Sprawozdanie z projektu indywidualnego Data Mining

Spis treści

Zadanie 1 - Analiza zbioru `Cars 93`, danych różnych modeli samochodów

a) Pierwsze operacje

Wczytano zbiór. Z wczytanego zbioru utworzono podzbiór, w którym zawarte są wskazane w poleceniu kolumny: *Min.Price, MPG.city, MPG.highway, Weight, Origin, Type*.

Następnie utworzono nowe zmienne w zbiorze, zgodnie z formułami zawartymi w poleceniu:

```
cars$Lp100km.city <- ((100*3.8)/(1.6*cars$MPG.city))
cars$Lp100km.highway <- ((100*3.8)/(1.6*cars$MPG.highway))
cars$Weight.kg <- cars$Weight * 0.4536
cars$Min.Price.PLN <- cars$Min.Price * 3.35
```

b) Statystyki próbkowe

Dla ceny wersji podstawowej w złotówkach obliczono kwantyl rzędu 0.95:

```
95% :
115.508
```

Dla wyznaczonego kwantyla wypisano ceny i modele samochodów droższe niż wyznaczona cena:

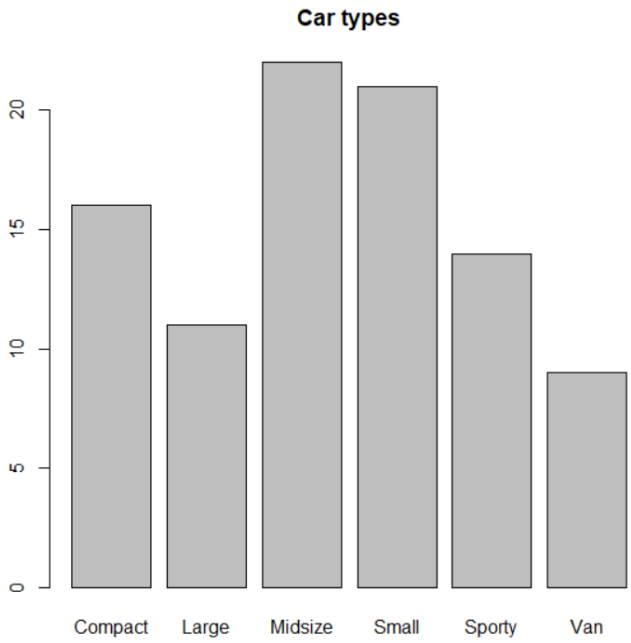
```
                  Make Min.Price.PLN
11   Cadillac Seville       125.625
19 Chevrolet Corvette       115.910
48       Infiniti Q45       152.090
50        Lexus SC300       116.245
59 Mercedes-Benz 300E       146.730
```

e) Wykres słupkowy i kołowy zmiennej *Type*

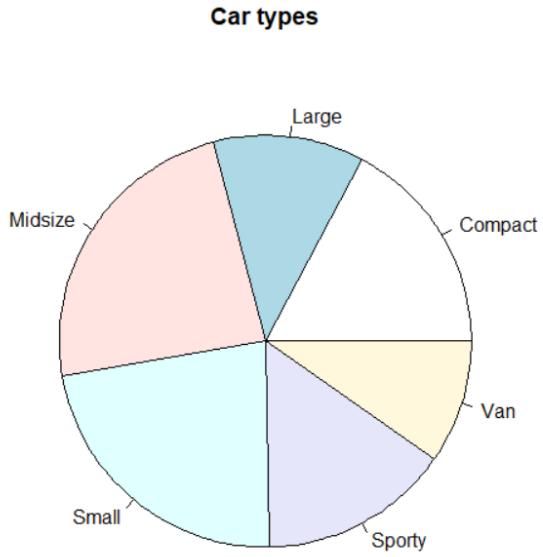
Sporządzono zestawienie ilości każdego typu samochodu:

```
     Type freq
1 Compact   16
2   Large   11
3 Midsize   22
4   Small   21
5  Sporty   14
6     Van    9
```

Dla zmiennej *Type* narysowano wykres słupkowy:



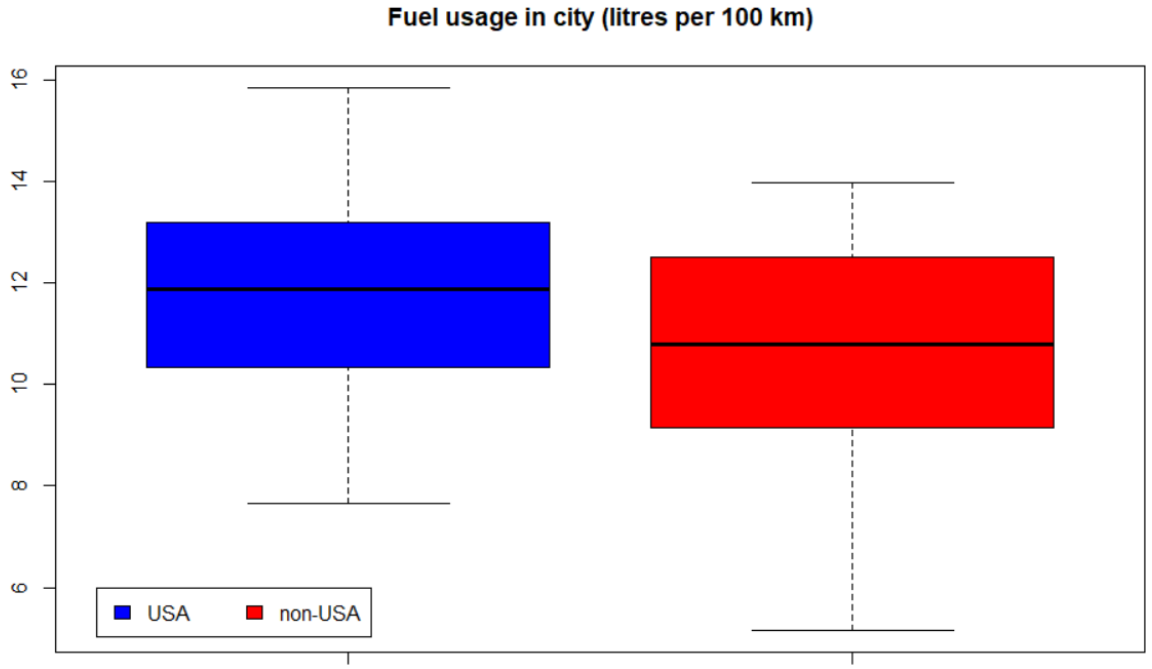
Oraz kołowy:



Wśród badanych samochodów 14 z nich zaliczono do kategorii sportowej.

f) Wykresy skrzynkowe zużycia benzyny w mieście

Sporządzono wykresy skrzynkowe dla zużycia benzyny podczas jazdy w mieście z podziałem na samochody amerykańskie i pozostałe.

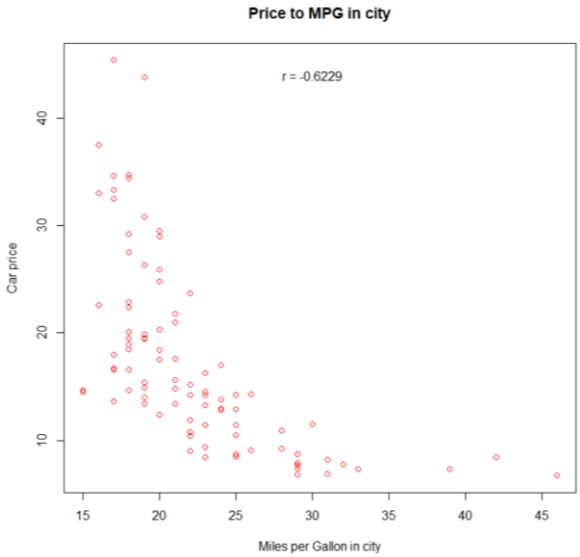


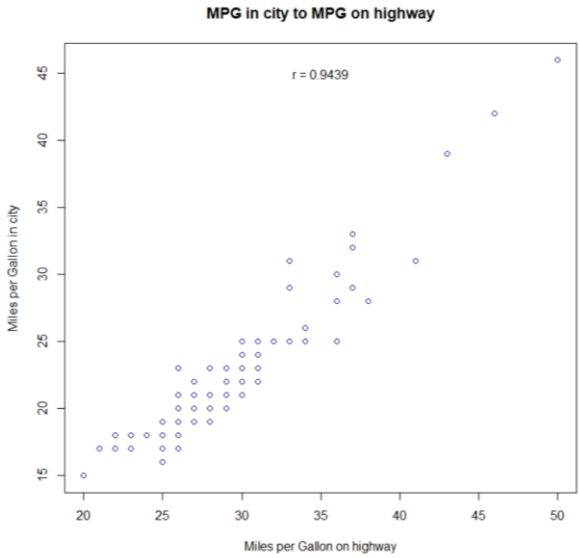
Ze sporządzonego wykresu można wyciągnąć wnioski, że samochody spoza USA są bardziej ekonomiczne w mieście – mają niższe spalanie, co widać po medianie oraz wartościach pierwszego i trzeciego kwartyla. Samochody amerykańskie mają mniejsze „wąsy", czyli wartości obserwowane nie odstające od pozostałych. Dodatkowo, mają mniejszy rozstęp międzykwartylowy, co oznacza, że wartość pierwszego i trzeciego kwartyla mają niższą różnicę niż dla samochodów spoza USA. Sugeruje to, że samochody nie pochodzące z Ameryki maja bardziej zróżnicowane spalanie w mieście.

g) Wykresy rozrzutu

Sporządzono wykresy rozrzutu:

- Ceny podstawowej wersji samochodu w funkcji jego zużycia benzyny w mieście
- Zużycia benzyny w mieście w funkcji zużycia benzyny na autostradzie





Analizując rozrzut przedstawionych obserwacji, można wywnioskować, że zarówno w pierwszym jak i w drugim przypadku mamy do czynienia z silnie skorelowanymi zmiennymi.

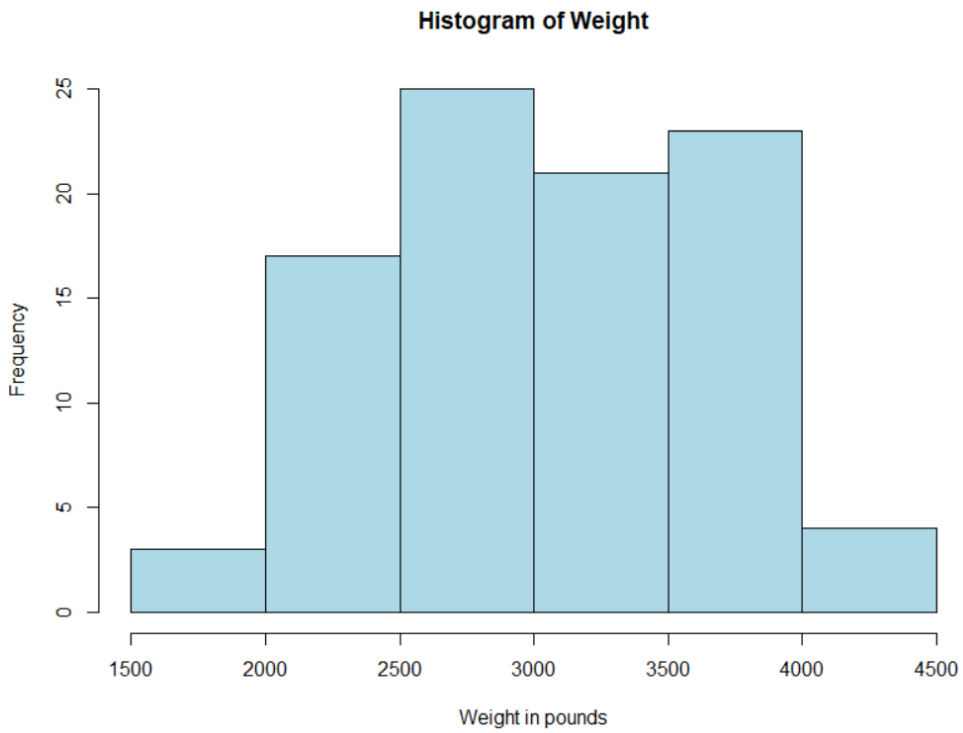
W pierwszym przypadku współczynnik korelacji wynosi *r* = - 0,6229. Jest on ujemny, co może świadczyć o odwrotnie proporcjonalnej zależności – im wyższa cena tym mniej mil na jeden galon paliwa. Wskazuje to na wniosek, że drogie samochody to również takie, które mają wysokie spalanie. Z kolei te tańsze pozwalają na przejechanie większej odległości na jednym galonie paliwa.

Na drugim wykresie mamy do czynienia ze współczynnikiem korelacji *r* bardzo bliskim wartości 1, bo wynosi 0,9439. Widać to również po punktach obserwacji naniesionych na wykresy, których ułożenie przypomina funkcję liniową. Zatem liczba mil na galon paliwa w mieście jest wprost proporcjonalna do wyników spalania na autostradzie. Prowadzi to do stwierdzenia, że jeśli samochód spala względnie dużo paliwa w mieście, to również na autostradzie będzie potrzebował więcej paliwa niż inne samochody. Wskazuje to na obserwację, że dla większości samochodów ze zbioru danych, iloraz spalania na autostradzie przez spalanie w mieście jest podobny i średnio wynosi 1.32.

```
> summary(cars$MPG.highway/cars$MPG.city)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.065   1.250   1.318   1.320   1.400   1.562 
```

h) Histogram częstości wagi samochodu

Sporządzono histogram częstości wagi samochodu w funtach:



Jak widać na histogramie, znaczna większość samochodów ma masę w przedziale od 2000 do 4000 funtów. Z kolei samochody spoza tego przedziału można uznać za będące wyjątkami w zbiorze, jest ich zaledwie 8,6 % wśród 93 samochodów, czyli jedynie 8. Z przedstawionych na wykresie przedziałów 500 funtów, najwięcej samochodów było w przedziale 2500 – 3000 funtów.

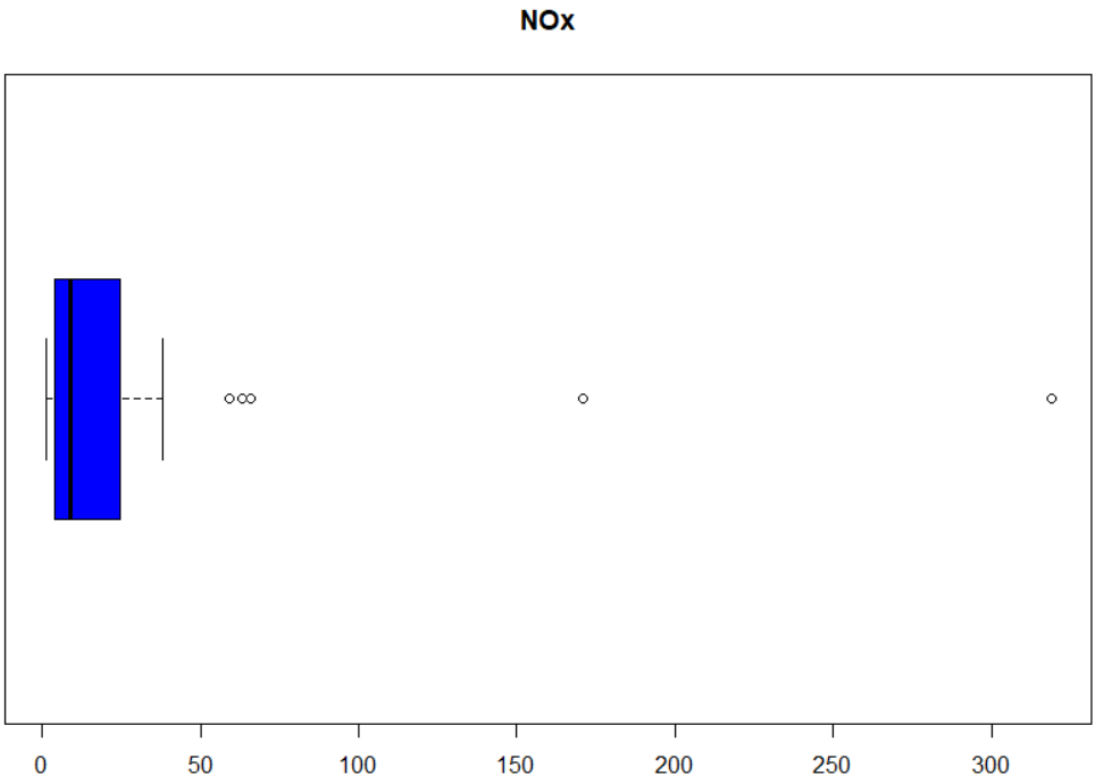
Zadanie 2 – Analiza zbioru `airpollution`, związku zanieczyszczeń powietrza ze śmiertelnością

a) Wczytanie danych, wstępna analiza, model *Mortality(NOx)*

Wyodrębniono wskazane w poleceniu zmienne i utworzono z nich podzbiór. Dokonano statystycznej analizy zmiennych:

```
   Mortality        Education       X.NonWhite        income          JanTemp
 Min.   : 790.7   Min.   : 9.00   Min.   : 0.80   Min.   :   40   Min.   :12.00
 1st Qu.: 898.4   1st Qu.:10.40   1st Qu.: 4.95   1st Qu.:29877   1st Qu.:27.00
 Median : 943.7   Median :11.05   Median :10.40   Median :32451   Median :31.50
 Mean   : 940.3   Mean   :10.97   Mean   :11.87   Mean   :32693   Mean   :33.98
 3rd Qu.: 983.2   3rd Qu.:11.50   3rd Qu.:15.65   3rd Qu.:35384   3rd Qu.:40.00
 Max.   :1113.2   Max.   :12.30   Max.   :38.50   Max.   :47966   Max.   :67.00
    JulyTemp          NOx
 Min.   :63.00   Min.   :  1.00
 1st Qu.:72.00   1st Qu.:  4.00
 Median :74.00   Median :  9.00
 Mean   :74.58   Mean   : 22.60
 3rd Qu.:77.25   3rd Qu.: 23.75
 Max.   :85.00   Max.   :319.00
```

Zmienna *NOx* opisująca stężenie tlenku azotanu ma rozstęp międzykwartylowy 4 – 23.75. Co ciekawe, maksymalna wartość wynosi aż 319. Jak widać na poniższym wykresie skrzynkowym, jest to wartość nietypowa, zaś typowe wartości mieszczą się w przedziale 1 – 45.

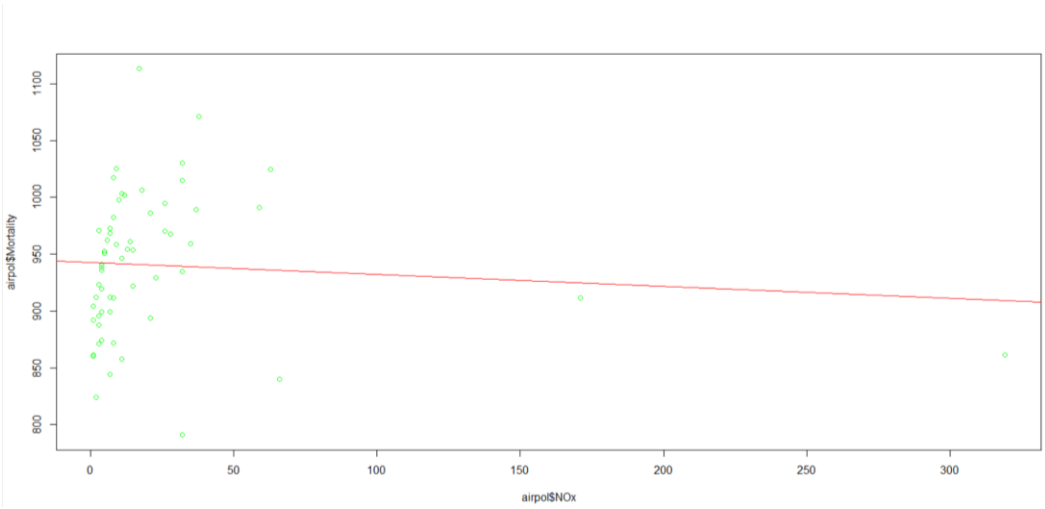


Dopasowano model liniowy objaśniający śmiertelność *Mortality* z pomocą zmiennej stężenia tlenku azotanu:

```
lmMortality <- lm(airpol$Mortality ~ airpol$NOx, data=airpol)
```

b) Weryfikacja modelu *Mortality(NOx)*

Współczynnik nachylenia prostej wyniósł *a* = -0.1043, zaś błąd standardowy *s* = 62.56. W celu zobrazowania tego jak dobrze model dopasował się do danych, naniesiono prostą regresji na wykres rozrzutu *Mortality* od *NOx*.



Jak widać na powyższym wykresie, model źle opisuje *Mortality* od *NOx*. Tak jak było wcześniej widać na wykresie skrzynkowym, wartości stężenia tlenku azotanu skupiają się w zakresie 1-45, z

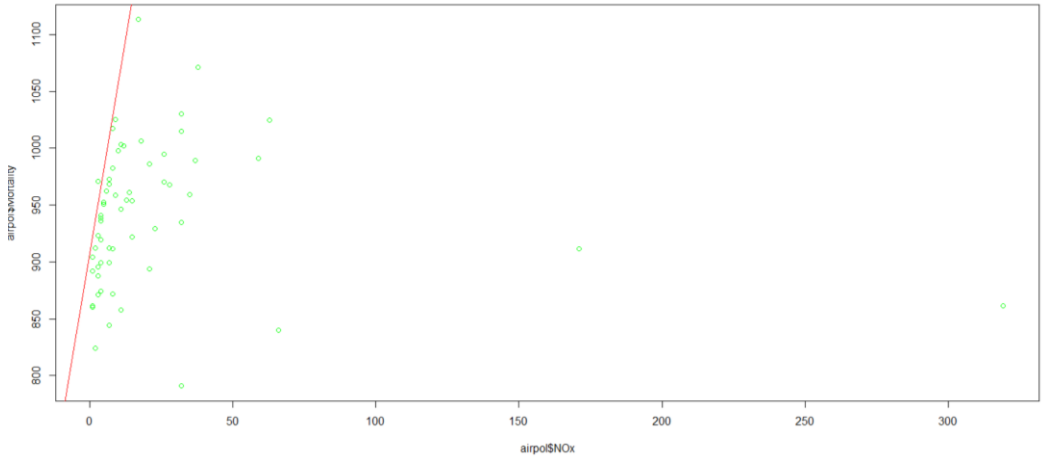
kolei wartości spoza są nietypowe. Model próbuje jednak dopasować się również do tych wartości, co może skutkować złym dopasowaniem.

c) Model *Mortality(log(NOx))*

Zmieniono skalę zmiennej objaśniającej *NOx* na logarytmiczną. Dopasowano model liniowy do tego atrybutu.

Współczynnik nachylenia prostej wyniósł a = 15.1, z kolei błąd standardowy s = 59.96. Jak widać są to zupełnie inne wartości niż w poprzednim przypadku. Prosta wyznaczona przy skali logarytmicznej ma wyższy współczynnik, co oznacza, jest bardziej stroma. Ponadto a jest dodatni, czyli wyznaczony model ma rosnące wartości, a nie malejące. Błąd standardowy dalej jest wysoki, natomiast zmalał względem poprzedniego modelu, a zatem zwiększyliśmy dokładność dopasowania.

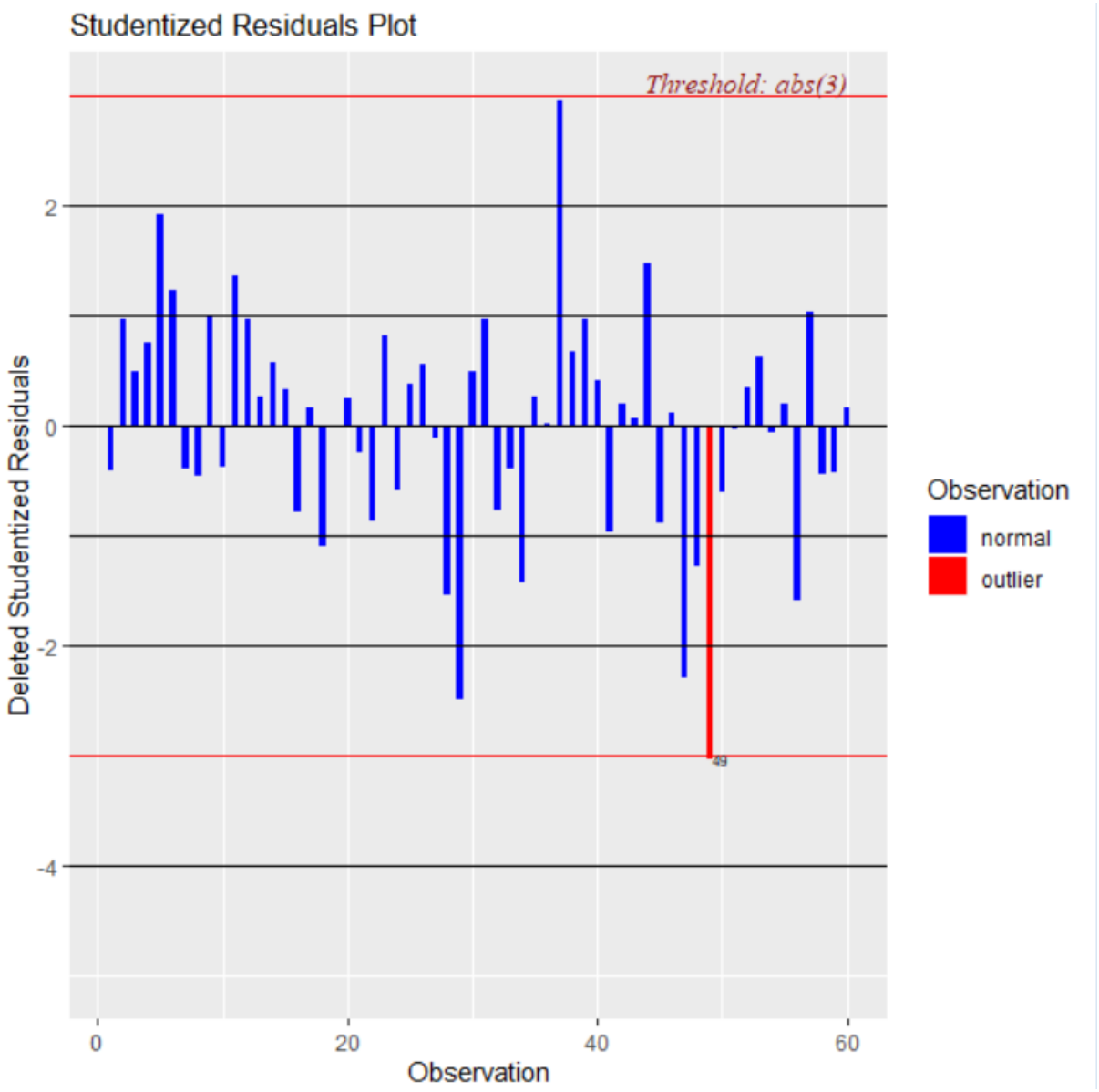
Poniżej naniesiono linię modelu regresji na wykres rozrzutu:



Jak widać, to dopasowanie skupia się bliżej przedziału międzykwartylowego niż poprzedni model, przez co można je traktować jako lepsze dopasowanie. Nie mniej jednak dalej błąd standardowy s jest wysoki.

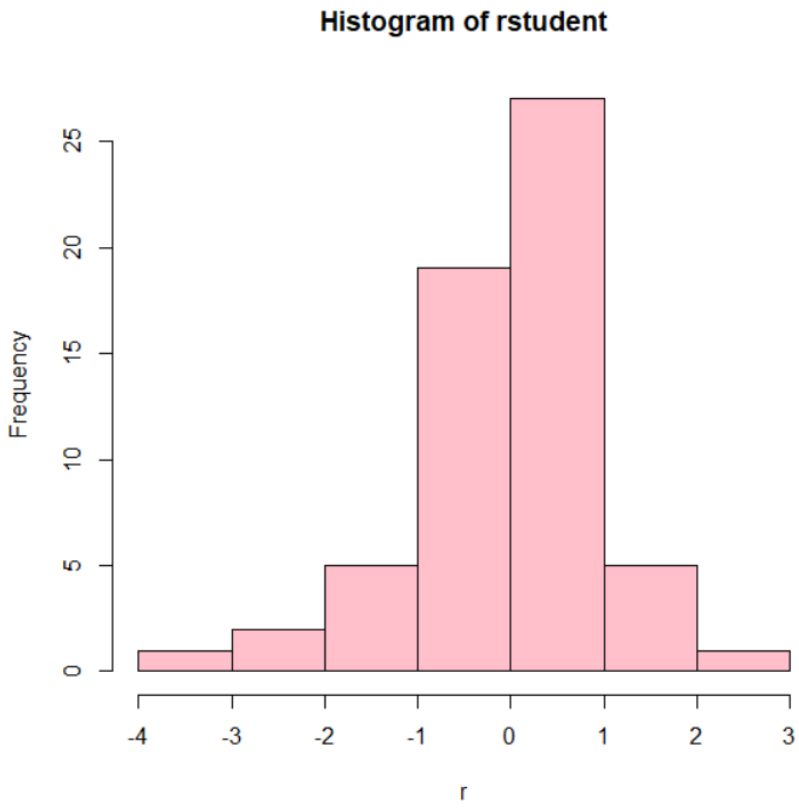
d) Model *Mortality(log(NOx))* bez obserwacji o dużych residuach studentyzowanych

Przy pomocy biblioteki *olsrr* wyznaczono wykres residuów studentyzowanych:

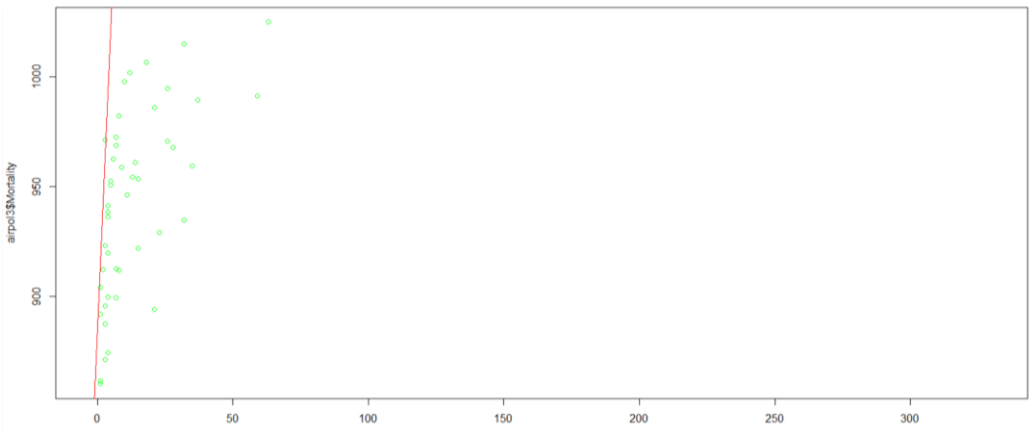


Powyższy wykres sugeruje, że jedynie obserwacja nr 49 jest nietypowa, natomiast obserwując model na wykresie rozrzutu można stwierdzić, że takich obserwacji jest więcej.

Wyznaczono histogram residuów studentyzowanych modelu:



Po analizie histogramu, przyjęto, że duże residua studentyzowane to te spoza zakresu (-1, 1). Usunięto obserwacje z tymi residuami. Dla tak okrojonego zbioru dopasowano nowy model regresji liniowej:



Błąd standardowy *s* dla tego modelu wyniósł 30.59, zatem najmniej spośród modeli dla tej zależności. Współczynnik nachylenia prostej tego modelu wyniósł a = 27.501 , zatem jest jeszcze bardziej stroma niż poprzednia.

Współczynnik determinacji R^2 opisuje jakość dopasowania modelu regresji do danych. Jego wartości należą do przedziału [0,1], im większa wartość tym lepsze dopasowanie. Dla wykonanych modeli współczynnik R^2 wyniósł:

- R^2 = 0.0871 dla modelu *Mortality(log(NOx))*
- R^2 = 0.4969 dla modelu *Mortality(log(NOx))* bez obserwacji o dużych residuach studentyzowanych

Jak widać, usunięcie obserwacji o dużych residuach studentyzowanych pozwoliło na prawie sześciokrotne poprawienie współczynnika determinacji. Taki model jest zatem lepiej dopasowany do naszych danych, jednak warto przy takiej operacji być ostrożnym, żeby uniknąć przeuczenia.

Zadanie 3 – Analiza zbioru `savings`, dotyczącego sytuacji ekonomicznej w 50 krajach (dane za lata 1960-1970)

a) Analiza wykorzystywanych zmiennych, model liniowy *Savings(dpi,ddpi,Pop15,Pop75)*

Przeprowadzono wstępną analizę statystyczną wykorzystywanych w zbiorze zmiennych:

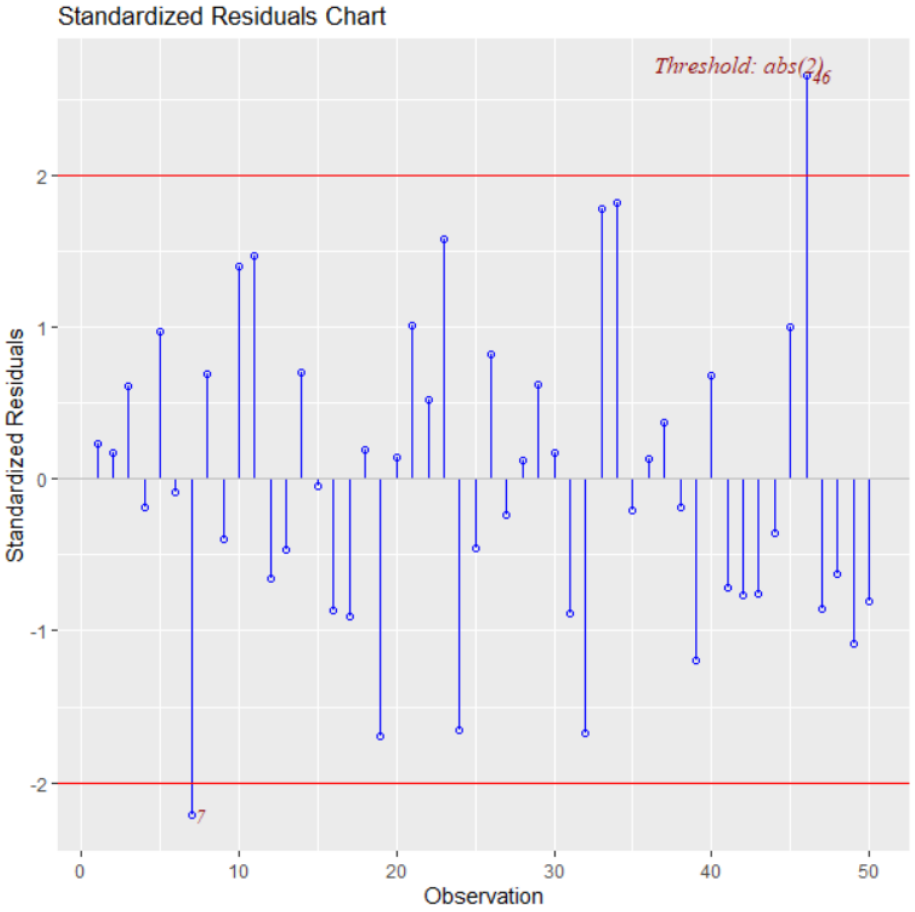
```
     Country       Pop15           Pop75            dpi              ddpi
Australia: 1   Min.   :21.44   Min.   :0.560   Min.   :  88.94   Min.   : 0.220
Austria  : 1   1st Qu.:26.22   1st Qu.:1.125   1st Qu.: 288.21   1st Qu.: 2.002
Belgium  : 1   Median :32.58   Median :2.175   Median : 695.66   Median : 3.000
Bolivia  : 1   Mean   :35.09   Mean   :2.293   Mean   :1106.79   Mean   : 3.758
Brazil   : 1   3rd Qu.:44.06   3rd Qu.:3.325   3rd Qu.:1795.62   3rd Qu.: 4.478
Canada   : 1   Max.   :47.64   Max.   :4.700   Max.   :4001.89   Max.   :16.710
(Other)  :44
    Savings
Min.   : 0.600
1st Qu.: 6.970
Median :10.510
Mean   : 9.671
3rd Qu.:12.617
Max.   :21.100
```

Jak widać po danych, zarówno dochód netto *dpi* jak i oszczędności *Savings* są bardzo zróżnicowane, mają znaczny rozstęp międzykwartylowy, a ich wartości maksymalne znacznie różnią się od mediany.

Dopasowano model linowy *Savings(dpi,ddpi,Pop15,Pop75).*

b) Wykres i analiza reszt modelu *Savings(dpi,ddpi,Pop15,Pop75)*

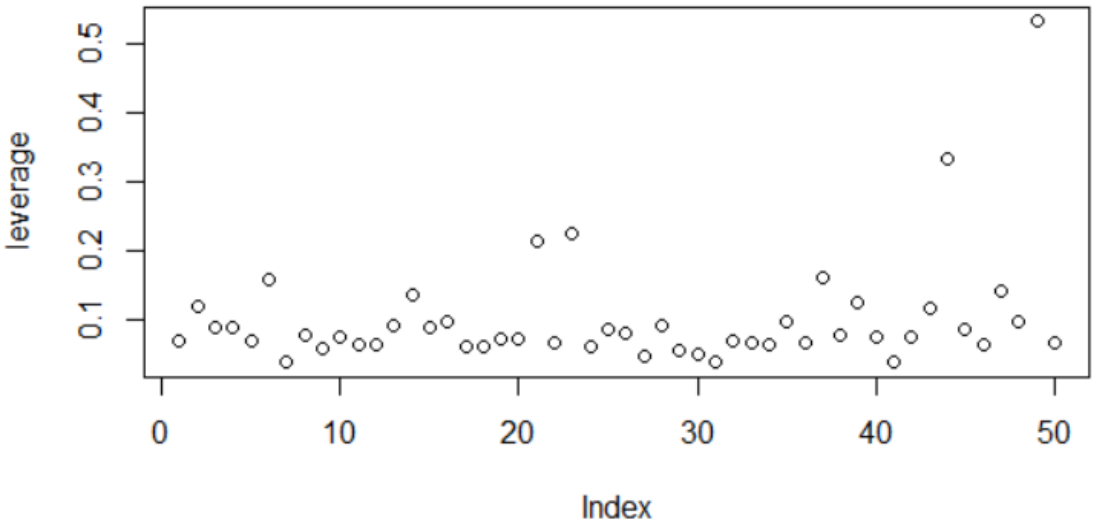
Za pomocą biblioteki *olsrr* wyznaczono wykres reszt modelu:



Jak widać na wykresie, najmniejsza reszta odpowiada krajowi nr 7, a więc Chile, zaś największa krajowi nr 46, Zambii. Z kolei najmniejszą wartość bezwzględną reszty ma kraj nr 15, czyli Niemcy.

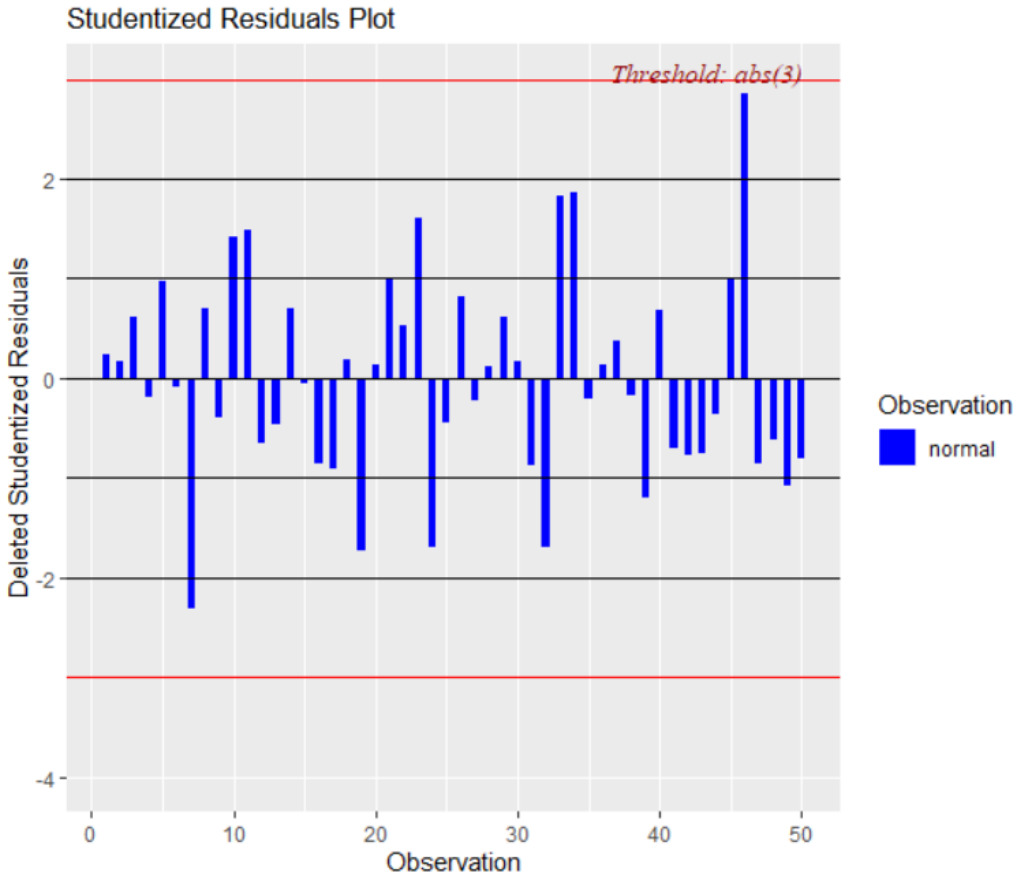
c) Analiza wartości dźwigni, reszty studentyzowane

Obliczone wartości dźwigni zostały naniesione na wykres:



Duża wartość dźwigni wystąpiła dla krajów nr 49, 44, 23 oraz 21. Są to kolejno kraje: Libia, USA, Japonia oraz Irlandia.

Wyznaczono reszty studentyzowane:

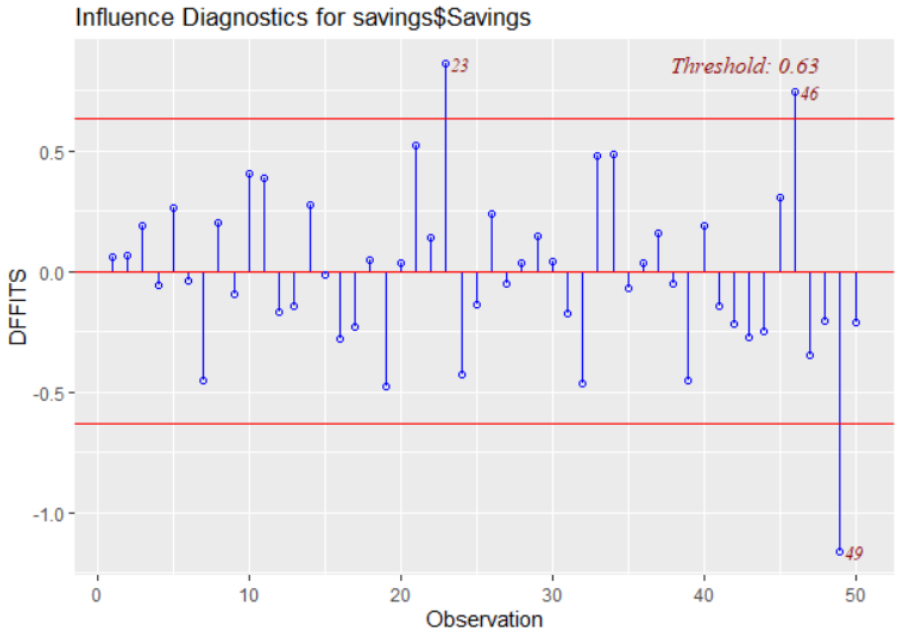


Duże wartości bezwzględne reszt studentyzowanych są dla kraju nr 46, 7, 34 oraz 33. Są to kolejno Zambia, Chile, Filipiny oraz Peru.

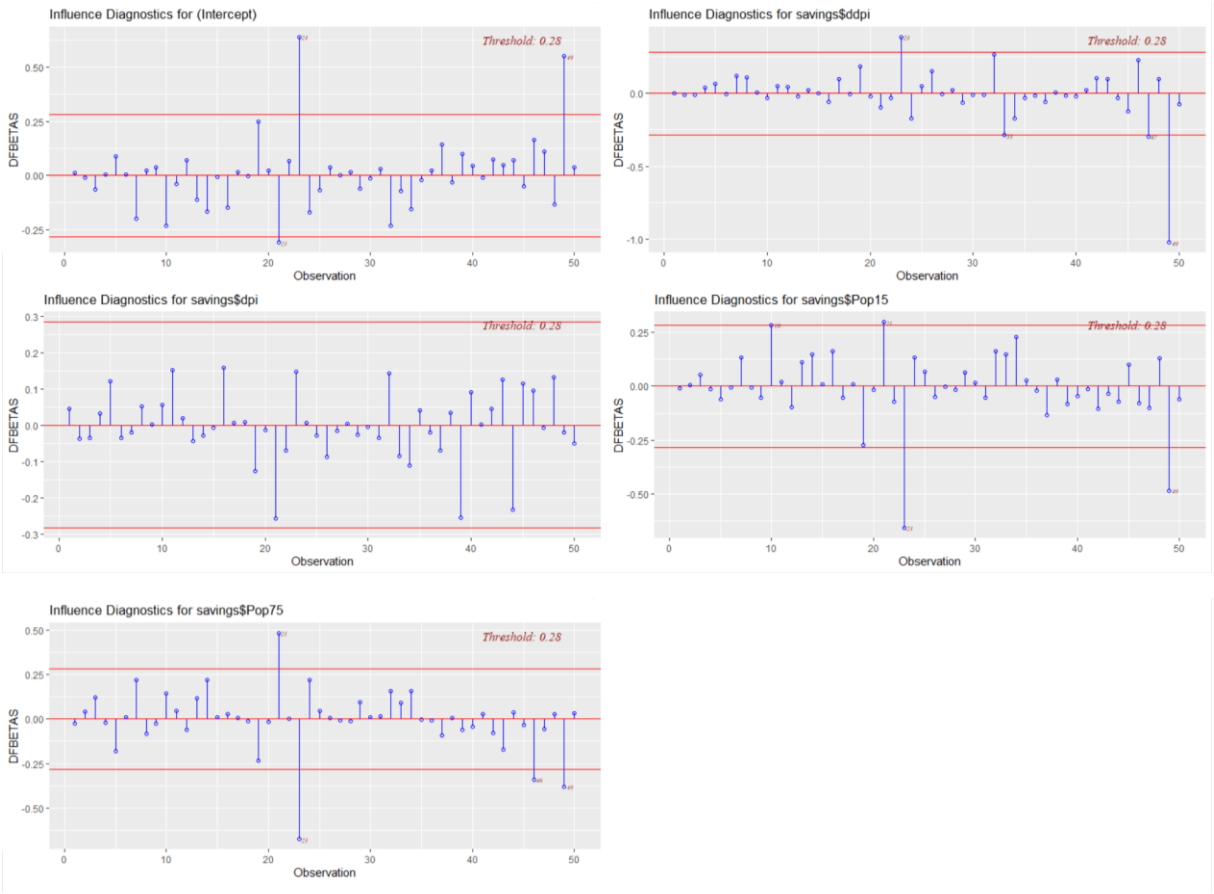
d) Miary DFFITS, DFBETAS oraz odległość Cooke'a

Miary DFFITS oraz DFBETAS wskazują na nietypowe wartości zmiennej objaśnianej *y*, czyli na obserwacje odstające.

Wyznaczono wartości miary DFFITS:



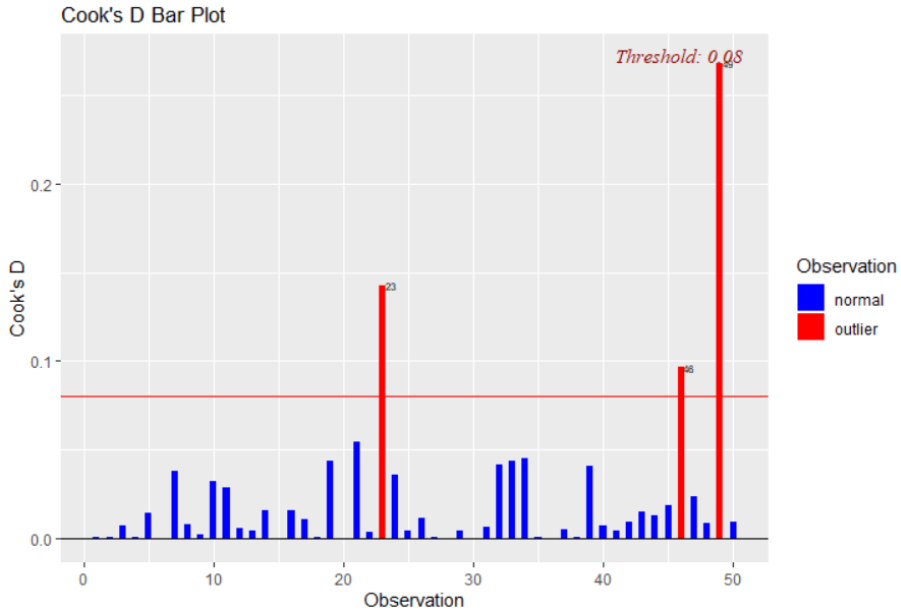
Oraz wartości DFBETAS:



Miara DFFITS jest wyznaczana dla całego modelu, z kolei DFBETAS dla każdej zmiennej opisującej osobno. Jako obserwacje nietypowe w statystyce DFFITS oznaczono wyniki państw nr 23, 46 oraz 49. Wskazania te dotyczą Japonii, Zambii oraz Libii, a zatem dotyczą państw, które mają duże reszty studentyzowane oraz wartości dźwigni.

We wskazaniach statystyki DFBETAS jako obserwacje nietypowe pojawiają się państwa nr 21, 23, 46 i 49, a zatem do obserwacji nietypowych z miary DFFITS doszła również Irlandia, która wcześniej została wymieniona w obserwacji dużej wartości dźwigni. Pojawiły się również obserwacje o nietypowej wartości tylko jednej ze zmiennych: dla *ddpi* obserwacja 33, czyli Peru gdzie tempo wzrostu dochodu wynosi zaledwie 0.57%, zaś dla zmiennej *Pop15* kraj nr 10, czyli Kostaryka, w której procent populacji poniżej 15 roku życia wyniósł aż 47.64.

Z kolei miara odległości Cooke'a wykrywa zarówno obserwacje wpływowe i odstające. Na poniższym wykresie przedstawiono te wartości:



Obserwacje wskazane jako nietypowe mają nr 23, 46 oraz 49, czyli są to wskazywane już wcześniej Japonia, Zambia i Libia. Przyjrzyjmy się bliżej tym obserwacjom:

	Country	Pop15	Pop75	dpi	ddpi	Savings
23	Japan	27.01	1.91	1257.28	8.21	21.10
46	Zambia	45.25	0.56	138.33	5.14	18.56
49	Libya	43.69	2.07	123.58	16.71	8.89

Japonia jest krajem, w którym współczynnik oszczędności jest najwyższym w całym zbiorze. Z kolei dochód netto przypadający na mieszkańca wcale nie jest maksimum w zbiorze, jest nawet poniżej średniej. Zrozumiałe jest zatem to, że przewidywana wartość odsetku oszczędności różni się od rzeczywistej.

Zambijczycy z kolei mają bardzo niski dochód netto przypadający na mieszkańca, poniżej 1. kwartyla w zbiorze, jednak ich procent oszczędności jest powyżej 3. kwartyla w zbiorze, zatem jest wyższy niż w większości państw. Dodatkowo, populacja ludności poniżej 15. roku życia to aż 45.25, co wpłynęło na fakt, że model przewidział niższy odsetek oszczędności niż jest w rzeczywistości.

Libia jest przypadkiem dosyć podobnym do Zambii – również ma młodą populację oraz niski dochód netto. Jednak na odstawanie od przewidywanej wartości wpływa również fakt, że mimo tych wskaźników ma również najwyższy w całym zbiorze procent tempa wzrostu dochodu, aż 16.71.

e) Model *Savings(dpi,ddpi,Pop15,Pop75)* bez obserwacji o najwyższym dystansie Cooke'a

Zgodnie z poleceniem wykonano model bez obserwacji o najwyższym dystansie Cooke'a, a więc bez wyników Libii.

Dla modelu z Libią:

```
Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007902
```

Dla modelu bez Libii:

```
Residual standard error: 3.795 on 44 degrees of freedom
Multiple R-squared:  0.3554,    Adjusted R-squared:  0.2968
F-statistic: 6.066 on 4 and 44 DF,  p-value: 0.0005616
```

Jak widać po powyższych statystykach, usunięcie Libii ze zbioru pozwoliło na nieznaczne poprawienie dopasowania modelu. Sprawdzono również jak będą wyglądać wyniki dopasowania dla regresji liniowej dopasowanej do zbioru bez obserwacji nietypowych z krajów analizowanych w poprzednim punkcie, a więc Japonii, Zambii i Libii:

```
Residual standard error: 3.441 on 42 degrees of freedom
Multiple R-squared:  0.3503,    Adjusted R-squared:  0.2885
F-statistic: 5.662 on 4 and 42 DF,  p-value: 0.0009741
```

Usunięcie kolejnych obserwacji również poprawiło dopasowanie modelu, natomiast również nie jest to znaczna poprawa jakości.

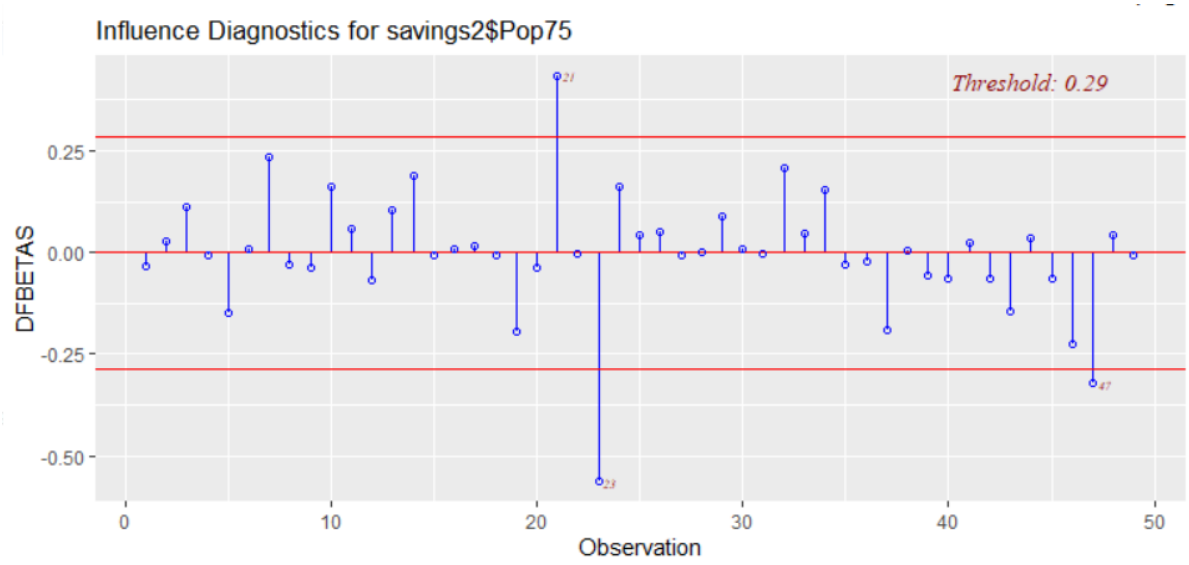
Zmienna *dpi* w modelu bez usuniętych obserwacji miała następujące współczynniki jakości dopasowania i istotności:

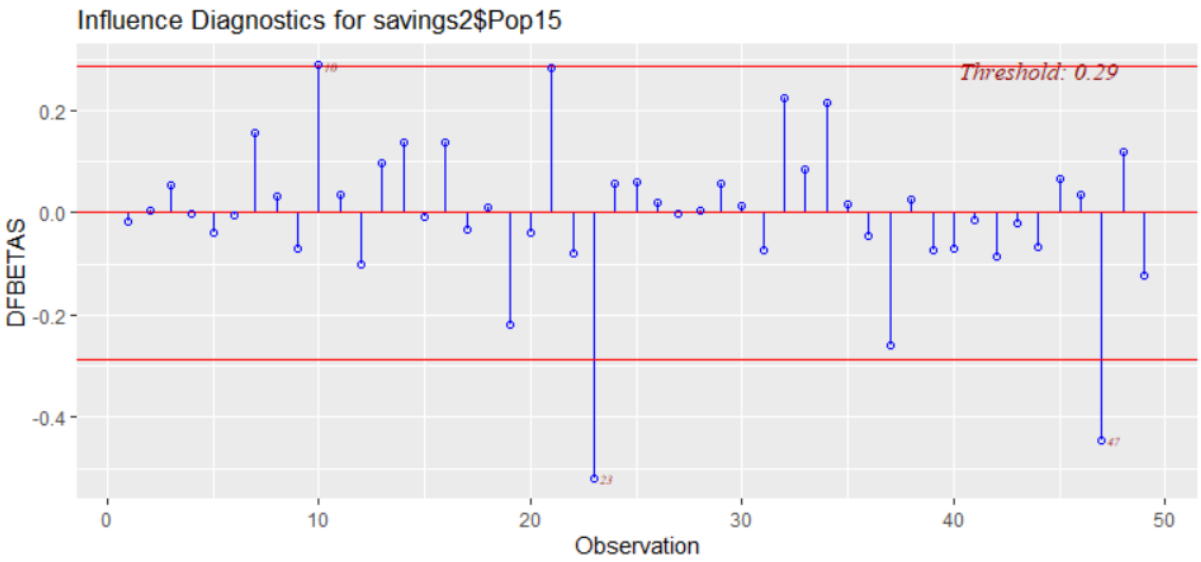
```
                Estimate Std. Error t value Pr(>|t|)
savings$dpi   -0.0003368  0.0009311  -0.362 0.719296
```

Wynik testu istotności wskazuje na to, że zmienna *dpi* nie wpływa na wartość *Savings* i model regresji liniowej – test `Pr(>|t|)` pokazuje, że istnieje aż 71% prawdopodobieństwa, że rozkład zmiennej *dpi* jest losowy i nie ma związku ze zmienną przewidywaną.

f) Wykres zmian wartości współczynników przy zmiennych *pop15, pop75*

Wyznaczono wykres zmian wartości współczynników przy zmiennych *pop15* oraz *pop75* w modelu z usuniętą obserwacją:





W przypadku obu zmiennych największy wpływ na model ma Japonia.

Zadanie 4 – zbiór `realest`, zależność ceny domu na przedmieściach Chicago od wybranych parametrów

a) Model *Price(Bedroom,Space,Room,Lot,Tax,Bathroom,Garage,Condition)*

Dopasowano model regresji do podanego zbioru danych. Podany model regresji miał następujące wyniki dopasowania:

```
Residual standard error: 7.337 on 17 degrees of freedom
Multiple R-squared:  0.7688,    Adjusted R-squared:   0.66
F-statistic: 7.065 on 8 and 17 DF,  p-value: 0.0003757
```

Współczynnik determinacji R^2 wyniósł 0.7688, zatem model regresji liniowej całkiem dobrze opisuje ceny domów w zależności od wybranych zmiennych.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.712572   9.514111   1.441   0.1677
Bedroom     -7.756208   3.109374  -2.494   0.0232 *
Space        0.011626   0.008981   1.295   0.2128
Room         5.097706   2.764303   1.844   0.0827 .
Lot          0.228063   0.195434   1.167   0.2593
Tax          0.003374   0.006859   0.492   0.6291
Bathroom     5.718372   4.276867   1.337   0.1988
Garage       3.613603   2.064997   1.750   0.0982 .
Condition   -2.162027   4.137400  -0.523   0.6080
```

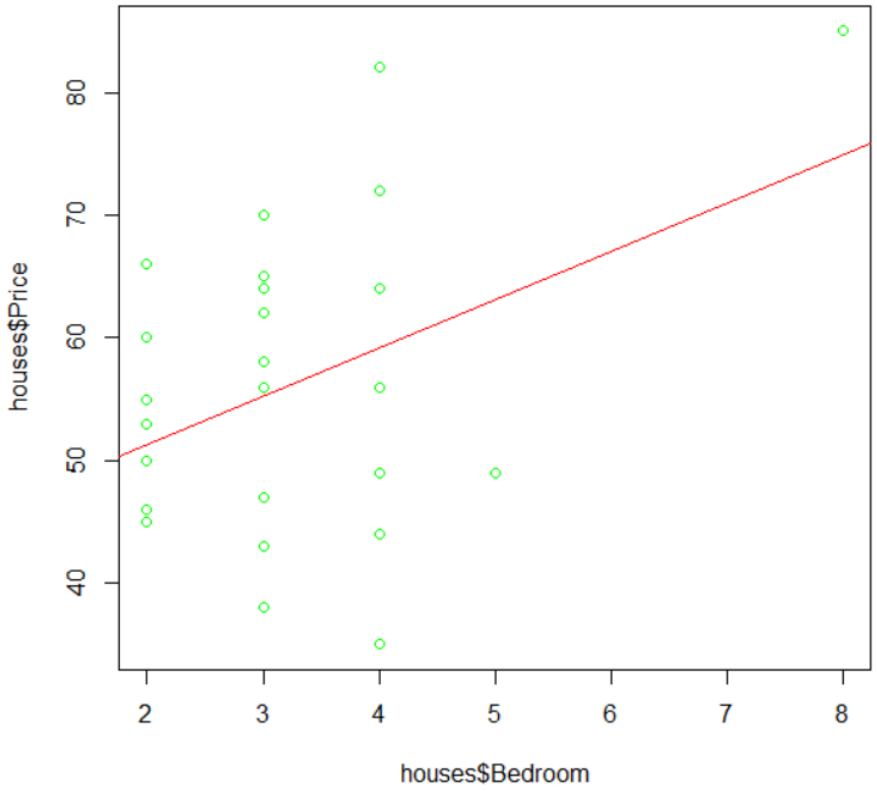
Wszystkie zmienne objaśniające, oprócz liczby sypialni są ustalone, mają wynik `Pr(>|t|)` powyżej 5% procent prawdopodobieństwa bycia losowo związanymi z ceną. W celu sprawdzenia wpływu liczby sypialni na cenę domu zwiększono ilość sypialni w każdym z rekordów zbioru uczącego i z pomocą wytrenowanego modelu przewidziano cenę dla domów z większą ilością sypialni:

	original	extra.bedroom	extra.bedroom...original
1	54.60339	46.84719	-7.756208
2	54.84794	47.09173	-7.756208
3	49.59620	41.83999	-7.756208
4	57.26890	49.51270	-7.756208
5	68.02621	60.27001	-7.756208
6	52.79164	45.03543	-7.756208
7	48.28970	40.53349	-7.756208
8	75.98696	68.23075	-7.756208
9	62.69083	54.93462	-7.756208
10	81.40150	73.64529	-7.756208
11	78.71303	70.95682	-7.756208
12	45.46288	37.70668	-7.756208
13	44.49659	36.74038	-7.756208
14	48.19841	40.44220	-7.756208
15	63.19758	55.44138	-7.756208
16	47.01449	39.25829	-7.756208
17	57.81247	50.05626	-7.756208
18	66.68484	58.92863	-7.756208
19	55.23538	47.47917	-7.756208
20	47.76297	40.00676	-7.756208
21	42.73852	34.98231	-7.756208
22	42.52067	34.76447	-7.756208
23	45.35031	37.59410	-7.756208
24	50.05983	42.30363	-7.756208
25	56.29809	48.54188	-7.756208
26	62.95066	55.19445	-7.756208

Okazuje się, że zwiększenie liczby sypialni wpłynęło na zmniejszenie się ceny domu, co można uznać za nieoczywiste – większa liczba pokojów powinna być atutem, a nie czymś co obniża cenę. Dla każdego z rekordów cena zmalała o 7.756208.

Taką pozornie błędną zmianę można tłumaczyć faktem, że w wytrenowanym modelu regresji liniowej zmienna *Bedroom* otrzymała współczynnik -7.756028. Należy pamiętać o tym, że ilość sypialni w zbiorze waha się pomiędzy 2 a 8, natomiast większość domów ma pomiędzy 2 a 4 sypialniami. Obserwując wykres rozrzutu można dojść do wniosku, że liczba sypialni różnie wpływa na cenę domów. Najtańszy w zbiorze ma 4 sypialnie, podobnie jak drugi najdroższy z domów, co może wzbudzić wątpliwości na temat czy sypialnie mogą wpływać liniowo na cenę domu.

W celu dalszego sprawdzenia tej zależności dopasowano model liniowy jedynie przy zmiennej *Bedroom*:




```
Residual standard error: 11.73 on 24 degrees of freedom
Multiple R-squared:  0.1655,    Adjusted R-squared:  0.1308
F-statistic: 4.761 on 1 and 24 DF,  p-value: 0.03914
```

Zarówno po graficznym jak i liczbowym opisie modelu widać, że związek liczby sypialni z ceną nie jest wprost liniowy. Porównano poprzednio uzyskane wyniki z przewidzianymi przez model:

	original	extra.bedroom	predicted.by.bedrooms	predicted.by.bedrooms.extra
1	54.60339	46.84719	51.32852	55.24910
2	54.84794	47.09173	51.32852	55.24910
3	49.59620	41.83999	55.24910	59.16968
4	57.26890	49.51270	55.24910	59.16968
5	68.02621	60.27001	55.24910	59.16968
6	52.79164	45.03543	59.16968	63.09025
7	48.28970	40.53349	63.09025	67.01083

```
8  75.98696      68.23075              55.24910              59.16968
9  62.69083      54.93462              59.16968              63.09025
10 81.40150      73.64529              59.16968              63.09025
11 78.71303      70.95682              74.85199              78.77256
12 45.46288      37.70668              51.32852              55.24910
13 44.49659      36.74038              55.24910              59.16968
14 48.19841      40.44220              59.16968              63.09025
15 63.19758      55.44138              59.16968              63.09025
16 47.01449      39.25829              51.32852              55.24910
17 57.81247      50.05626              55.24910              59.16968
18 66.68484      58.92863              59.16968              63.09025
19 55.23538      47.47917              51.32852              55.24910
20 47.76297      40.00676              59.16968              63.09025
21 42.73852      34.98231              55.24910              59.16968
22 42.52067      34.76447              55.24910              59.16968
23 45.35031      37.59410              51.32852              55.24910
24 50.05983      42.30363              51.32852              55.24910
25 56.29809      48.54188              51.32852              55.24910
26 62.95066      55.19445              55.24910              59.16968
```

Ceny przewidziane dla modelu bazującego jedynie na liczbie sypialni są oczywiście znacznie mniej zróżnicowane, co jest związane z faktem, że liczba sypialni w zbiorze również nie jest urozmaicona. Po przewidzianych wartościach widać jednak, że w modelu *Price(Bedroom)*, dla zwiększonej ilości sypialni w zbiorze przewidywane ceny domów urosły, przeciwnie do omawianego wcześniej modelu korzystającego z większej ilości zmiennych.

Na fakt, że w modelu *Price(Bedroom,Space,Room,Lot,Tax,Bathroom,Garage,Condition)* zmienna *Bedroom* otrzymała ujemny współczynnik, zaś w modelu *Price(Bedroom)* może wpływać fakt, że w pierwszym modelu na przewidywaną cenę wpływały również inne zmienne. Dodatkowo, na wyżej omawianym wykresie rozrzutu widać było, że przy tej samej liczbie sypialni ceny mogą być zarówno wysokie, jak i niskie.

b) Przewidywanie ceny dla konkretnego domu z wykorzystaniem modelu

Z pomocą uzyskanego modelu przewidziano cenę domu z parametrami podanymi w poleceniu:

```
predict(lmHouses,data.frame(Bedroom=3,Space=1500,Room=8,Lot=40,Tax=1
000,Bathroom=5,Garage=1,Condition=0))
```

Cena ta została oszacowana na 93.3675. Jest to większa cena niż najdroższego domu w zbiorze. Mamy tu jednak do czynienia z domem, który dla zmiennych *Room* i *Bathroom* ma wartości powyżej trzeciego kwartyla, a w naszym modelu liniowym te zmienne mają duże znaczne dodatnie współczynniki. Analizując współczynniki przy zmiennych w naszym modelu, liczba łazienek ma

największy pozytywny wpływ na cenę domu. Nasz przykładowy dom ma 5 łazienek, podczas gdy w zbiorze ich maksimum wynosi 3. Również liczba pokoi ma duży pozytywny wpływ na cenę domu, a w naszym przykładzie wynosi 8 – jedynie dwa domy w zbiorze mają więcej.

Zadanie 5 – zbiór `gala_data` zawierający informację o liczbie gatunków żółwi na danych wyspach archipelagu Galapagos

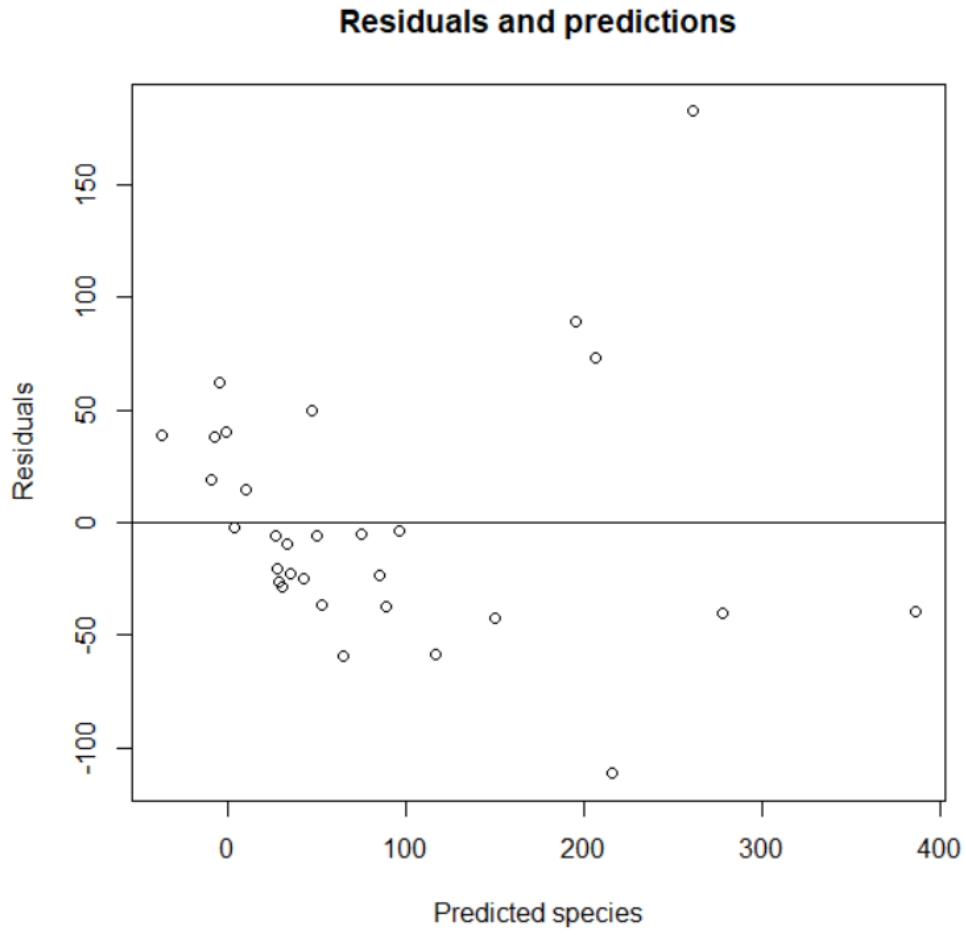
a) Model *Species(Area,Elevation,Nearest,Scruz,Adjacent)*

Dopasowano model liniowy do wybranych zmiennych. Zestawiono współczynniki opisujące model:

```
Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Współczynnik determinacji R^2 wyniósł 0.7658, przez co można uznać, że model dość dobrze dopasował się do danych.

Według założeń modelu regresji liniowej, residua nie powinny rosnąć wraz ze wzrostem wartości prognozowanych, powinny utrzymywać stałą wartość. Sporządzono wykres residuów w zależności od wartości przewidywanej:



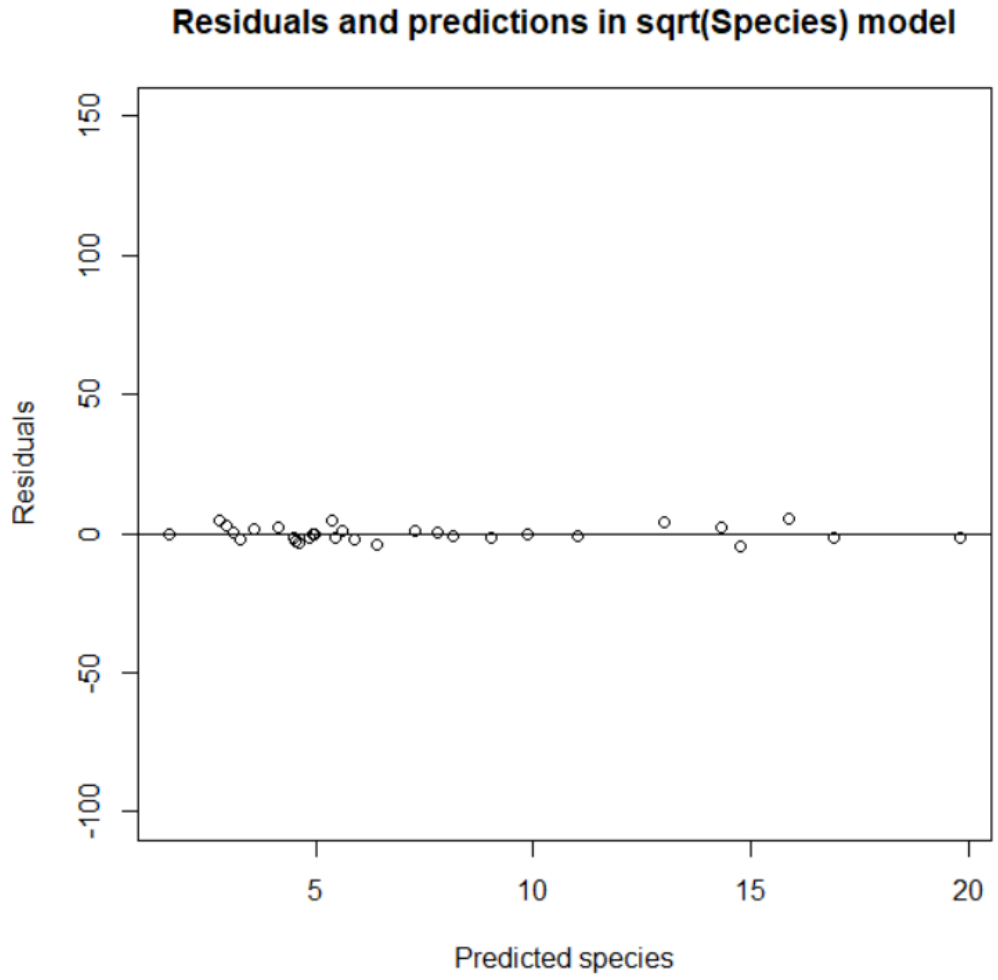
Jak widać na wykresie, residua rosną wraz ze zmienną objaśnianą.

b) Poprawa modelu

W celu usunięcia problemu zmiennej wariancji residuów spierwiastkowano zmienną objaśnianą. Sporządzono model oparty na takiej zmiennej:

```
Residual standard error: 2.774 on 24 degrees of freedom
Multiple R-squared:  0.7827,    Adjusted R-squared:  0.7374
F-statistic: 17.29 on 5 and 24 DF,  p-value: 2.874e-07
```

Wskaźniki tego modelu wyglądają lepiej niż w pierwszym przypadku. Współczynnik R^2 wyniósł 0.7827, zatem jest to zmiana, która wskazuje na poprawę dopasowania. Błąd standardowy *s* wynosił 60.98 dla pierwszego modelu, z kolei teraz zaledwie 2.774. Sprawdzono czy udało się rozwiązać problem zmienności residuów:



Spierwiastkowanie zmiennej objaśnianej pozwoliło na wyeliminowanie zmienności residuów, które teraz utrzymują się w przedziale [-4;4] niezależnie od wartości prognozowanej.

W nowym modelu sprawdzono, która ze zmiennych objaśniających ma najwyższą wartość w teście `p - value`:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.3919243  0.8712678   3.893 0.000690 ***
Area        -0.0019718  0.0010199  -1.933 0.065080 .
```

```
Elevation    0.0164784  0.0024410   6.751 5.55e-07 ***
Nearest      0.0249326  0.0479495   0.520 0.607844
Scruz       -0.0134826  0.0097980  -1.376 0.181509
Adjacent    -0.0033669  0.0008051  -4.182 0.000333 ***
```

Najwyższą wartość w teście uzyskała zmienna *Nearest*, opisująca odległość od najbliższej wyspy. Sporządzono model bez tej zmiennej:

```
Residual standard error: 4.627 on 25 degrees of freedom
Multiple R-squared:  0.3701,    Adjusted R-squared:  0.2693
F-statistic: 3.672 on 4 and 25 DF,  p-value: 0.01743
```

Ta operacja nie wpłynęła pozytywnie na dopasowanie modelu, poprzednia ilość zmiennych była bardziej optymalna.

	Species(Area,Elevation,Nearest,Scruz,Adjacent)	*Species(Area,Elevation,Nearest,Scruz,Adjacent) ze spierwiastkowanym Species*	*Species(Area,Elevation,Scruz,Adjacent) ze spierwiastkowanym Species*
R^2	0.7658	0.7827	0.3701
R^2_{Adj}	0.7171	0.7374	0.2693

Jak wskazano we wcześniejszej analizie, najlepsza wartość współczynnika determinacji została uzyskana przy modelu ze spierwiastkowaną zmienną objaśnianą i ze wszystkimi dostępnymi zmiennymi objaśniającymi.

Zadanie 6 – klasyfikacja irysów w zbiorze `irys` z wykorzystaniem drzew decyzyjnych

a) Sporządzenie drzewa, analiza modelu

Po wczytaniu danych, podzielono je na zbiór treningowy i testowy. Wytrenowano drzewo:

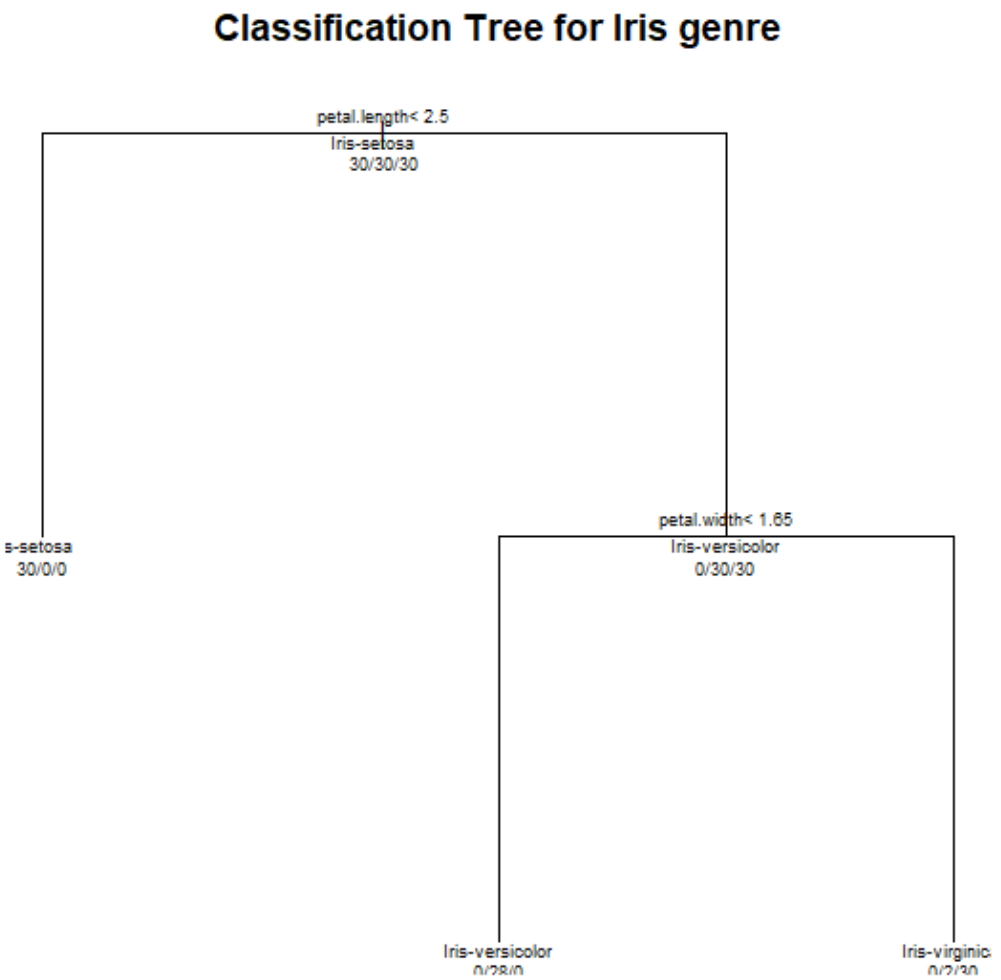
```
Classification tree:
rpart(formula = class ~ sepal.length + sepal.width + petal.length +
    petal.width, data = train, method = "class")
Variables actually used in tree construction:
[1] petal.length petal.width
Root node error: 60/90 = 0.66667
n= 90

CP nsplit      rel error  xerror     xstd
1 0.50000      0  1.000000 1.18333 0.064526
```

```
2 0.46667       1   0.500000 0.83333 0.078567
3 0.01000       2   0.033333 0.05000 0.028382
```

Drzewo w formie tekstowej:

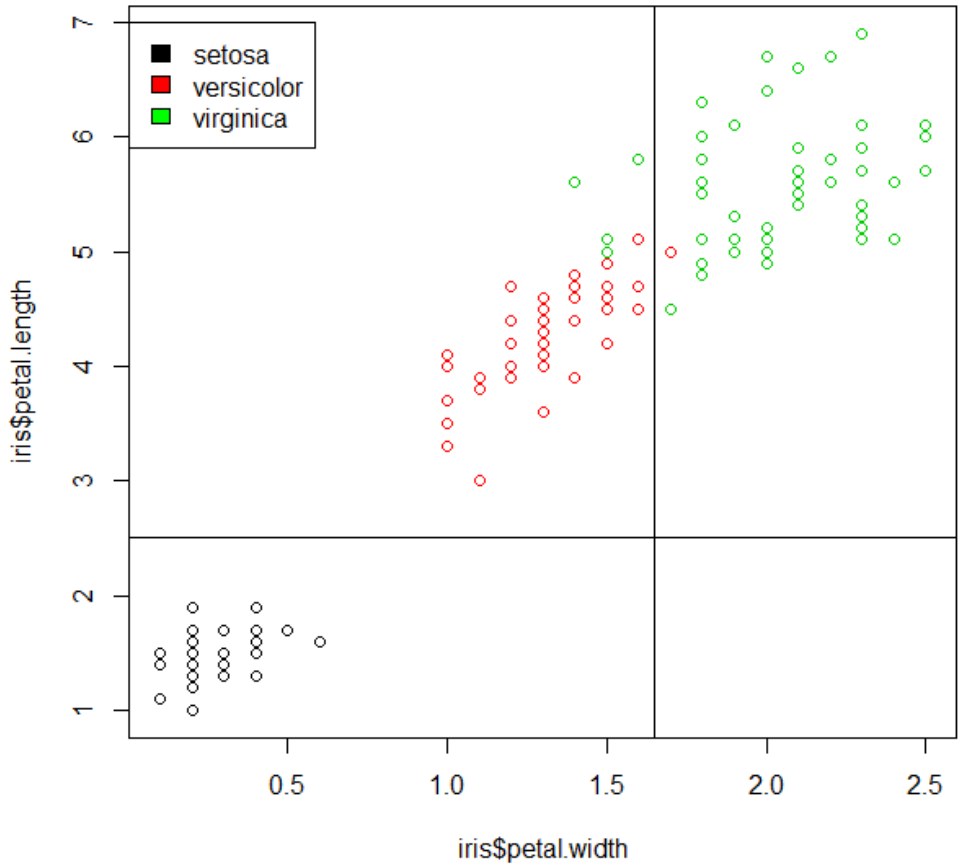
```
             class  Iris Iris Iris
     Iris-setosa [1.00  .00  .00] when petal.length <  2.5
 Iris-versicolor [ .00 1.00  .00] when petal.length >= 2.5 & petal.width <  1.7
  Iris-virginica [ .00  .06  .94] when petal.length >= 2.5 & petal.width >= 1.7
```

W formie graficznej:



Wytrenowane drzewo użyło jedynie 2 z 4 zmiennych do sklasyfikowania irysów – *petal.length* oraz *petal.width*. W pierwszej regule przyjęło, że jeśli długość płatka jest mniejsza niż 2.5, to irys jest z gatunku *setosa*. W przeciwnym wypadku sprawdzana jest szerokość płatka – jeśli jest mniejsza niż 1.65, to irys jest z gatunku *versicolor*, jeśli nie to jest to *virginica*.

Sporządzono wykres rozrzutu długości płatka od jego szerokości, dla całego zbioru irysów. Na wykres ten naniesiono również granice reguł decyzyjnych przyjętych przez drzewo decyzyjne:



Jak widać na powyższym wykresie, to że drzewo decyzyjne skorzystało jedynie z tych dwóch zmiennych jest uzasadnione – sporządzone reguły decyzyjne oparte na szerokości i długości płatka irysa pozwalają na dość skuteczne wyznaczenie jego gatunku.

b) Macierz błędów, trafność modelu

Wygenerowano macierz błędów:

pred	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	20	0	0
Iris-versicolor	0	20	4
Iris-virginica	0	0	16

Model drzewa decyzyjnego błędnie zaklasyfikował 4 z 60 przypadków zbioru testowego, co daje trafność 93.3%. Zaklasyfikowanie irysów do gatunku *setosa* było nieomylne – reguła sprawdzająca długość płatka jest trafna. Z kolei przy rozróżnieniu gatunku *versicolor* od *virginica* pojawił się błąd – 4 kwiaty z gatunku *virginica* zostały błędnie zaklasyfikowane przez model do gatunku *versicolor*.

Zadanie 7 – klasyfikacja irysów w zbiorze `irys` z wykorzystaniem klasyfikatora *k*-NN

a) Normalizacja danych, algorytm 3-najbliższych sąsiadów

Po wczytaniu danych znormalizowano je za pomocą metody Min-Max. Następnie podzielono na zbiór treningowy i testowy. Uruchomiono algorytm 3-najbliższych sąsiadów. Zbiór testowy ma 37 elementów.

Sprawdzono czy zmiana liczby sąsiadów wpłynie na zmianę klasyfikacji, i faktycznie tak też się stało – przy liczbie sąsiadów k=7 jeden z przypadków został zaklasyfikowany inaczej.

b) Ewaluacja klasyfikatora

Macierz błędu dla klasyfikatora 3-NN:

	test_class		
nn3	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	12	0	0
Iris-versicolor	0	12	1
Iris-virginica	0	1	11

Trafność tego klasyfikatora wyniosła 94.59%.