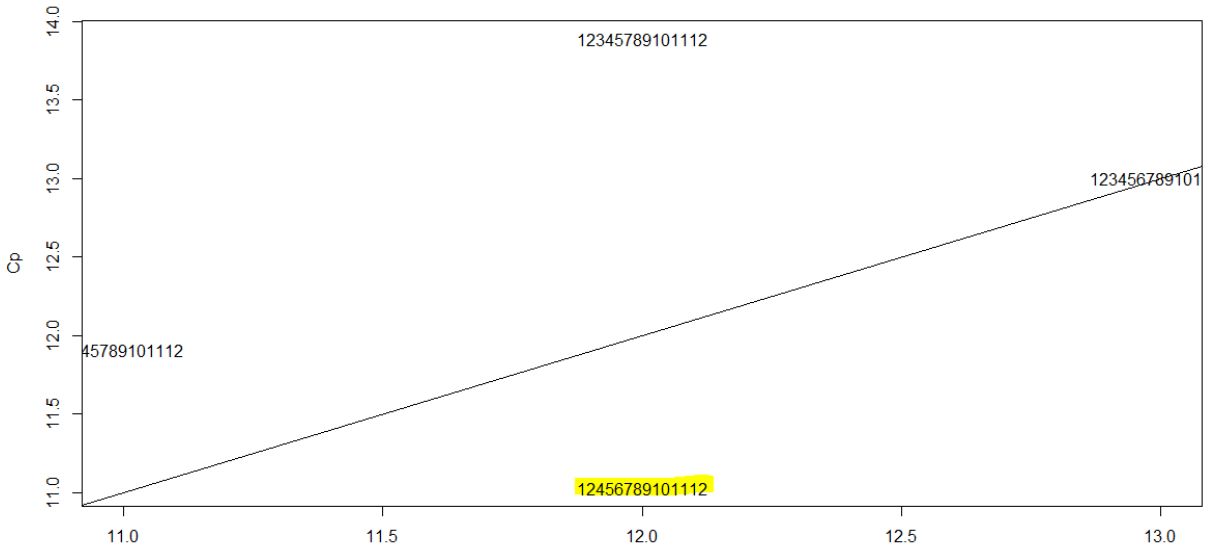
```
Step:  AIC=-924.35
sq_medv ~ crim + zn + nox + rm + age + dis + rad + tax + ptratio +
    b + lstat

          Df Sum of Sq    RSS     AIC
<none>                  70.739 -924.35
- age      1    0.4270 71.166 -923.40
- zn       1    1.0014 71.741 -919.46
- b        1    2.3652 73.105 -910.23
- nox      1    3.2217 73.961 -904.52
- rad      1    4.3513 75.091 -897.10
- tax      1    4.6116 75.351 -895.40
- crim     1    5.7101 76.450 -888.31
- dis      1    7.3446 78.084 -877.94
- ptratio  1    9.9201 80.660 -862.04
- rm       1   10.8299 81.569 -856.55
- lstat    1   16.5699 87.309 -823.22

Call:
lm(formula = sq_medv ~ crim + zn + nox + rm + age + dis + rad +
    tax + ptratio + b + lstat, data = BH2)

Coefficients:
(Intercept)         crim           zn          nox           rm
  6.2493595   -0.0165515    0.0029862   -1.3974457    0.3111721
        age          dis          rad          tax      ptratio
 -0.0018508   -0.1131425    0.0283561   -0.0015480   -0.0870257
          b        lstat
  0.0008706   -0.0458244
```

Summary modelu:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.27867 -0.22016 -0.04006  0.19828  1.70612

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.2493595  0.4213038  14.833  < 2e-16 ***
crim        -0.0165515  0.0026646  -6.212 1.14e-09 ***
zn           0.0029862  0.0011480   2.601  0.00958 **
nox         -1.3974457  0.2995102  -4.666 4.00e-06 ***
rm           0.3111721  0.0363751   8.555  < 2e-16 ***
age         -0.0018508  0.0010896  -1.699  0.09005 .
dis         -0.1131425  0.0160604  -7.045 6.51e-12 ***
rad          0.0283561  0.0052294   5.422 9.35e-08 ***
tax         -0.0015480  0.0002773  -5.582 3.99e-08 ***
ptratio     -0.0870257  0.0106293  -8.187 2.45e-15 ***
b            0.0008706  0.0002178   3.998 7.40e-05 ***
lstat       -0.0458244  0.0043307 -10.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3847 on 478 degrees of freedom
Multiple R-squared:  0.7956,    Adjusted R-squared:  0.7909
F-statistic: 169.2 on 11 and 478 DF,  p-value: < 2.2e-16
```

Jak widzimy zmienna age nadal wydaje się być nieistotna statystycznie. Gdy również ją usuniemy, R kwadrat pogarsza się jedynie o 12 tysięcznych - wydaje się to być bardzo małą stratą więc będziemy mieć to na uwadze zastanawiając się czy również usunąć tą zmienną z modelu.

Postanawiamy również zobaczyć inną metodą które zmienne byłyby najlepsze w dalszym modelowaniu. Zastosujemy kryterium CP, wybieramy taki model który ma jak najniższą wartość CP przy jednoczesnym braku obciążenia (CP>p uznajemy za obciążony).



(najlepszy model zaznaczony na żółto - zmienne 1,2,4,5,6,7,8,9,10,11,12 - a więc również model bez zmiennej indus). Używając metody adjR2 uzyskaliśmy taki sam wynik.

Zauważamy, że zmienna age jest mało istotna, z uwagi na dużą ilość pozostałych wpływowych zmienny postanawiamy sprawdzić czy wyeliminowanie jej wpłynie negatywnie na model.

Dopasowanie modelu i wstępna diagnostyka.

Zmienna zależną jest MEDV czyli średnia wartość domów. Celem projektu jest zbadanie, które zmienne wpływają na cenę oraz sprawdzenie w jakim stopniu wpływają one na zmienną zależną.

Summary z pełnego modelu po dokonanych transformacjach:

```
Call:
lm(formula = sq_medv ~ ., data = BH2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2789 -0.2183 -0.0402  0.2006  1.7049

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.2453235  0.4223560  14.787  < 2e-16 ***
crim        -0.0165752  0.0026707  -6.206 1.18e-09 ***
zn           0.0029667  0.0011545   2.570   0.0105 *
indus       -0.0008949  0.0050904  -0.176   0.8605
nox         -1.3819112  0.3125659  -4.421 1.22e-05 ***
rm           0.3105408  0.0365888   8.487 2.69e-16 ***
age         -0.0018545  0.0010909  -1.700   0.0898 .
dis         -0.1137609  0.0164571  -6.913 1.53e-11 ***
rad          0.0281162  0.0054096   5.197 3.00e-07 ***
tax         -0.0015260  0.0003043  -5.015 7.50e-07 ***
ptratio     -0.0867531  0.0107525  -8.068 5.83e-15 ***
b            0.0008693  0.0002181   3.986 7.77e-05 ***
lstat       -0.0457500  0.0043556 -10.504  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3851 on 477 degrees of freedom
Multiple R-squared:  0.7956,    Adjusted R-squared:  0.7905
F-statistic: 154.7 on 12 and 477 DF,  p-value: < 2.2e-16
```

po zastosowaniu metody “both” - z pozostawioną zmienną age:

```
Call:
lm(formula = sq_medv ~ . - indus, data = BH2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27867 -0.22016 -0.04006  0.19828  1.70612

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.2493595  0.4213038  14.833  < 2e-16 ***
crim        -0.0165515  0.0026646  -6.212 1.14e-09 ***
zn           0.0029862  0.0011480   2.601  0.00958 **
nox         -1.3974457  0.2995102  -4.666 4.00e-06 ***
rm           0.3111721  0.0363751   8.555  < 2e-16 ***
age         -0.0018508  0.0010896  -1.699  0.09005 .
dis         -0.1131425  0.0160604  -7.045 6.51e-12 ***
rad          0.0283561  0.0052294   5.422 9.35e-08 ***
tax         -0.0015480  0.0002773  -5.582 3.99e-08 ***
ptratio     -0.0870257  0.0106293  -8.187 2.45e-15 ***
b            0.0008706  0.0002178   3.998 7.40e-05 ***
lstat       -0.0458244  0.0043307 -10.581  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3847 on 478 degrees of freedom
Multiple R-squared:  0.7956,    Adjusted R-squared:  0.7909
F-statistic: 169.2 on 11 and 478 DF,  p-value: < 2.2e-16
```

Postanawiamy sprawdzić jak sprawuje się model po odrzuceniu również zmiennej age (jak widzimy nie wydaje się ona istotna):

```
Call:
lm(formula = sq_medv ~ . - indus - age, data = BH2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.28079 -0.22312 -0.03717  0.19785  1.69803

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.3200026  0.4200702  15.045  < 2e-16 ***
crim        -0.0165657  0.0026698  -6.205 1.19e-09 ***
zn           0.0031943  0.0011437   2.793 0.005431 **
nox         -1.5431757  0.2875235  -5.367 1.25e-07 ***
rm           0.2979874  0.0356072   8.369 6.44e-16 ***
dis         -0.1053280  0.0154177  -6.832 2.56e-11 ***
rad          0.0290163  0.0052252   5.553 4.66e-08 ***
tax         -0.0015512  0.0002778  -5.583 3.97e-08 ***
ptratio     -0.0885296  0.0106132  -8.341 7.88e-16 ***
b            0.0008486  0.0002178   3.896 0.000112 ***
lstat       -0.0485646  0.0040269 -12.060  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3855 on 479 degrees of freedom
Multiple R-squared:  0.7944,    Adjusted R-squared:  0.7901
F-statistic: 185.1 on 10 and 479 DF,  p-value: < 2.2e-16
```

R kwadrat spada jedynie o 12 tysiączne, co jest bardzo mało różnicą, gdy rozważamy usunięcie zmiennej niezależnej. Postanawiamy więc ją odrzucić.

Kontynuujemy pracę z tym modelem. Pamiętając wygląd macierzy korelacji (dość silną korelację pomiędzy niektórymi zmiennymi niezależnymi) chcemy sprawdzić współliniowość zmiennych niezależnych, badamy to przez współczynnik wzrostu wariancji (VARIANCES VARIATION FACTOR). Przyjmujemy, że nie mamy silnej zależności, gdy VIF równa się około 1. Nie istnieje ścisłe ograniczenie, lecz przyjmujemy, że występuje silna zależność gdy VIF>10. Nasz wynik jest następujący:

```
> vif(reg_less)
    crim       zn      nox       rm      dis      rad      tax  ptratio
1.784764 2.246692 3.704817 1.780189 3.482967 6.788062 7.176430 1.651292
       b    lstat
1.339766 2.677692
```

Jak widzimy zmienne rad i tax mają dość wysoki współczynnik (podobne wnioski mogliśmy wyciągnąć z macierzy korelacji). Tak więc spróbujemy usunąć jedną z tych zmiennych i sprawdzić wpływ tej operacji na model.

Z naszego modelu usuwamy zmienną tax.

```
Call:
lm(formula = sq_medv ~ . - indus - age - tax, data = BH2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.28631 -0.23194 -0.03606  0.19926  1.73717

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.1127601  0.4313763  14.170  < 2e-16 ***
crim        -0.0160906  0.0027510  -5.849 9.16e-09 ***
zn           0.0018246  0.0011516   1.584   0.1138
nox         -1.9075424  0.2886845  -6.608 1.04e-10 ***
rm           0.3221503  0.0364370   8.841  < 2e-16 ***
dis         -0.0971149  0.0158223  -6.138 1.75e-09 ***
rad          0.0065954  0.0034464   1.914   0.0562 .
ptratio     -0.0985393  0.0107844  -9.137  < 2e-16 ***
b            0.0009107  0.0002242   4.061 5.71e-05 ***
lstat       -0.0498001  0.0041452 -12.014  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3974 on 480 degrees of freedom
Multiple R-squared:  0.781,     Adjusted R-squared:  0.7769
F-statistic: 190.2 on 9 and 480 DF,  p-value: < 2.2e-16
```

Wyniki są bardzo interesujące. Okazuje się, że po usunięciu zmiennej niezależnej tax - pełnowartościowa stawka podatku od nieruchomości za 10 tys USD - inne zmienne przestają być istotne w modelu. Może to potwierdzać silną korelację tax z innymi zmiennymi, co prowadziło do zaburzenia analizy. Postanawiamy więc prowadzić dalsze badania bez tej zmiennej. Badamy jeszcze raz współczynnik wzrostu wariancji i jak widzimy przyjmuje on teraz dość niskie wartości dla każdej zmiennej:

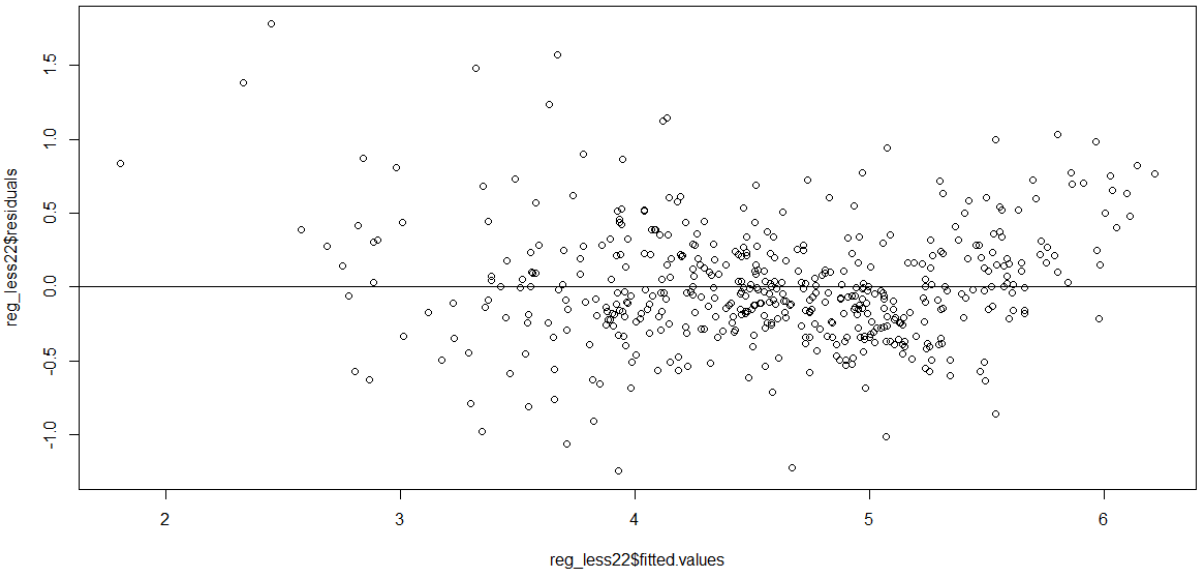
```
> vif(reg_less2)
    crim       zn      nox       rm      dis      rad  ptratio        b
1.782951 2.143308 3.513935 1.753889 3.451257 2.778385 1.604168 1.336274
   lstat
2.669605
```

Postanawiamy usunąć z modelu zmienne zn i rad jako, że wydają się one nieistotne (p-value w modelu powyżej >0.05). Wynik jeszcze bardziej pomniejszonego modelu jest jak poniżej:

```
Call:
lm(formula = sq_medv ~ . - indus - age - tax - zn - rad, data = BH2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2463 -0.2387 -0.0413  0.2129  1.7802

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.8414932  0.4082735  14.308  < 2e-16 ***
crim        -0.0131990  0.0024815  -5.319 1.60e-07 ***
nox         -1.6959983  0.2676278  -6.337 5.38e-10 ***
rm           0.3424947  0.0357837   9.571  < 2e-16 ***
dis         -0.0820660  0.0136993  -5.991 4.10e-09 ***
ptratio     -0.0951962  0.0094397 -10.085  < 2e-16 ***
b            0.0008295  0.0002221   3.734 0.000211 ***
lstat       -0.0491267  0.0041564 -11.820  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3994 on 482 degrees of freedom
Multiple R-squared:  0.7778,    Adjusted R-squared:  0.7746
F-statistic: 241.1 on 7 and 482 DF,  p-value: < 2.2e-16
```

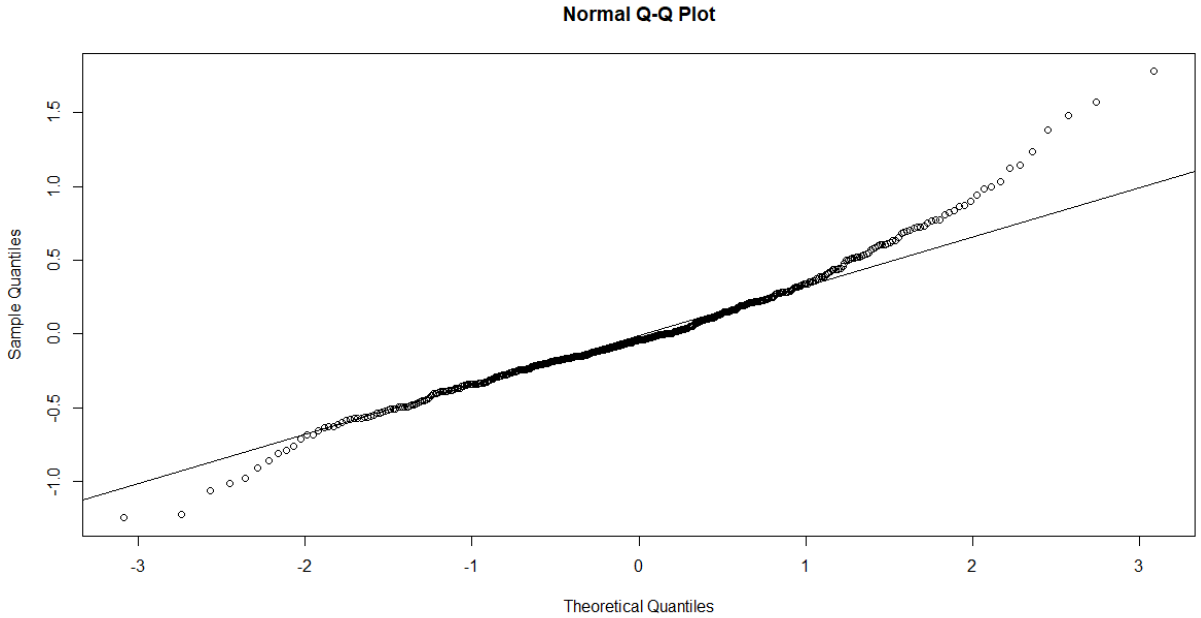
Jak widzimy R kwadrat spadł ledwo o 30 tysięcznych, co utwierdza nas w przekonaniu, że usunięcie kolejnych dwóch zmiennych było dobrą decyzją. Uzyskaliśmy dość dobry rezultat 7 zmiennych niezależnych, które są istotne. Musimy pamiętać, że zaczynaliśmy z liczbą 13, więc dość zgrabnie pozbyliśmy się nadmiarowych informacji, które mogłyby utrudniać analizę.

Na tym pomniejszonym modelu ponownie tworzymy kolejne wykresy:
-wykres residuów od estymowanego Y:

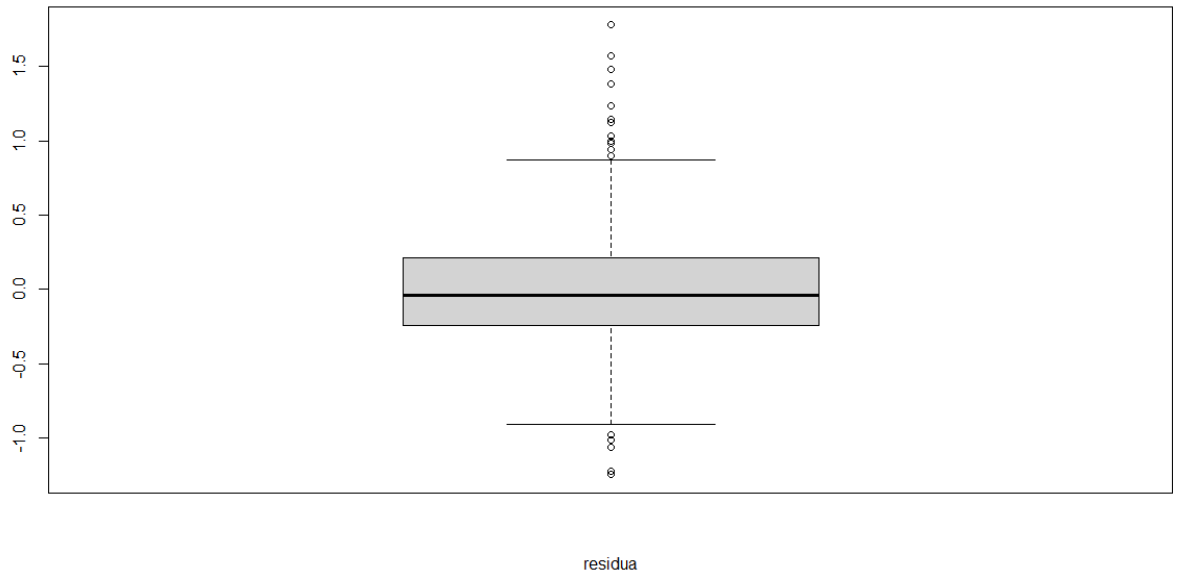


nie zauważamy żadnej istotnej zależności.

-wykres qqnorm residuów:



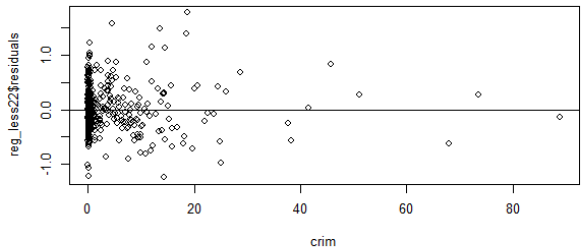
Jak widzimy ogony nadal są odstające, lecz ułożone dość symetrycznie. Potwierdzamy to boxplotem reszt:

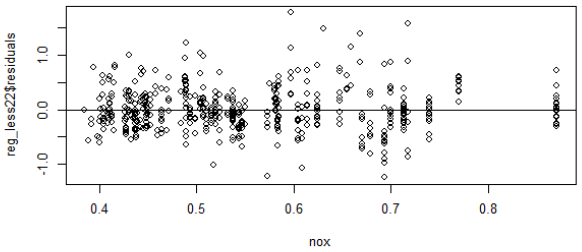


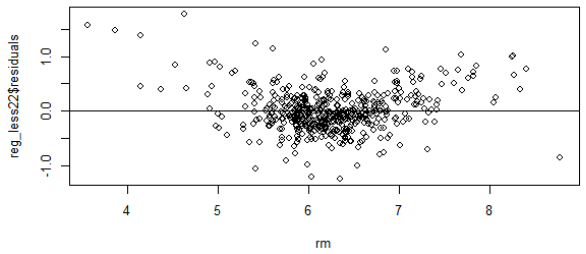
Nie jest to rozkład normalny (potwierdziliśmy to również testem normalności Shapiro-Wilka) lecz średnia jest bardzo bliska zeru, oraz nie obserwujemy skośności rozkładu.

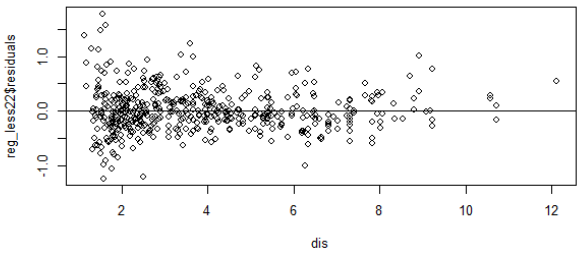
```
        Shapiro-Wilk normality test

data:  reg_less22$residuals
W = 0.96717, p-value = 5.175e-09
```

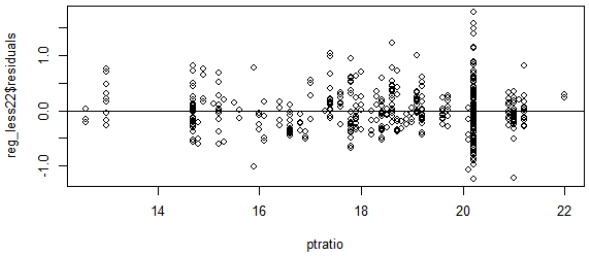
- Kolejne wykresy reszt od zmiennych niezależnych pozostawionych w modelu:

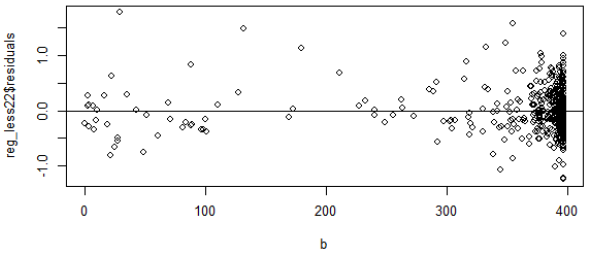


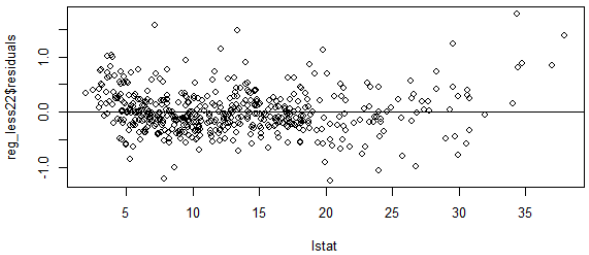






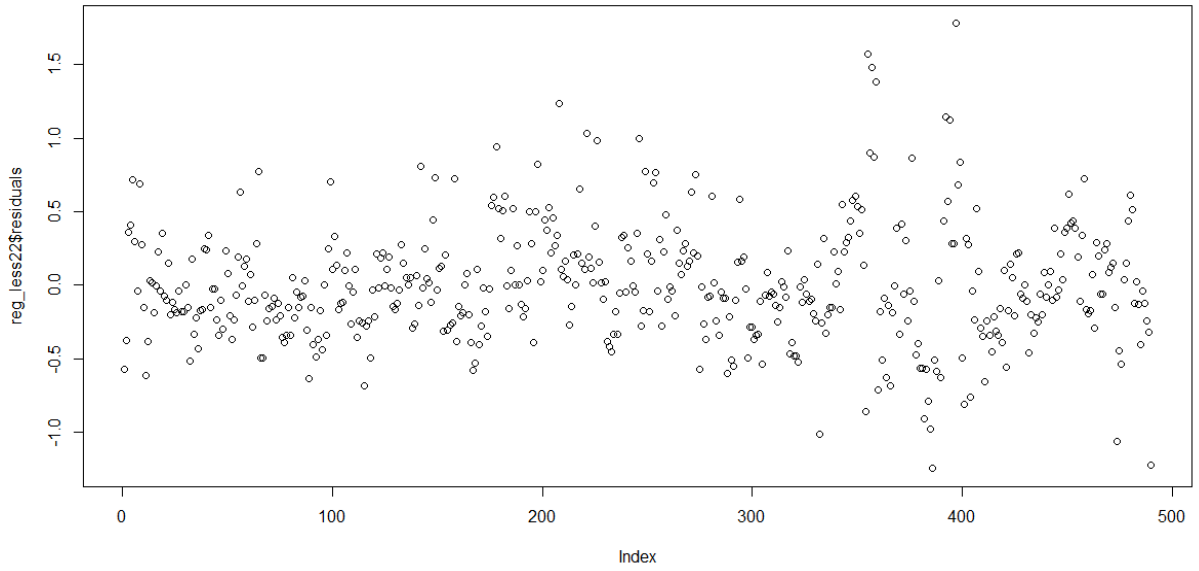






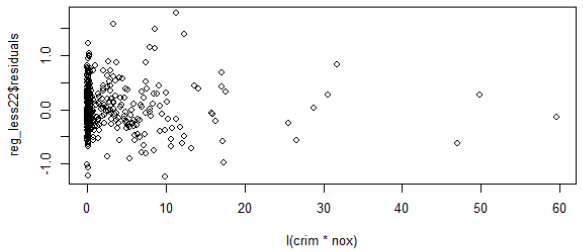
Nie zauważamy żadnej niepokojącej zależności pomiędzy resztami a zmiennymi niezależnymi.

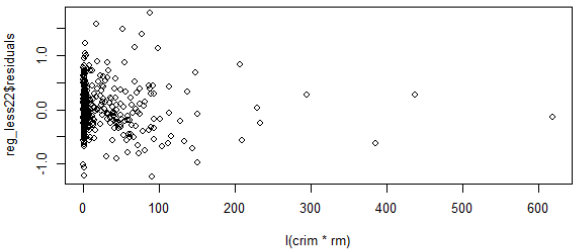
- Wykres reszt

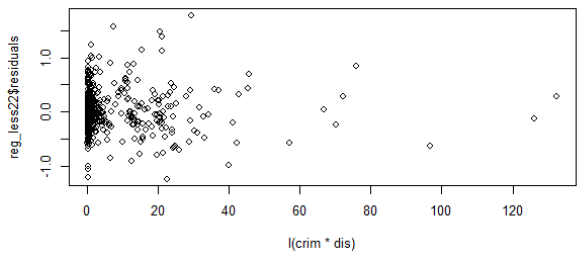


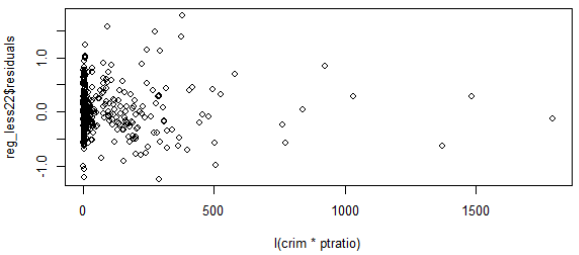
Postanawiamy sprawdzić również interakcje, interpretujemy to jako wspólne działanie jednej i drugiej zmiennej:

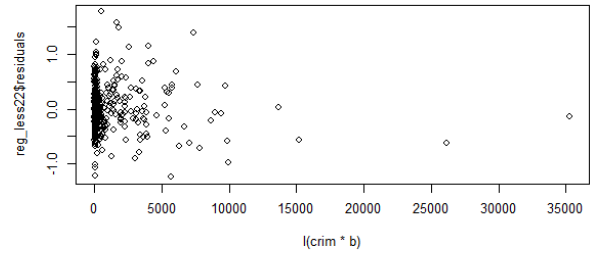
- interakcja zmiennej crim i kolejnych zmiennych niezależnych:

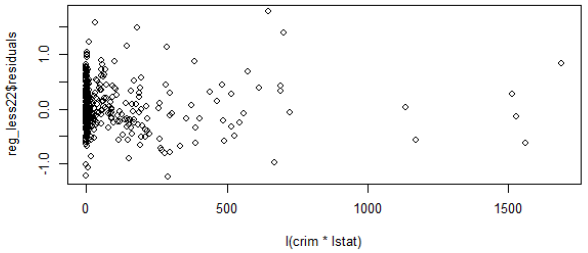




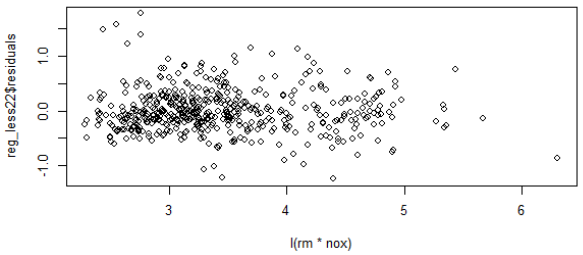


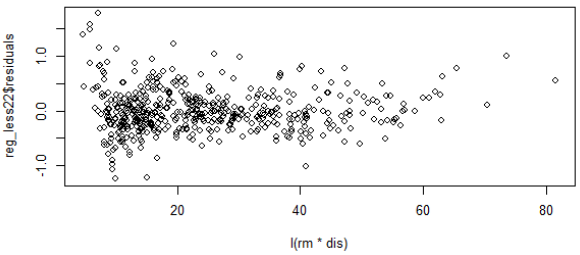


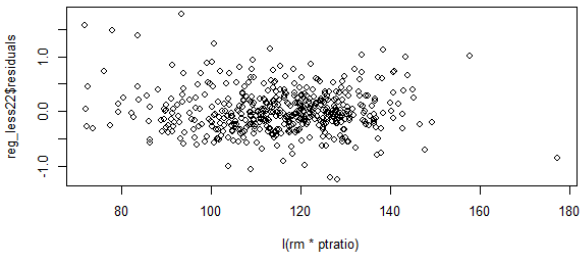


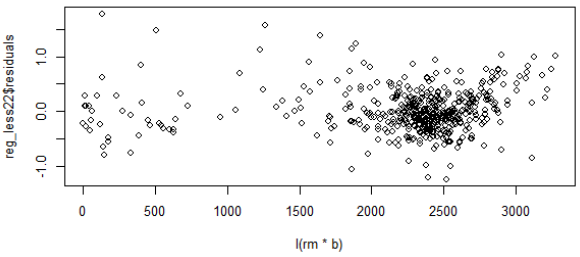


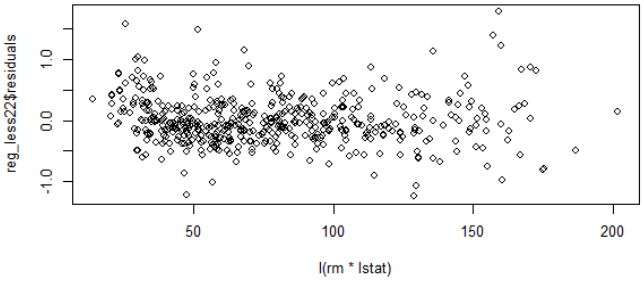
- zmiennej rm i kolejnych zmiennych niezależnych:



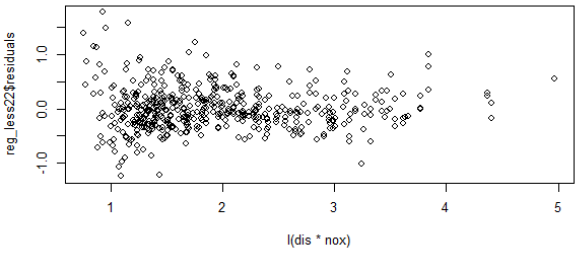


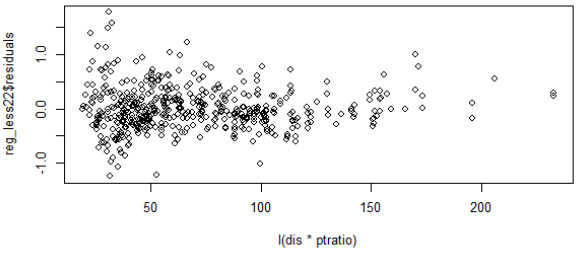


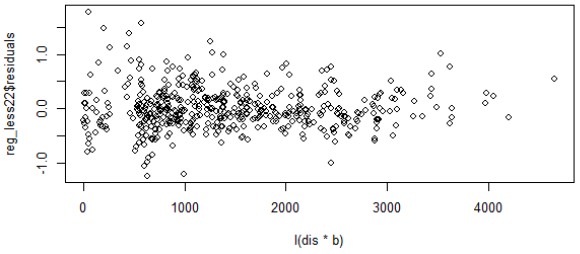


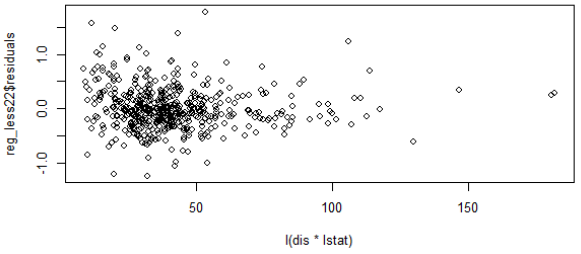


- zmiennej dis z pozostałymi zmiennymi:









Zauważamy możliwą lekką liniową zależność w wykresie I(rm*lstat). Spróbujemy sprawdzić jak dodanie tej interakcji oraz na przykład I(dis*nox) wpłynie na wyniki modelu:

- model z dodatkową interakcją rm*lstat plus dis*nox:

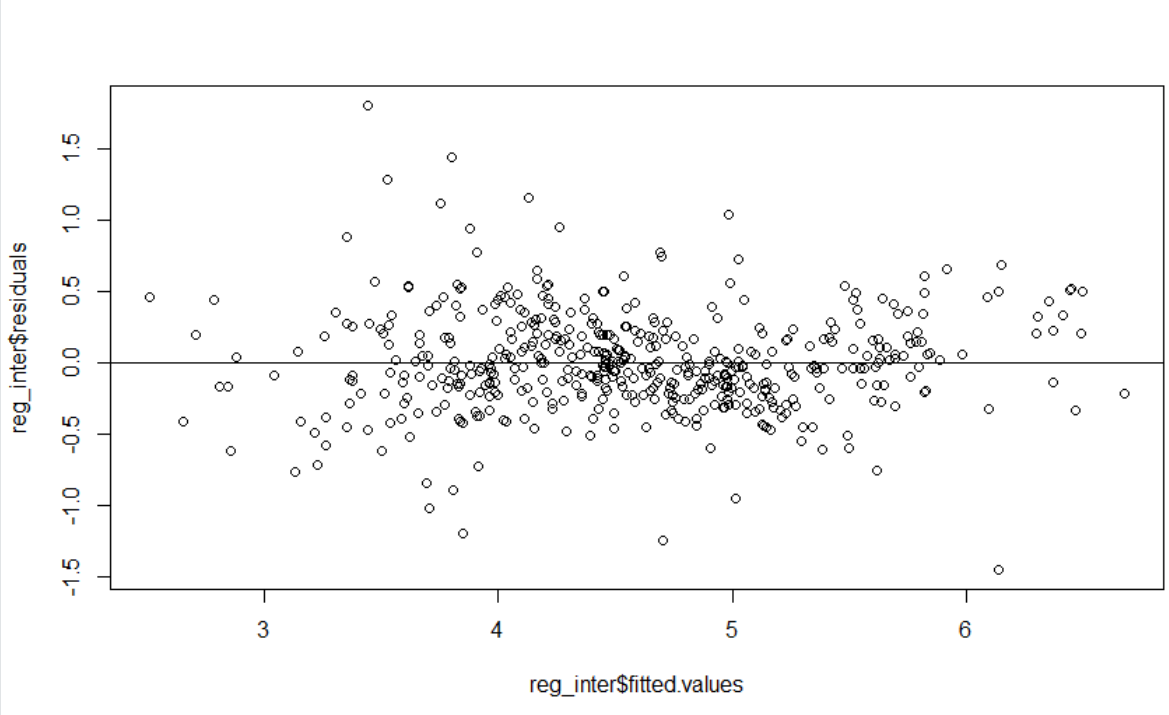
```
Call:
lm(formula = sq_medv ~ . - indus - age - tax - zn - rad + I(rm *
    lstat) + I(dis * nox), data = BH2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43731 -0.20758 -0.03332  0.19250  1.85669

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.4232649  0.4454663   7.685 8.72e-14 ***
crim          -0.0159138  0.0022995  -6.921 1.45e-11 ***
nox           -1.7597834  0.3557916  -4.946 1.05e-06 ***
rm             0.7062217  0.0451999  15.624  < 2e-16 ***
dis           -0.1801063  0.0606988  -2.967  0.00316 **
ptratio       -0.0758741  0.0089667  -8.462 3.21e-16 ***
b              0.0004210  0.0002009   2.095  0.03668 *
lstat          0.1438952  0.0174274   8.257 1.46e-15 ***
I(rm * lstat) -0.0336288  0.0029655 -11.340  < 2e-16 ***
I(dis * nox)   0.2479906  0.1412366   1.756  0.07975 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3554 on 480 degrees of freedom
Multiple R-squared:  0.8248,    Adjusted R-squared:  0.8215
F-statistic: 251.1 on 9 and 480 DF,  p-value: < 2.2e-16
```

Co nas pozytywnie zaskakuje interakcja rm*lstat wydaje się istotna oraz R kwadrat dość istotnie wzrasta, to bardzo dobry rezultat. I(dis*nox) jest już mniej istotna, więc sprawdzamy model bez niej:

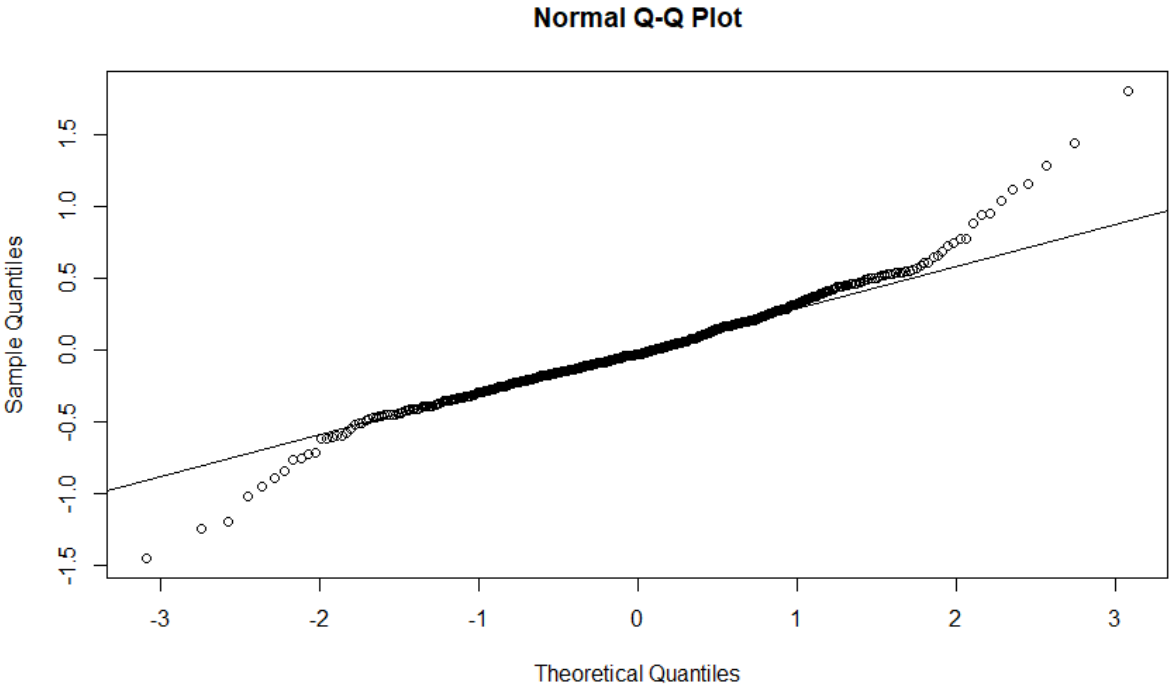
```
Call:
lm(formula = sq_medv ~ . - indus - age - tax - zn - rad + I(rm *
    lstat), data = BH2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4552 -0.2043 -0.0316  0.1919  1.8047

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.2301344  0.4326070   7.467 3.89e-13 ***
crim          -0.0168799  0.0022375  -7.544 2.30e-13 ***
nox           -1.2998569  0.2413020  -5.387 1.12e-07 ***
rm             0.6982087  0.0450661  15.493  < 2e-16 ***
dis           -0.0757048  0.0122311  -6.190 1.29e-09 ***
ptratio       -0.0717719  0.0086757  -8.273 1.29e-15 ***
b              0.0004273  0.0002013   2.122   0.0343 *
lstat          0.1382893  0.0171695   8.054 6.34e-15 ***
I(rm * lstat) -0.0327821  0.0029324 -11.179  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3562 on 481 degrees of freedom
Multiple R-squared:  0.8237,    Adjusted R-squared:  0.8207
F-statistic: 280.8 on 8 and 481 DF,  p-value: < 2.2e-16
```

Tak jak widzimy, R2 spada jedynie 15 tysięcznych, więc możemy nie brać pod uwagę I(dis*nox). Interakcje rm*lstat pozostawiamy w modelu. A więc jeszcze raz wykonujemy kolejne wykresy reszt, by upewnić się, że założenia modelu regresji liniowej nadal są spełnione w naszym zmodyfikowanym modelu:

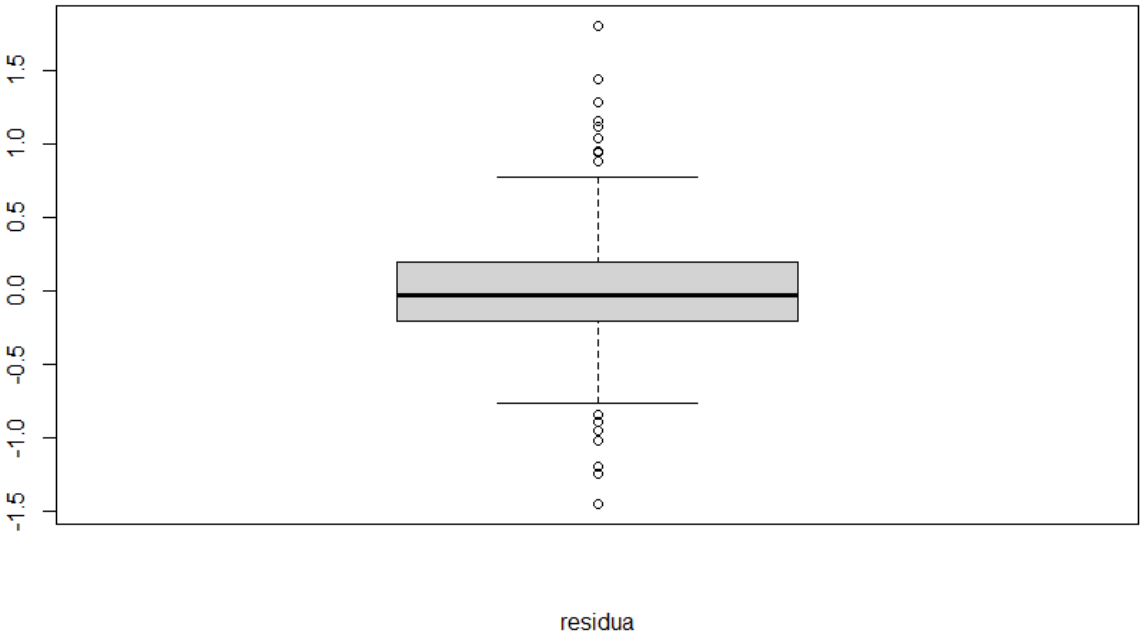
- reszty od estymowanego Y:



- qqnorm:



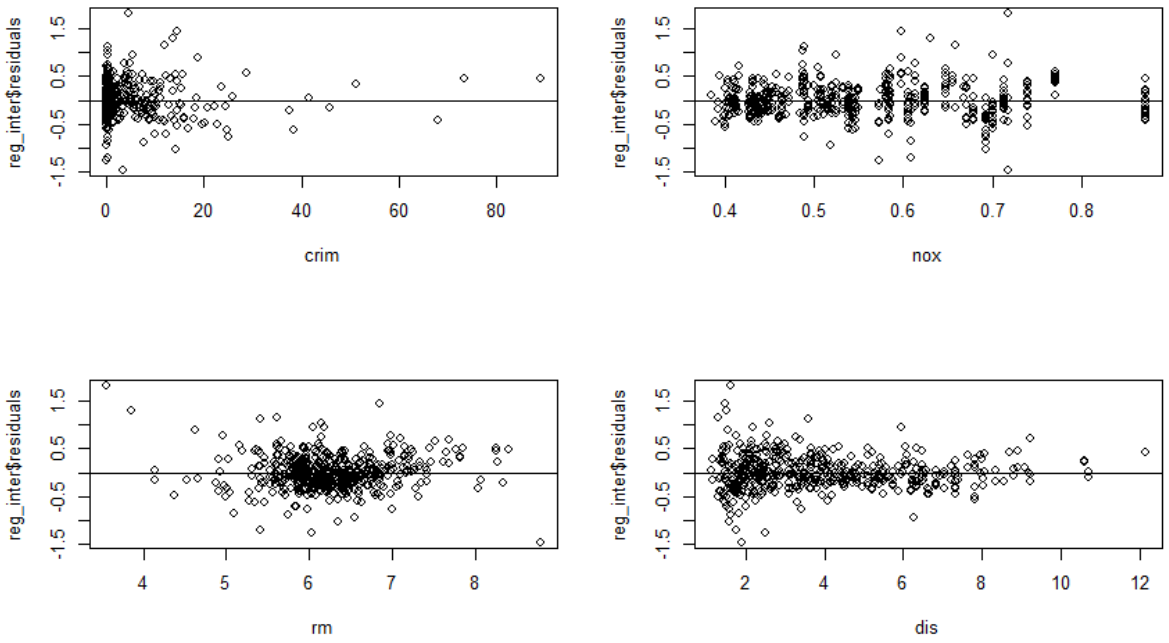
- boxplot reszt:

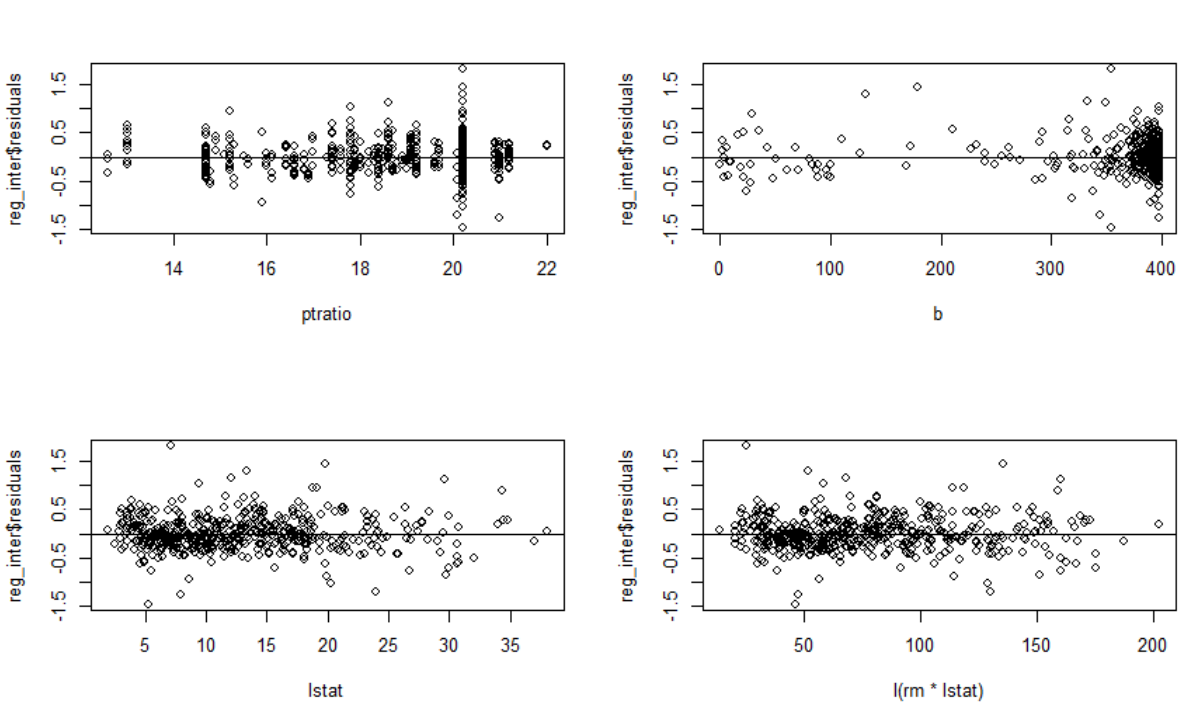


- test normalności reszt (nie mają rozkładu normalnego, ale nadal średnia około 0 i symetrycznie rozłożone)

```
        Shapiro-Wilk normality test

data:  reg_inter$residuals
W = 0.96309, p-value = 9.508e-10
```

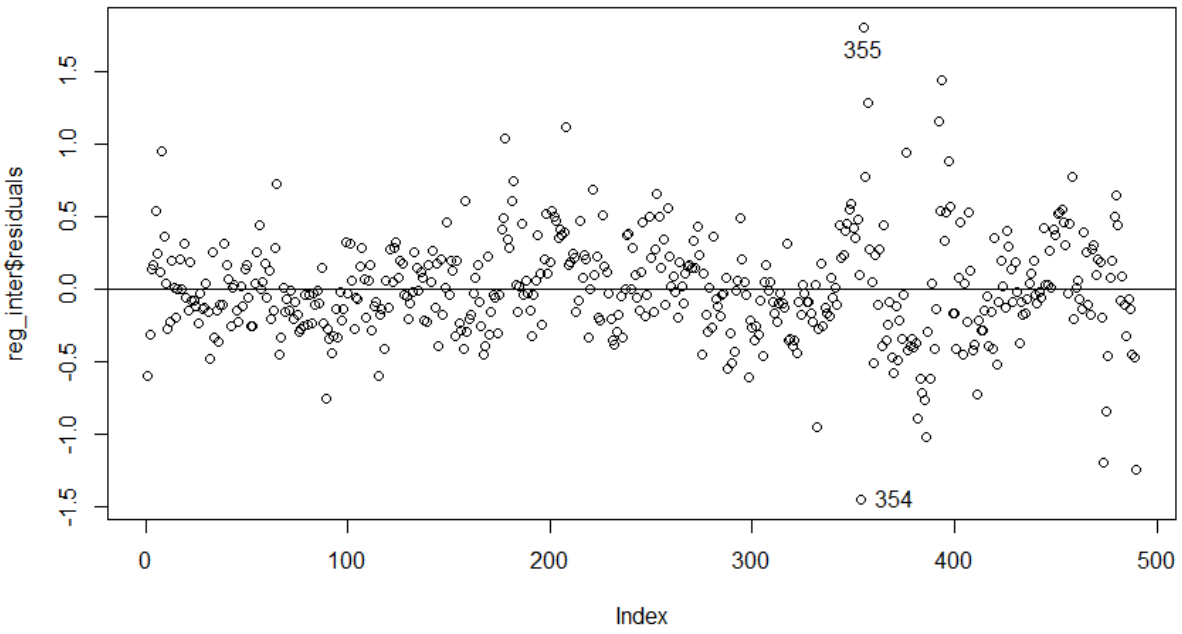
- reszty od zmiennych niezależnych:



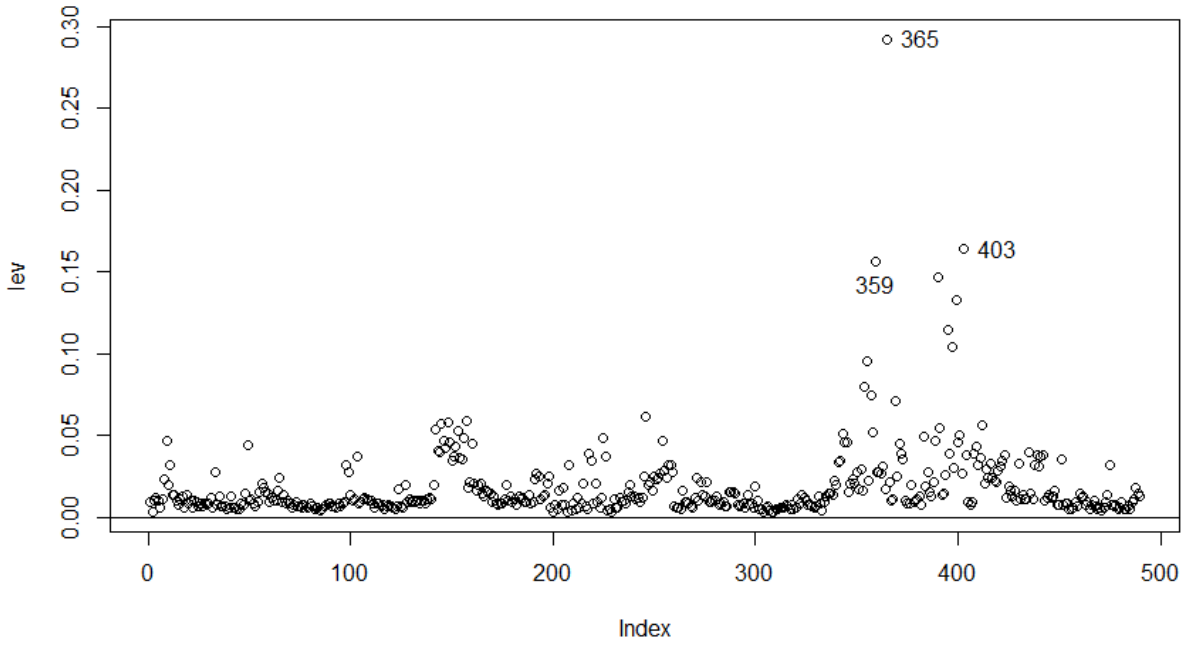


Wszystkie wykresy wydają się w porządku, nie zauważamy żadnej podejrzanej zależności.

- Wykres residuów z zaznaczonymi skrajnymi wartościami reszt (odpowiednio -1.455182 1.804688):



- Wykres hatvalues (wskazuje potencjał dźwigni, czyli obserwacji odstających ze względu na X, skrajne wartości odpowiednio: 0.1561639; 0.2923291; 0.1638508):



Ostatecznie sprawdzamy jeszcze które obserwacje są obserwacjami wpływowymi. Robimy to za pomocą funkcji DFFITS, nasza próba liczy 490 obserwacji i 8 parametrów, a więc przyjmujemy obserwację za wpływową gdy wartość bezwzględna z DFFITS jest większa niż 2*sqrt(P/n), gdzie P-liczba parametrów, n-liczba obserwacji. U nas dostajemy 44 obserwacje wpływowe na 490 ogółem. Są to kolejno obserwacje wypisane poniżej jako TRUE:

```
> wplywowe<-abs(dffits(reg_inter))>2*sqrt(P/n)
> wplywowe[wplywowe==T]
   8   65  145  149  157  182  215  229  234  254  262  263  267  354  365  366  367  368
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 381  385  386  388  399  400  401  402  404  406  408  410  411  412  413  414  417  419
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
 420  425  427  437  467  490  491  506
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> table(wplywowe)
wplywowe
FALSE  TRUE
  446    44
```

Jako, że na początku przy przeglądzie danych sprawdziliśmy, że nie mamy żadnych obserwacji niemożliwych, nie zamierzamy usuwać żadnej z nich. Nie są one w żaden sposób według nas odstające niepoprawnie- tak, że mogłyby zaburzyć analizę. Pozostawiamy nasz model w ostatecznej formie:

```
Call:
lm(formula = sq_medv ~ . - indus - age - tax - zn - rad + I(rm *
    lstat), data = BH2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4552 -0.2043 -0.0316  0.1919  1.8047

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.2301344  0.4326070   7.467 3.89e-13 ***
crim          -0.0168799  0.0022375  -7.544 2.30e-13 ***
nox           -1.2998569  0.2413020  -5.387 1.12e-07 ***
rm             0.6982087  0.0450661  15.493  < 2e-16 ***
dis           -0.0757048  0.0122311  -6.190 1.29e-09 ***
ptratio       -0.0717719  0.0086757  -8.273 1.29e-15 ***
b              0.0004273  0.0002013   2.122   0.0343 *
lstat          0.1382893  0.0171695   8.054 6.34e-15 ***
I(rm * lstat) -0.0327821  0.0029324 -11.179  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3562 on 481 degrees of freedom
Multiple R-squared:  0.8237,    Adjusted R-squared:  0.8207
F-statistic: 280.8 on 8 and 481 DF,  p-value: < 2.2e-16
```

Wnioski

Analiza wskazuje, że ceny mieszkań w Bostonie mocno zależą od następujących zmiennych:
CRIM - wskaźnik przestępczości na mieszkańca według miast.
NOX - stężenie tlenków azotu (ilość na 10 mln).
RM - średnia liczba pokoi na mieszkanie.
DIS - odległości do pięciu bostońskich centrów zatrudnienia.
PTRATIO - stosunek uczniów do nauczycieli.
LSTAT - % niższego statusu ludności
RM * LSTAT
oraz średnio od
B - 1000(Bk - 0.63)^2 gdzie Bk to odsetek czarnoskórych według miasta.

Uzyskaliśmy dość dobry rezultat 7 zmiennych niezależnych, które są istotne redukując początkowe 13 zmiennych o połowę.