Regresja liniowa – projekt

Aleksandra Mazurek, Radosław Kachel

Źródło: Helmut Spaeth,
Mathematical Algorithms for Linear Regression,
Academic Press, 1991,
ISBN 0-12-656460-4.

S Weisberg,
Applied Linear Regression,
New York, 1980, pages 32-33

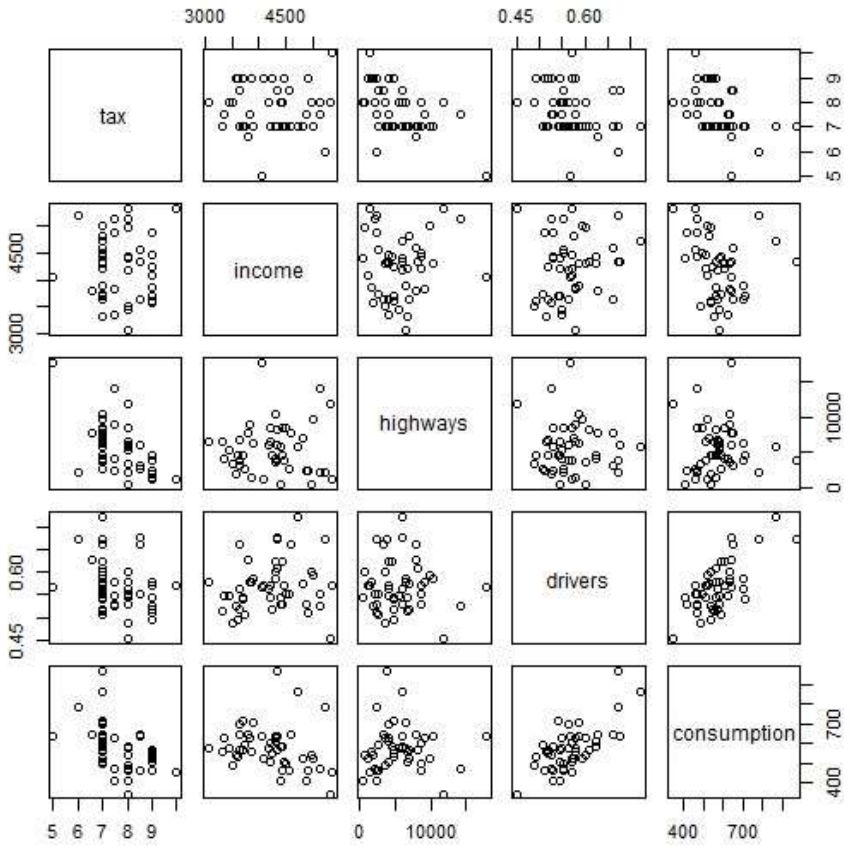
https://people.sc.fsu.edu/~jburkardt/datasets/regression/x16.txt

Opis datasetu: Dane przedstawiają ilość zużycia benzyny w ciągu jednego roku w 48 stanach US. Zmienne opisują kolejno:

X1 - *tax* - podatek od benzyny (cent/galon),
X2 - *income* – przychód na osobę (dolar),
X3 - *highways* – łączna długość autostrad w danym stanie (mile),
X4 - *drivers* – odsetek osób z prawem jazdy,
Y - *consumption* – zużycie benzyny (miliony galonów).

Jako zmienną zależną wybraliśmy *consumption*.

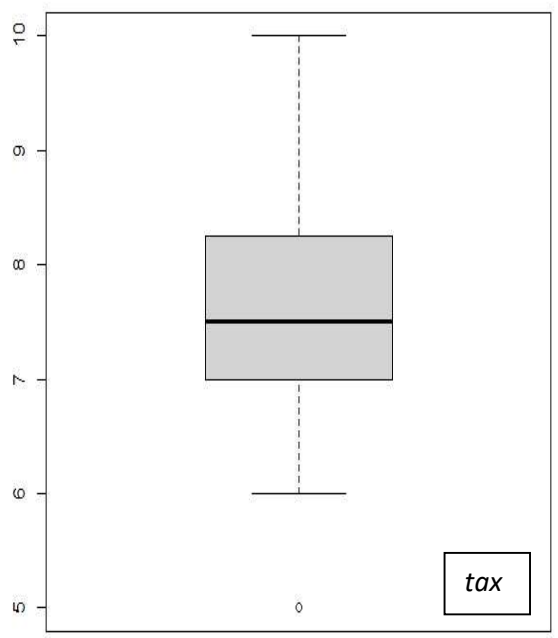
1. Przegląd danych

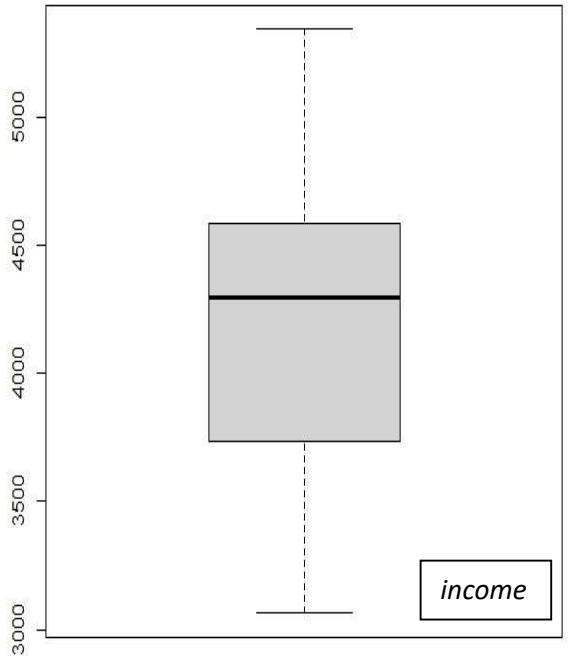


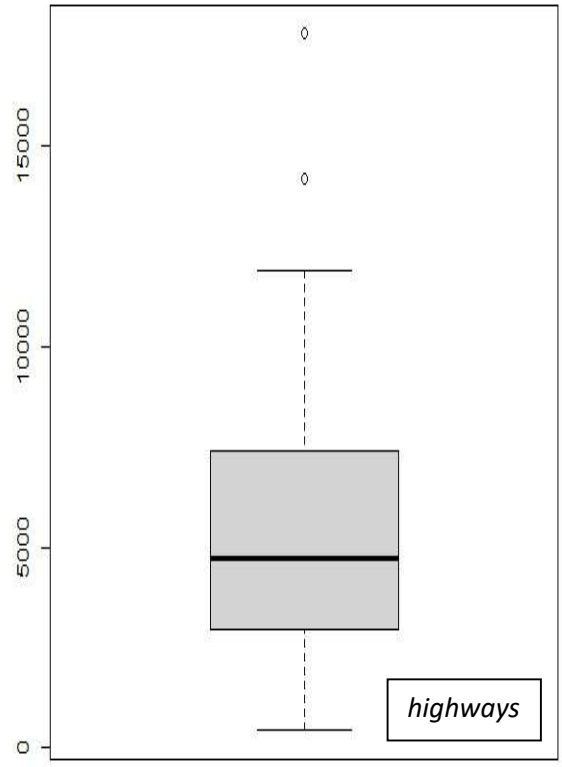

```
> cor(petrol)
                    tax      income    highways    drivers consumption
tax          1.00000000  0.01266516 -0.52213014 -0.2880372 -0.45128028
income       0.01266516  1.00000000  0.05016279  0.1570701 -0.24486207
highways    -0.52213014  0.05016279  1.00000000 -0.0641295  0.01904194
drivers     -0.28803717  0.15707008 -0.06412950  1.0000000  0.69896542
consumption -0.45128028 -0.24486207  0.01904194  0.6989654  1.00000000
```

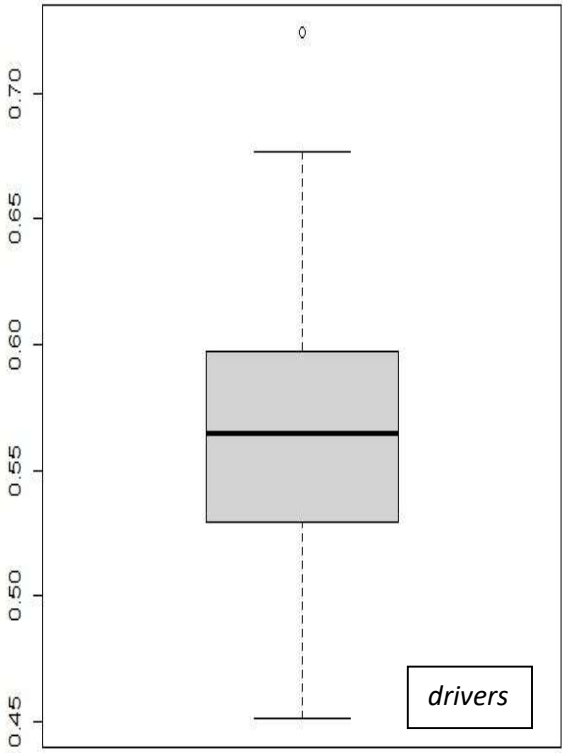
Na powyższych wykresach można zauważyć lekką zależność liniową między zmiennymi *drivers* i *consumption*, co potwierdza macierz korelacji, natomiast pozostałe wykresy nie sugerują innych zależności ze zmienną *consumption*. Dużą korelację widać również między zmiennymi niezależnymi *tax* oraz *highways* co może sugerować, iż jedna z tych zmiennych w późniejszym etapie wyjaśni drugą, więc może się okazać nieistotna w regresji.

```
> summary(petrol)
      tax             income       highways       drivers       consumption
 Min.   : 5.000   Min.   :3063   Min.   :  431   Min.   :0.4510   Min.   :344.0
 1st Qu.: 7.000   1st Qu.:3739   1st Qu.: 3110   1st Qu.:0.5298   1st Qu.:509.5
 Median : 7.500   Median :4298   Median : 4736   Median :0.5645   Median :568.5
 Mean   : 7.668   Mean   :4242   Mean   : 5565   Mean   :0.5703   Mean   :576.8
 3rd Qu.: 8.125   3rd Qu.:4579   3rd Qu.: 7156   3rd Qu.:0.5952   3rd Qu.:632.8
 Max.   :10.000   Max.   :5342   Max.   :17782   Max.   :0.7240   Max.   :968.0
```

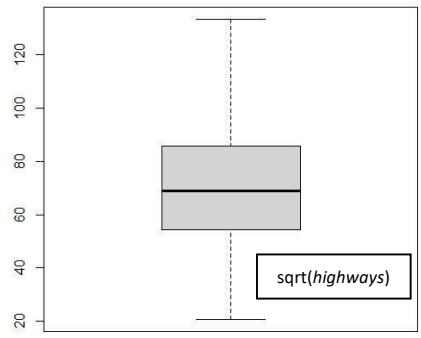

10
9
8
7
6
5
tax

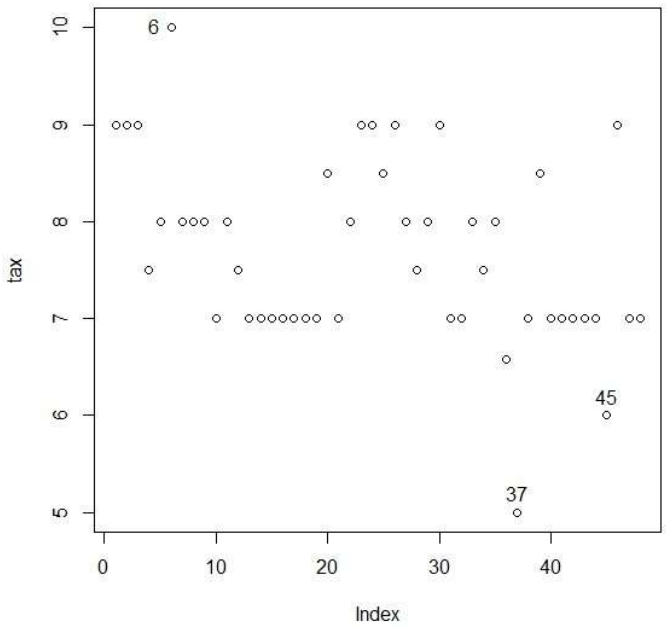

5000
4500
4000
3500
3000
income

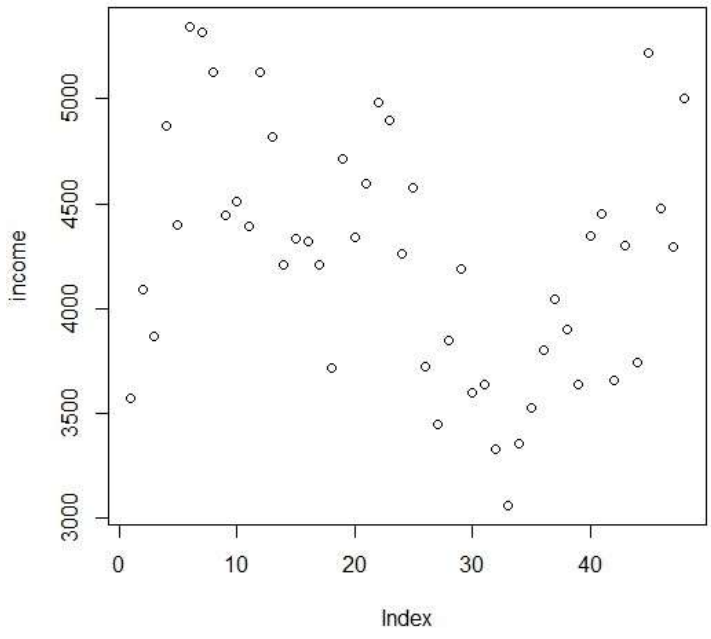

15000
10000
5000
0
highways

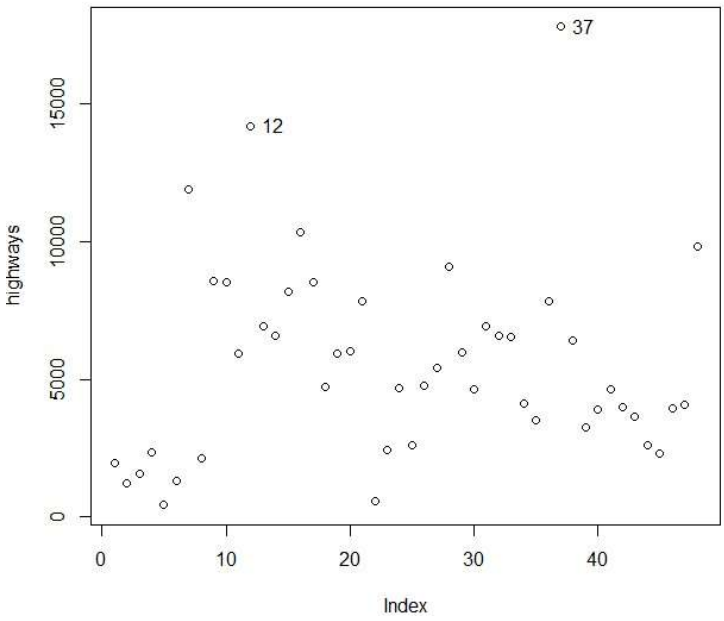

0.70
0.65
0.60
0.55
0.50
0.45
drivers

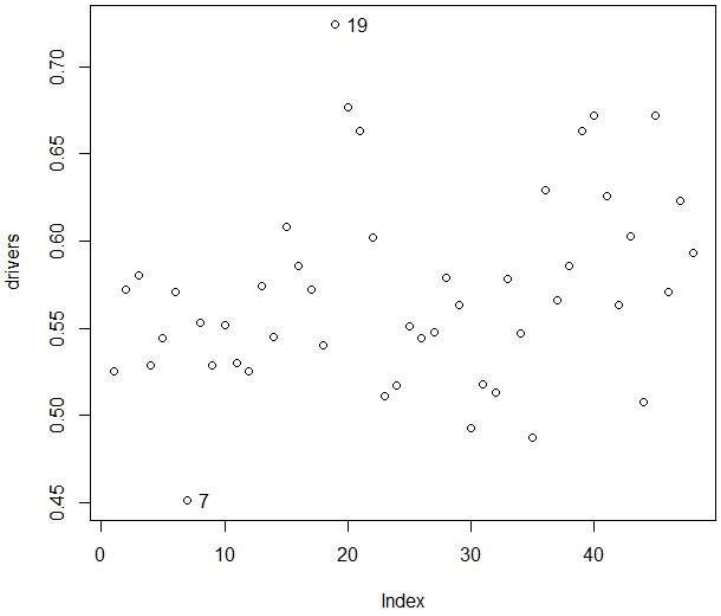
Zmienne posiadają niewiele obserwacji odstających, rozkłady wydają się być w miarę symetryczne poza zmienną *highways* oraz *tax*, jednakże w zmiennej *tax* wynika to ze specyfiki podatków (dyskretność wartości). Na zmiennej *highways* spróbowaliśmy transformacji log oraz sqrt. Logarytm niewiele zmienił, aczkolwiek pierwiastek usunął obserwacje odstające poprawiając wykres pudełkowy (rys. ->), ale i tak przy badaniu modelu regresji z sqrt(*highways*) analiza przy pomocy kryterium Akaike odrzucała zmienną sqrt(*highways*).

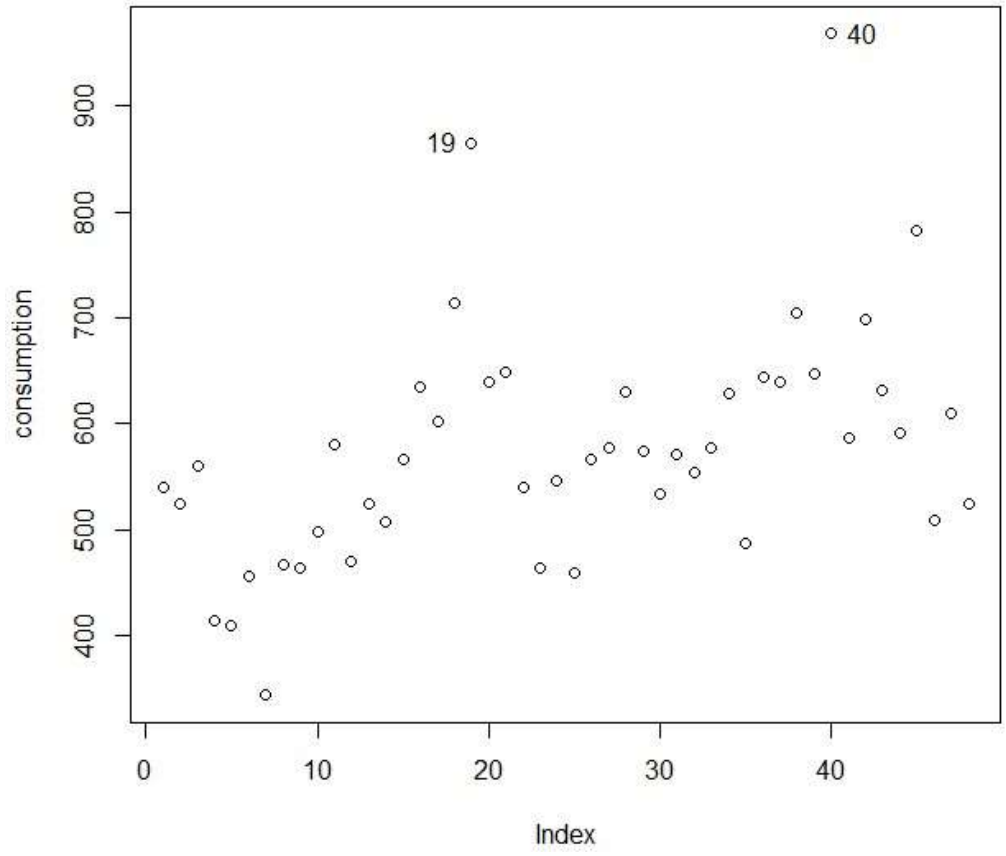












Na powyższych wykresach można wstępnie zidentyfikować podejrzane obserwacje odstające. Dla zmiennej *income* nie stwierdziliśmy obecności obserwacji odstających, natomiast dla pozostałych zmiennych niezależnych zauważamy pewne nieco odstające wartości. Podobnie dla zmiennej zależnej niektóre obserwacje można uznać za odstające.

2. Model pełny i redukcja ilości zmiennych objaśniających

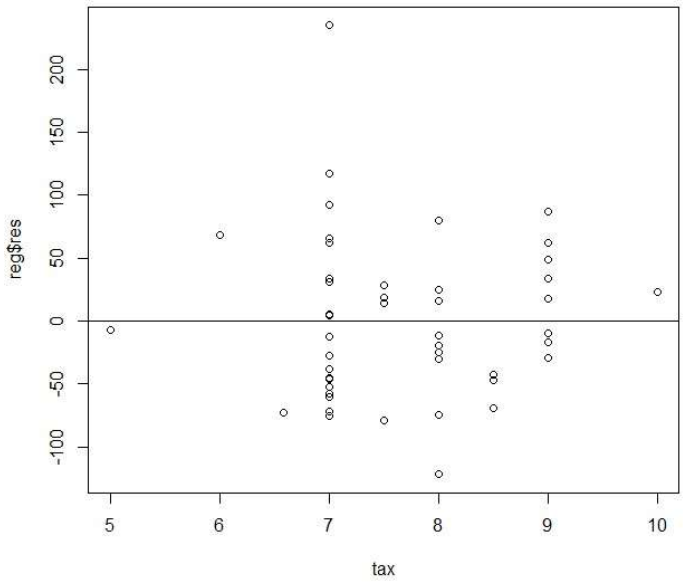
Model pełny jest postaci: **Y = A0 + A1*X1 + A2*X2 + A3*X3 + A4*X4**

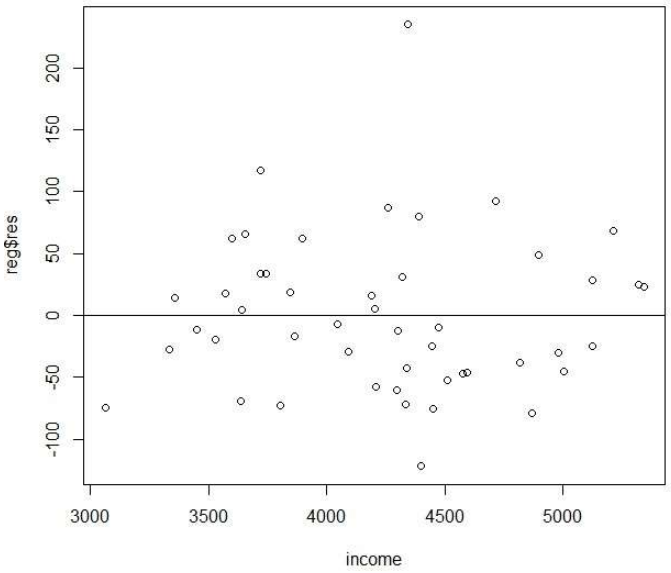
```
> reg<-lm(consumption~.,petrol)
> summary(reg)

Call:
lm(formula = consumption ~ ., data = petrol)

Residuals:
    Min      1Q  Median      3Q     Max
-122.03  -45.57  -10.66   31.53  234.95

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.773e+02  1.855e+02   2.033 0.048207 *
tax         -3.479e+01  1.297e+01  -2.682 0.010332 *
income      -6.659e-02  1.722e-02  -3.867 0.000368 ***
highways    -2.426e-03  3.389e-03  -0.716 0.477999
drivers      1.336e+03  1.923e+02   6.950 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.31 on 43 degrees of freedom
Multiple R-squared:  0.6787,    Adjusted R-squared:  0.6488
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

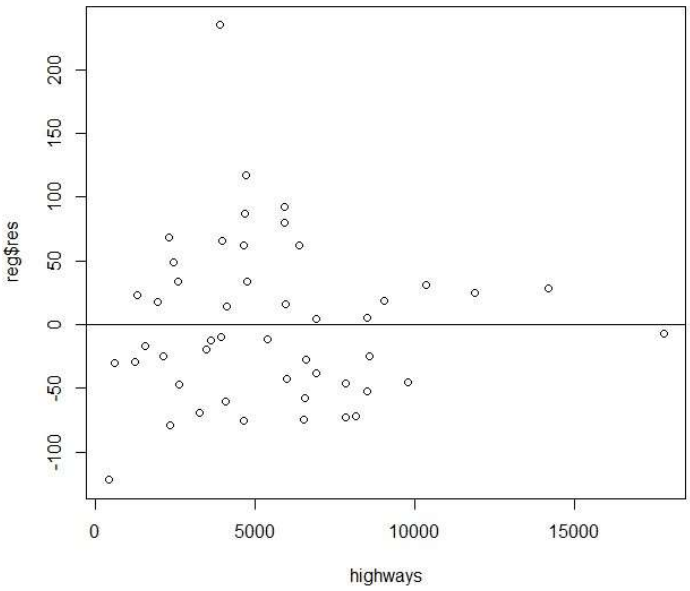
Z powyższego podsumowania modelu regresji otrzymaliśmy, że najbardziej istotne są zmienne *income* i *drivers*, a zmienna *tax* oraz wyraz wolny są nieco mniej istotne. Zmienna *highways* według modelu jest nieistotna, co może być spowodowane wcześniej zaobserwowaną korelacją ze zmienną *tax*, która może ją wyjaśniać. Wartość R^2 mogłaby być lepsza, natomiast ze względu na fakt, że dane opisują ludzkie zachowania związane z kwestią woli i wyboru, co powoduje, że są cięższe do przewidzenia, wartość ta jest zadowalająca.

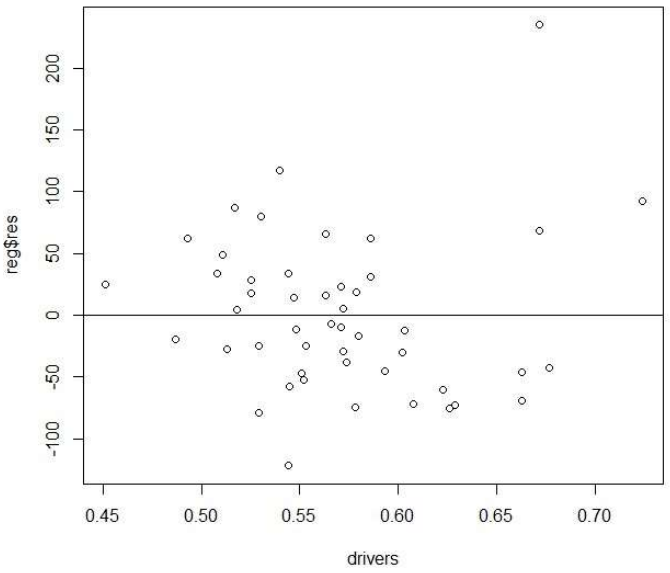
```
> vif(reg)
     tax   income highways  drivers
1.625676 1.043274 1.496937 1.216355
```

Wartości współczynnika inflacji wariancji dla modelu pełnego nie przekraczają 10 oraz są bliskie 1, więc są zadowalające. Przy zmiennych *highways* oraz *tax*, możemy zaobserwować nieco wyższe współczynniki, co ponownie sugeruje zależność tych zmiennych.

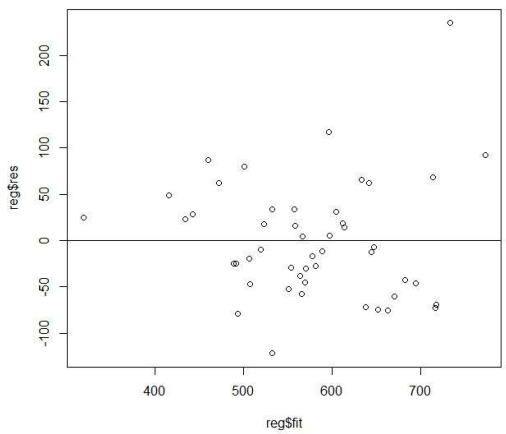




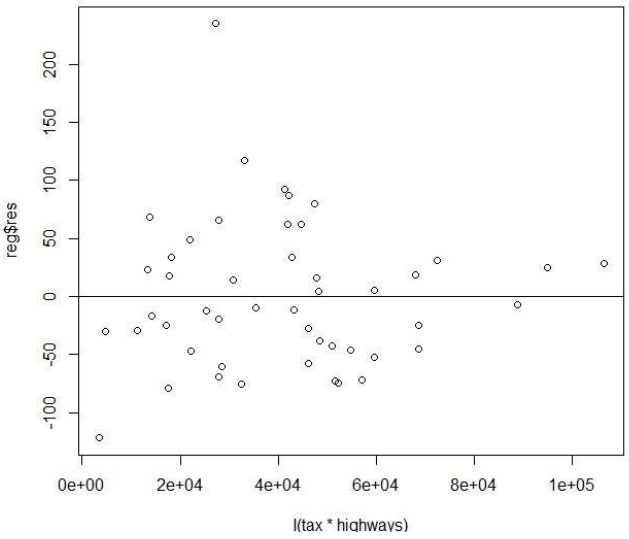


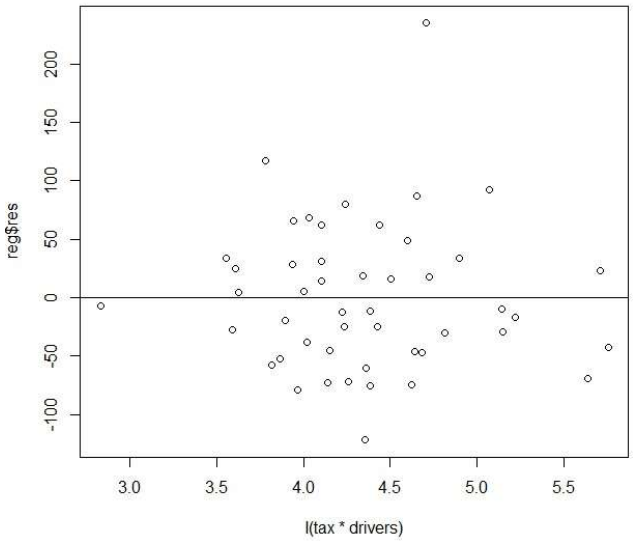


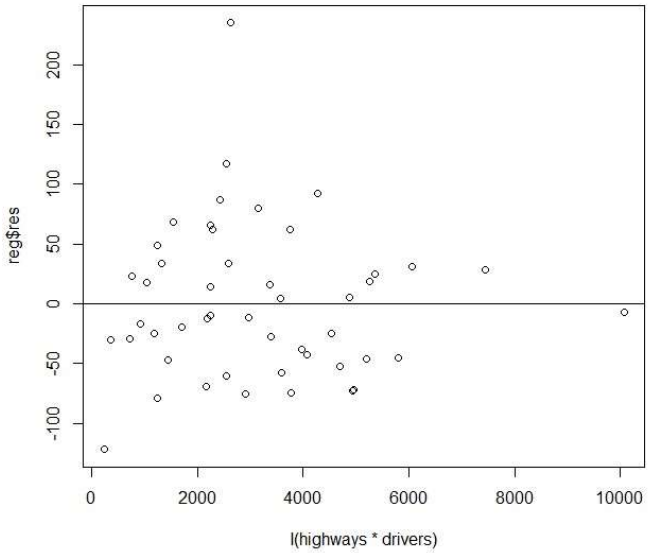
Na wykresach reszt pierwszych dwóch zmiennych nie zauważamy zależności, a wariancja wydaje się być stała. Na kolejnych dwóch wykresach wariancja również wydaje się być stała, pomimo tego, że skrajne obserwacje mogłyby sugerować malenie wariancji oraz pewną zależność funkcyjną. Mimo, że uważamy, że ilość obserwacji wzbudzająca podejrzenia jest zbyt mała aby stwierdzić niestałość wariancji lub zależność, podjęliśmy próby implementacji regresji dla stransformowanych zmiennych *highways* i *drivers* (log, sqrt). Transformacje te nie przyniosły jednak poprawy wykresów reszt od podanych zmiennych.

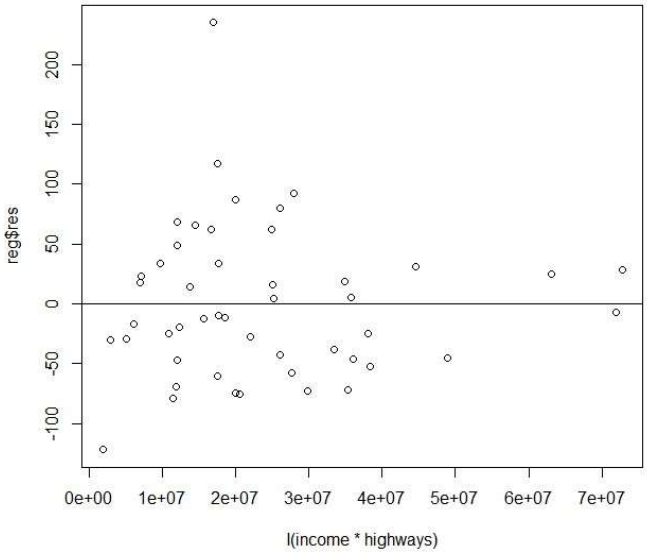


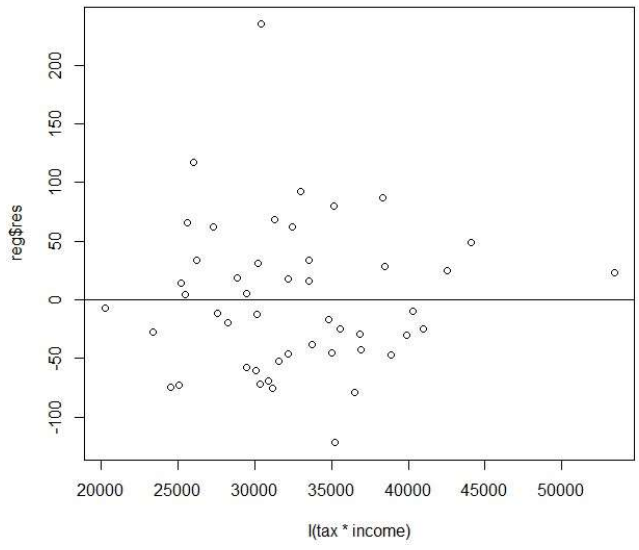
Na wykresie reszt od wartości dopasowanych nie zauważamy zależności.

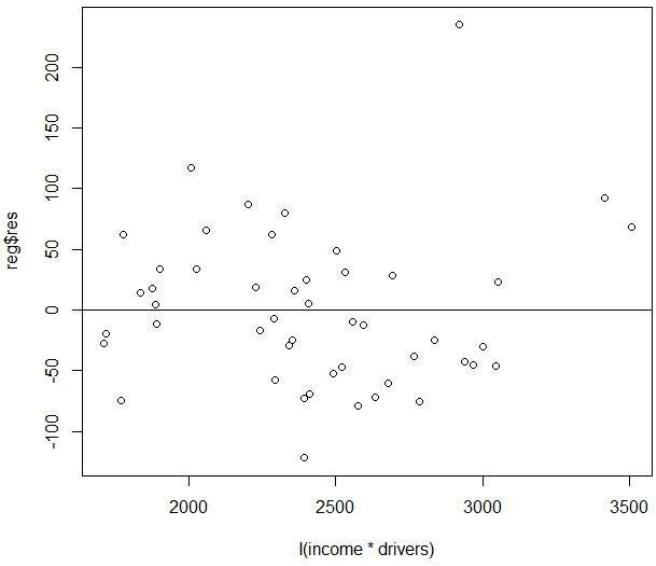

reg$res
200
150
100
50
0
-50
-100
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
l(tax * highways)


reg$res
200
150
100
50
0
-50
-100
3.0
3.5
4.0
4.5
5.0
5.5
l(tax * drivers)

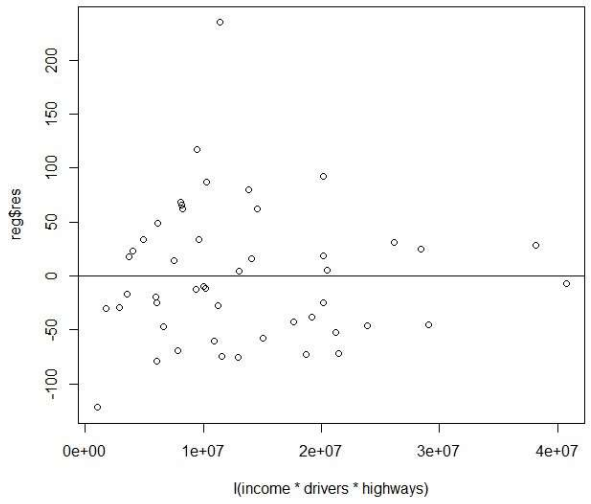

reg$res
200
150
100
50
0
-50
-100
0
2000
4000
6000
8000
10000
l(highways * drivers)

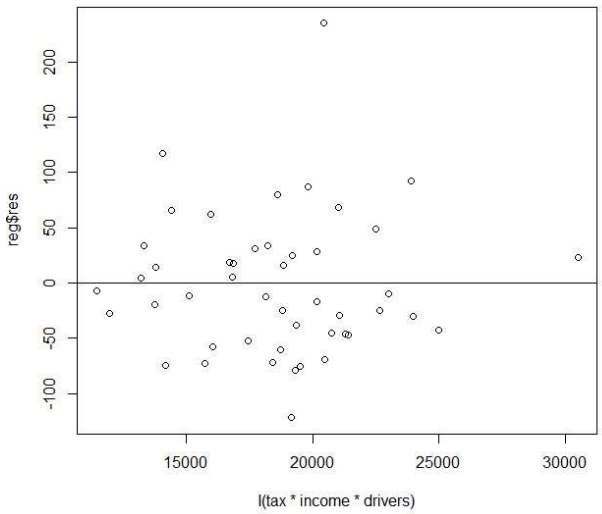

reg$res
200
150
100
50
0
-50
-100
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
7e+07
l(income * highways)

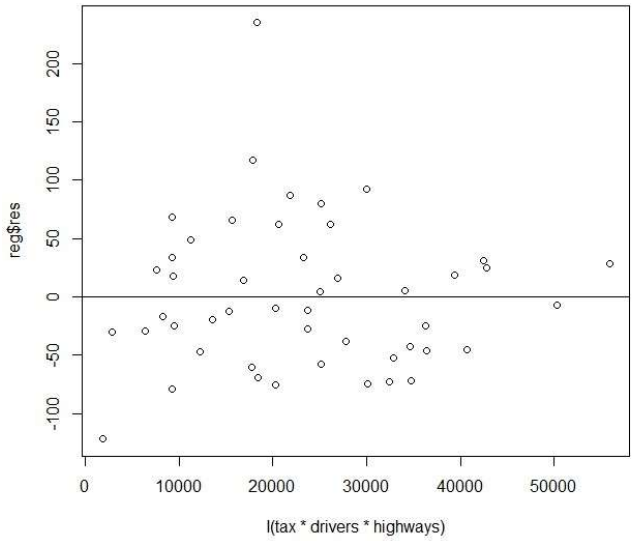

reg$res
200
150
100
50
0
-50
-100
20000
25000
30000
35000
40000
45000
50000
l(tax * income)

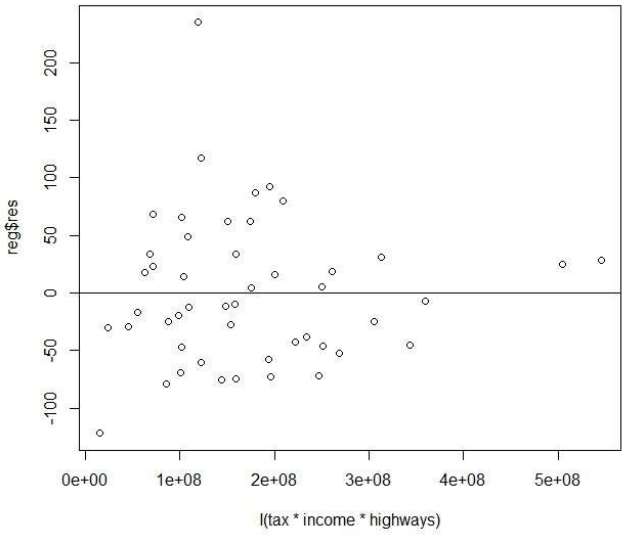

reg$res
200
150
100
50
0
-50
-100
2000
2500
3000
3500
l(income * drivers)

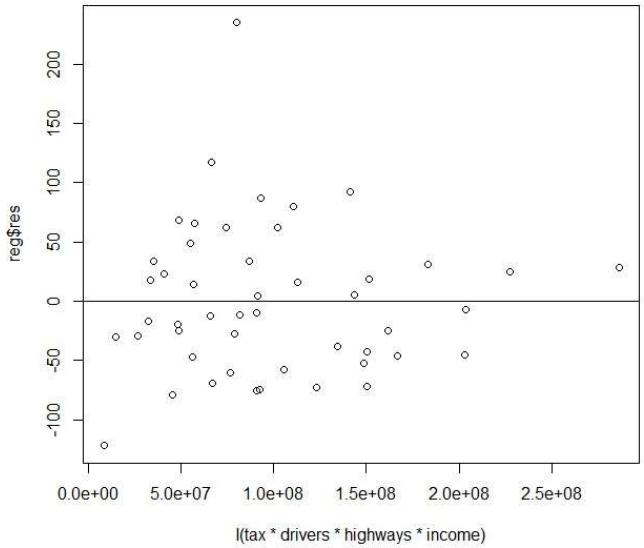
Ze względu na wysokie wartości zmiennej *highways*, wykresy interakcji pozostałych zmiennych z nią wyglądają podobnie jak wykres samej zmiennej *highways*, stąd to co może wyglądać jak malejąca wariancja na tych wykresach jest tylko pozorne i nie jest powodem do zmartwień. Na wykresie reszt od interakcji *income*drivers* możemy doszukiwać się pewnych zależności, co mogłoby wstępnie sugerować potrzebę dodania tejże interakcji do modelu.



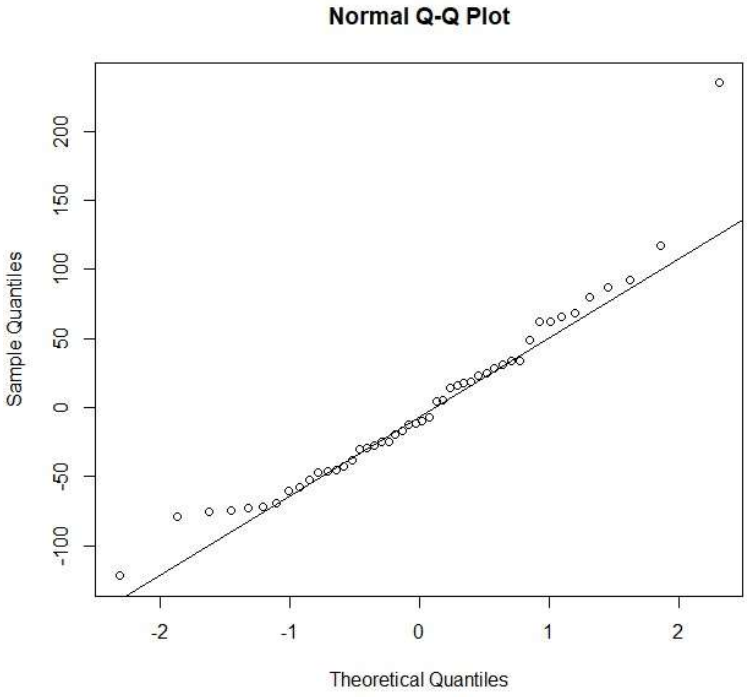


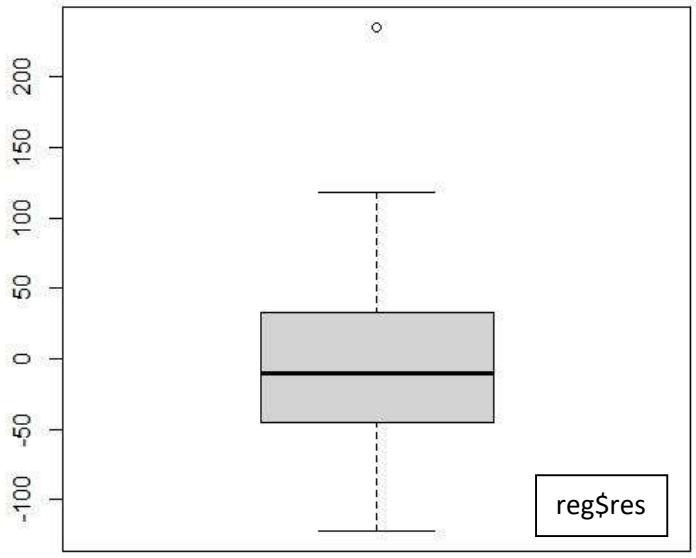






Wykresy reszt od interakcji potrójnych oraz poczwórnej nie wzbudzają podejrzeń, więc w następnym kroku rozważymy tylko model z podwójnymi interakcjami.



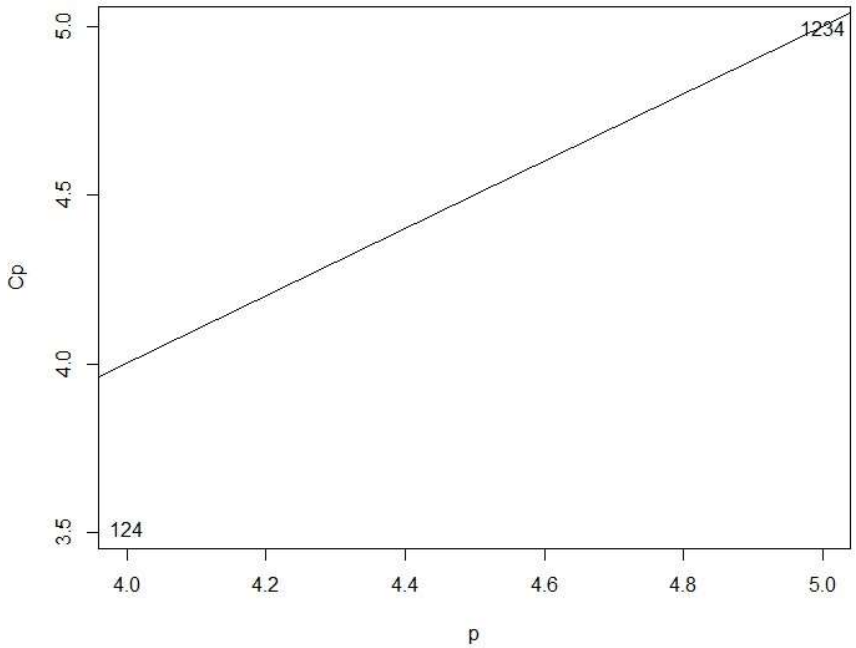



```
> shapiro.test(reg$res)

        Shapiro-Wilk normality test

data:  reg$res
W = 0.93918, p-value = 0.0151
```

Po analizie wykresu kwantyl-kwantyl reszt oraz teście Shapiro-Wilka stwierdzamy, że należy odrzucić hipotezę o normalności reszt.

```
> step(reg)
Start:  AIC=407.37
consumption ~ tax + income + highways + drivers

            Df Sum of Sq    RSS    AIC
- highways   1      2252 191302 405.94
<none>                   189050 407.37
- tax        1     31632 220682 412.80
- income     1     65729 254779 419.69
- drivers    1    212355 401405 441.51

Step:  AIC=405.94
consumption ~ tax + income + drivers

            Df Sum of Sq    RSS    AIC
<none>                   191302 405.94
- tax        1     33742 225044 411.74
- income     1     69532 260834 418.82
- drivers    1    243586 434889 443.36

Call:
lm(formula = consumption ~ tax + income + drivers, data = petrol)

Coefficients:
(Intercept)          tax       income      drivers
  307.32790    -29.48381     -0.06802   1374.76841
```

```
> reg0<-lm(consumption~1,petrol)
> step(reg0,scope=c(upper=consumption~tax+income+highways+drivers,lower=consumption~1))
Start:  AIC=453.87
consumption ~ 1

           Df Sum of Sq    RSS    AIC
+ drivers   1    287448 300918 423.68
+ tax       1    119823 468543 444.94
+ income    1     35277 553090 452.90
<none>                  588366 453.87
+ highways  1       213 588153 455.85

Step:  AIC=423.68
consumption ~ drivers

           Df Sum of Sq    RSS    AIC
+ income    1     75874 225044 411.74
+ tax       1     40084 260834 418.82
<none>                  300918 423.68
+ highways  1      2410 298509 425.30
- drivers   1    287448 588366 453.87

Step:  AIC=411.74
consumption ~ drivers + income

           Df Sum of Sq    RSS    AIC
+ tax       1     33742 191302 405.94
<none>                  225044 411.74
+ highways  1      4362 220682 412.80
- income    1     75874 300918 423.68
- drivers   1    328045 553090 452.90

Step:  AIC=405.94
consumption ~ drivers + income + tax

           Df Sum of Sq    RSS    AIC
<none>                  191302 405.94
+ highways  1      2252 189050 407.37
- tax       1     33742 225044 411.74
- income    1     69532 260834 418.82
- drivers   1    243586 434889 443.36

Call:
lm(formula = consumption ~ drivers + income + tax, data = petrol)

Coefficients:
(Intercept)      drivers       income          tax
  307.32790   1374.76841     -0.06802    -29.48381
```



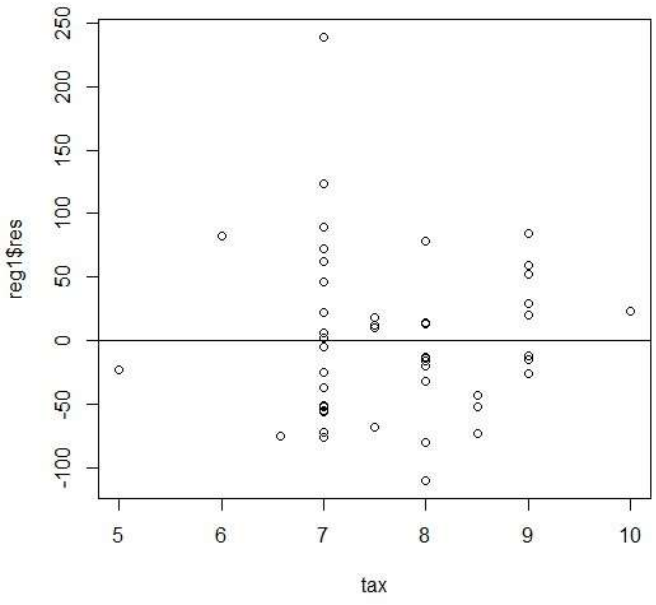

```
> leaps(x,y,method="adjr2")->ra
> maxadjr(ra)
  1,2,4 1,2,3,4     2,4
  0.653   0.649   0.601
```

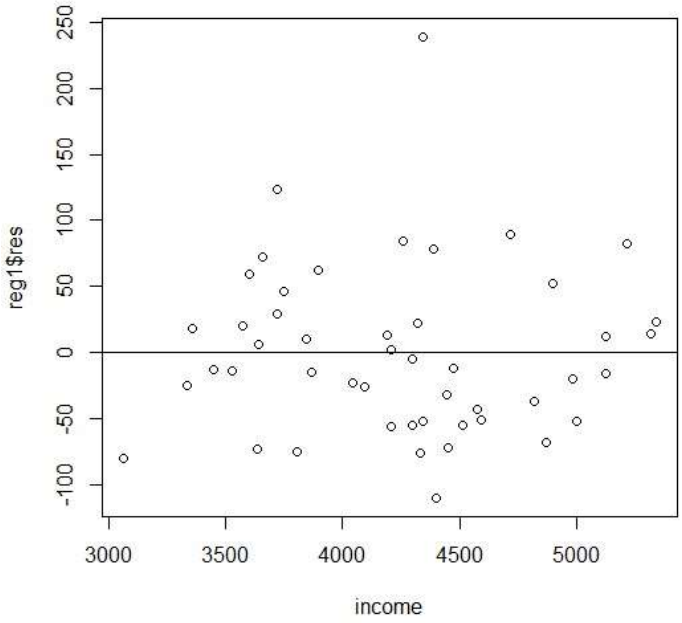
Kryteria Akaike, Mallowa oraz Maximum Adjusted R^2 sugerują jednogłośnie, że najlepszy jest model pełny z wyłączeniem zmiennej *highways*. Po raz kolejny mamy powód do stwierdzenia, że zmienna *highways* może być wyjaśniona przez inne.

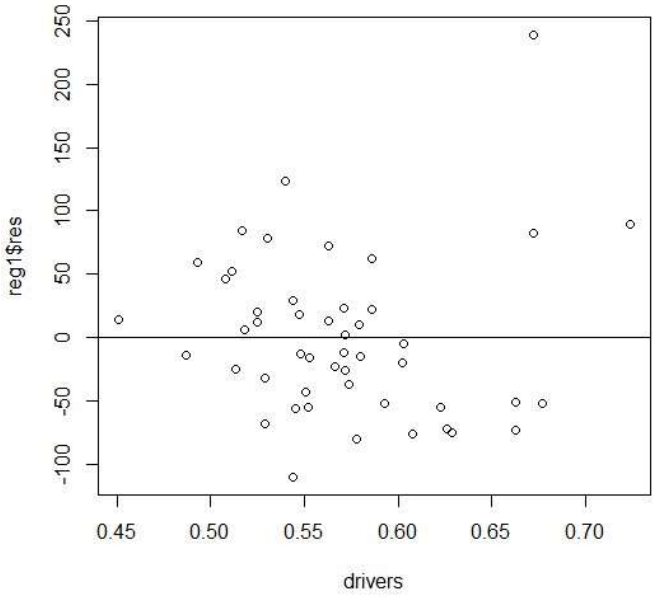
Na tym etapie decydujemy się na model pełny z wykluczeniem zmiennej *highways*.

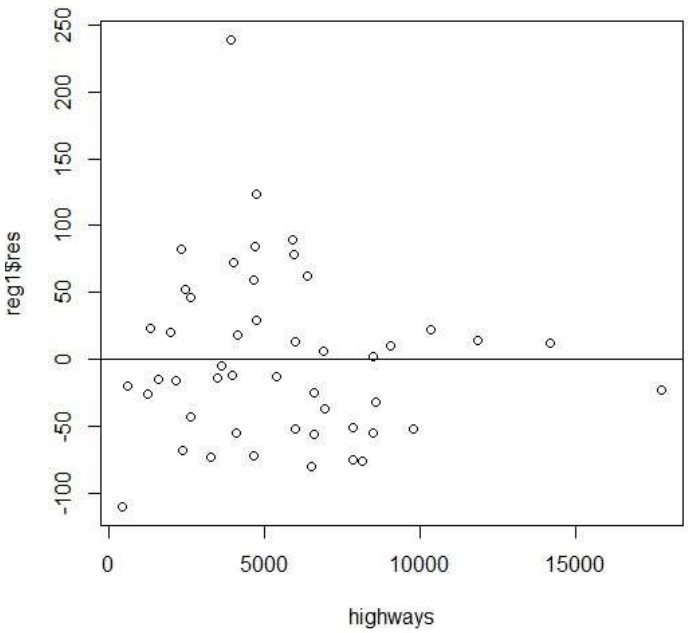
```
> reg1<-lm(consumption~.-highways,petrol)
> summary(reg1)

Call:
lm(formula = consumption ~ . - highways, data = petrol)

Residuals:
    Min      1Q  Median      3Q     Max
-110.10  -51.22  -12.89   24.49  238.77

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  307.32790  156.83067   1.960  0.05639 .
tax          -29.48381   10.58358  -2.786  0.00785 **
income        -0.06802    0.01701  -3.999  0.00024 ***
drivers     1374.76841  183.66954   7.485 2.24e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.94 on 44 degrees of freedom
Multiple R-squared:  0.6749,    Adjusted R-squared:  0.6527
F-statistic: 30.44 on 3 and 44 DF,  p-value: 8.235e-11
```

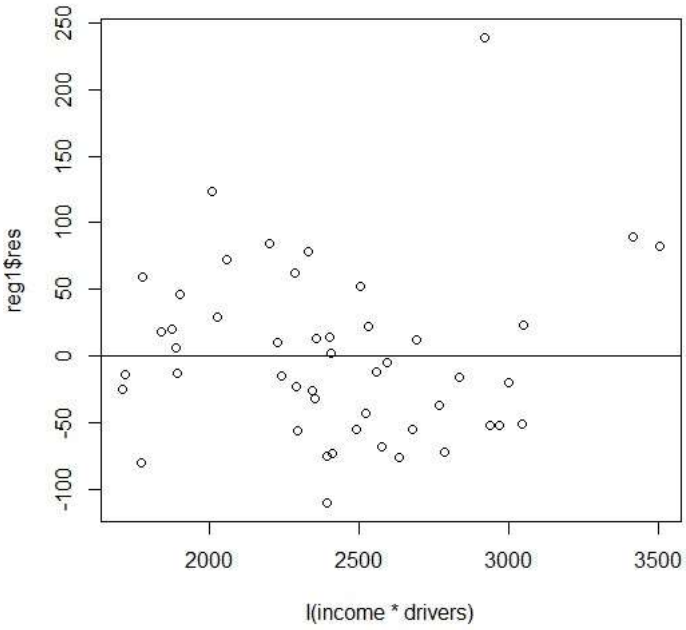
Zwiększyła się istotność zmiennej *tax*, a R^2 nieco spadło.

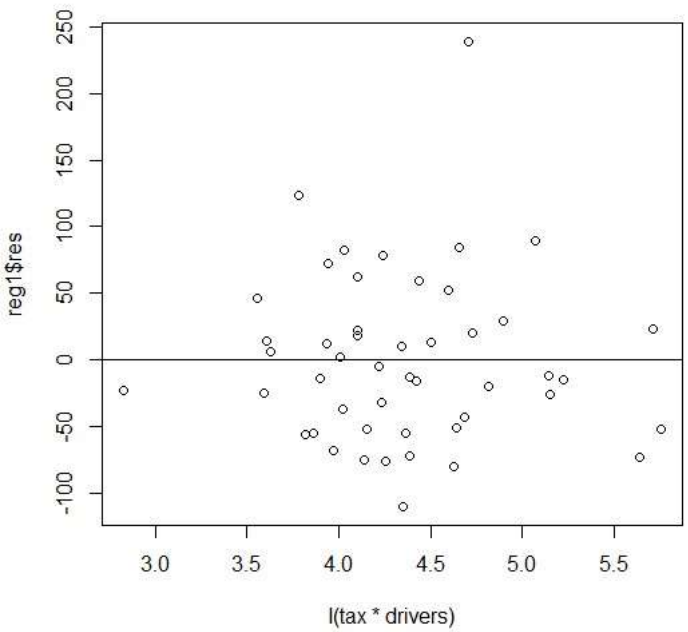


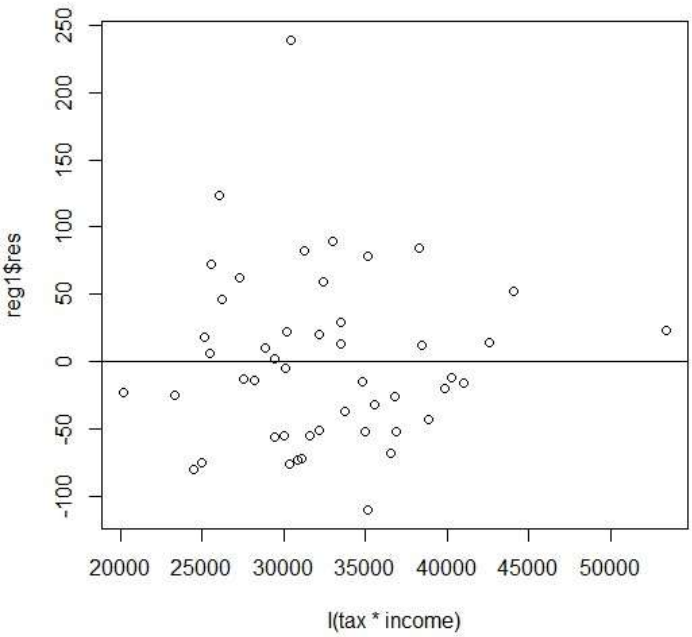


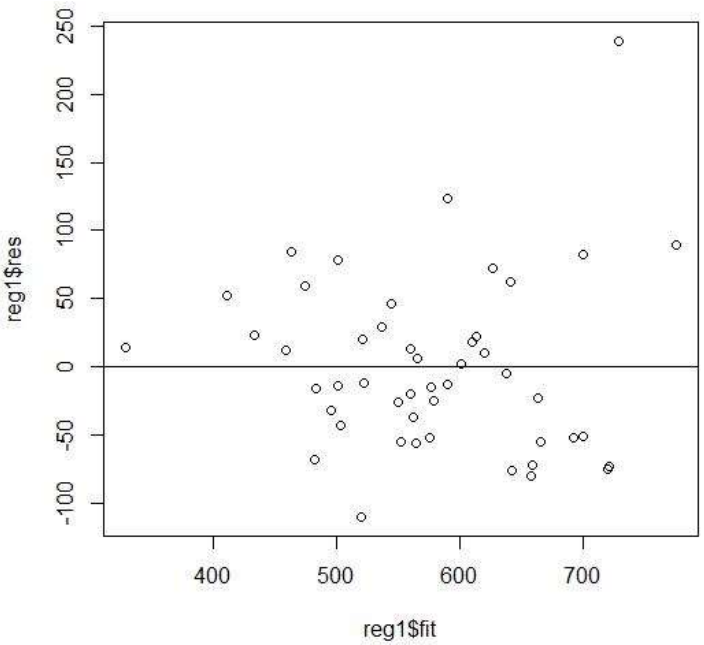

reg1$res
drivers
0.45
0.50
0.55
0.60
0.65
0.70
-100
-50
0
50
100
150
200
250


reg1$res
highways
0
5000
10000
15000
-100
-50
0
50
100
150
200
250

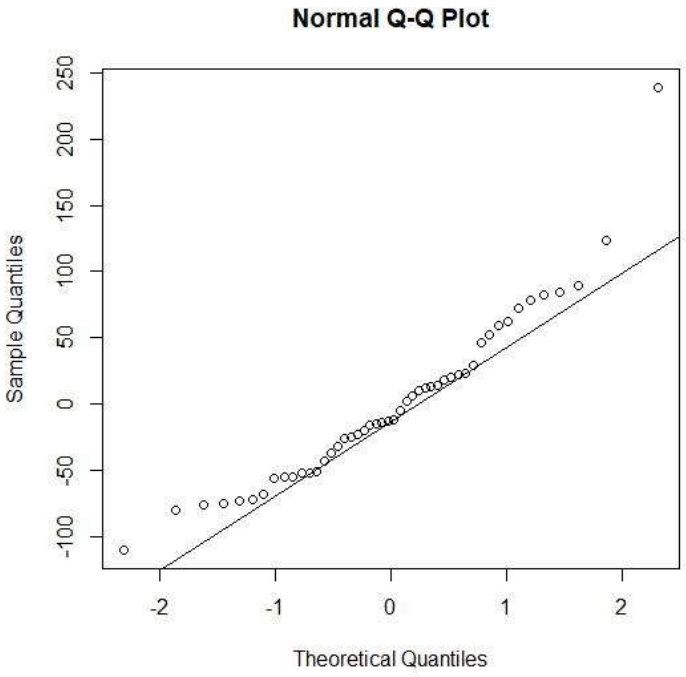

reg1$res
l(income * drivers)
2000
2500
3000
3500
-100
-50
0
50
100
150
200
250

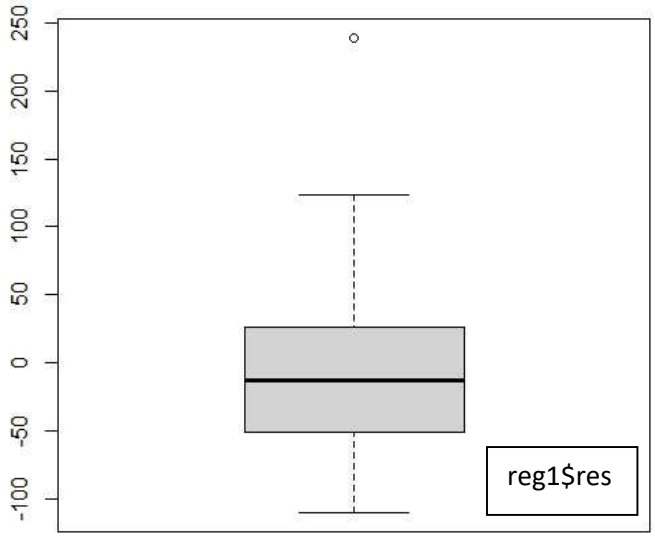

reg1$res
l(tax * drivers)
3.0
3.5
4.0
4.5
5.0
5.5
-100
-50
0
50
100
150
200
250


reg1$res
l(tax * income)
20000
25000
30000
35000
40000
45000
50000
-100
-50
0
50
100
150
200
250


reg1$res
reg1$fit
400
500
600
700
-100
-50
0
50
100
150
200
250

Wykresy reszt od zmiennych niezależnych, interakcji oraz wartości dopasowanych nie uległy widocznym zmianom.






```
> shapiro.test(reg1$res)

        Shapiro-Wilk normality test

data:  reg1$res
W = 0.9282, p-value = 0.005858
```

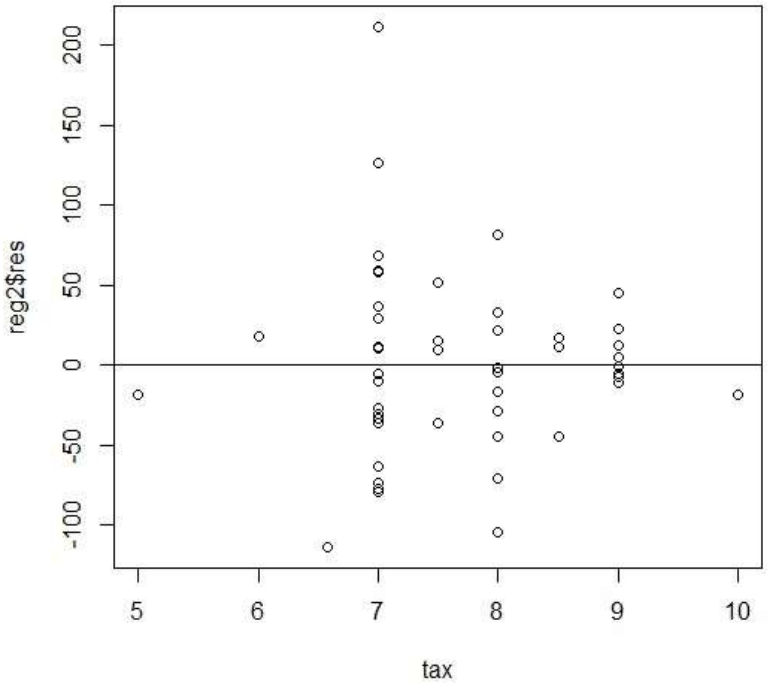
Tak jak poprzednio, odrzucamy hipotezę o normalności reszt.

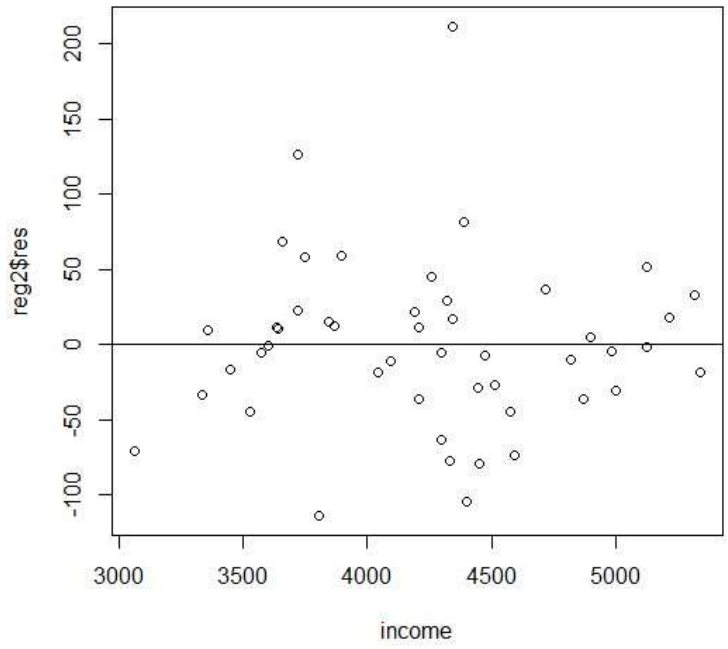
3. Wybór modelu

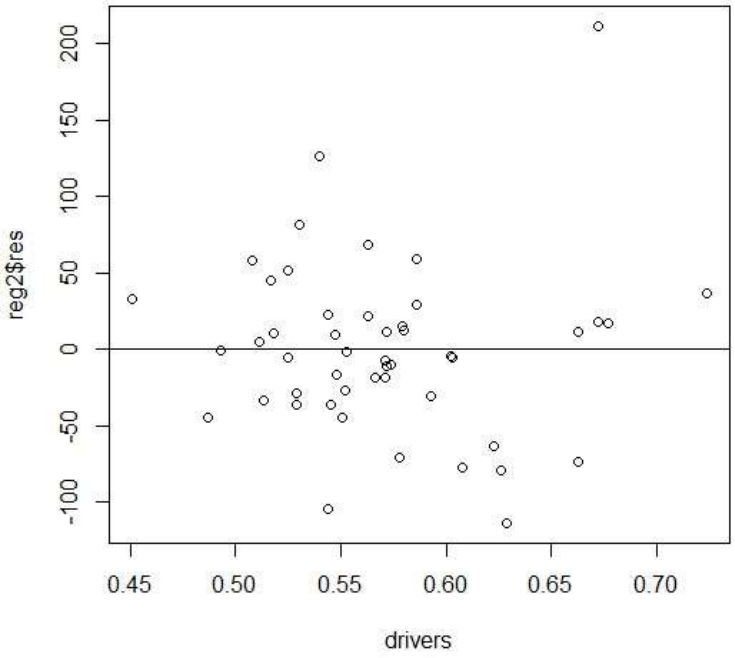
Do modelu bez zmiennej *highways* dokładamy interakcje podwójne zmiennych niezależnych występujących w tym modelu.

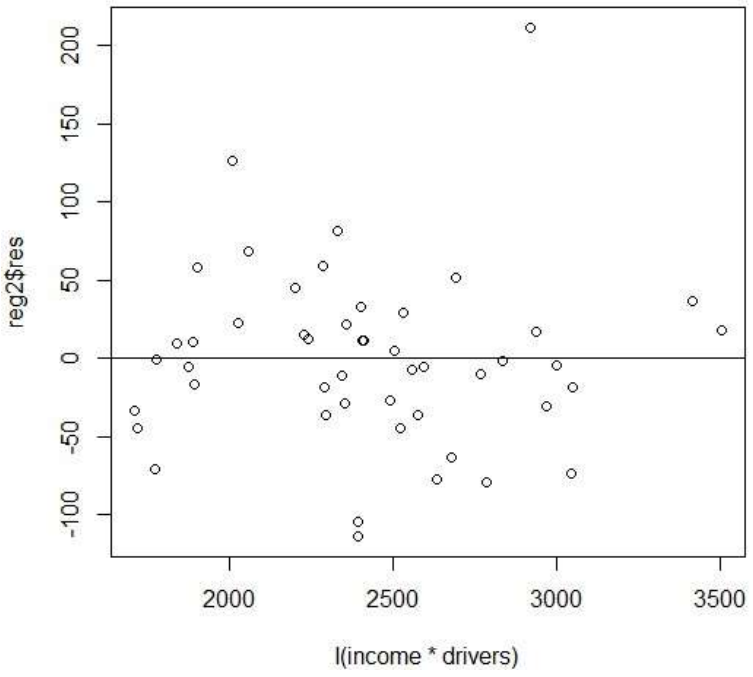
```
> reg2<-lm(consumption~tax*income*drivers-tax:income:drivers,petrol)
> summary(reg2)

Call:
lm(formula = consumption ~ tax * income * drivers - tax:income:drivers,
    data = petrol)

Residuals:
     Min       1Q   Median       3Q      Max
-113.284  -31.274   -2.641   22.195  211.434

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -978.02977 1455.89099  -0.672  0.50550
tax             231.83143  130.28628   1.779  0.08259 .
income           -0.43056    0.23117  -1.863  0.06970 .
drivers        5160.21514 2636.22431   1.957  0.05713 .
tax:income        0.02566    0.01782   1.440  0.15744
tax:drivers    -647.71335  238.76552  -2.713  0.00971 **
income:drivers    0.26756    0.31394   0.852  0.39903
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.75 on 41 degrees of freedom
Multiple R-squared:  0.7428,    Adjusted R-squared:  0.7052
F-statistic: 19.74 on 6 and 41 DF,  p-value: 1.118e-10
```

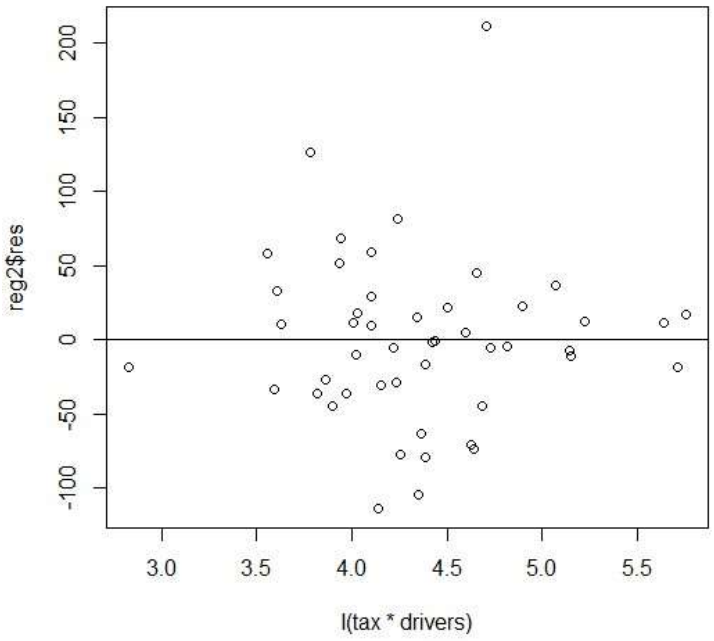
Zauważamy znaczący wzrost R^2, natomiast wiele zmiennych straciło na istotności. Najbardziej istotna jest zmienna *tax*drivers*.

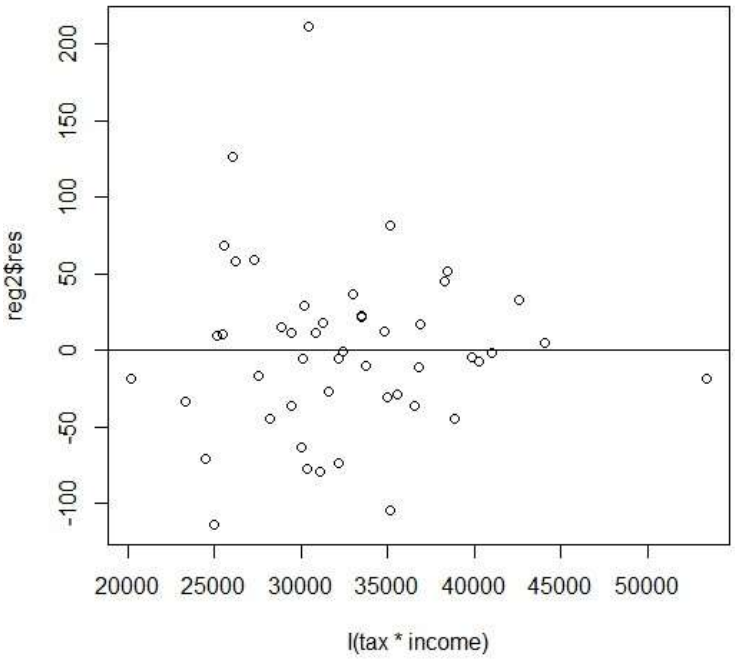












Wykresy reszt od niektórych zmiennych wydają się mieć podejrzanie malejącą wariancję, natomiast nie możemy na razie stwierdzić czy jest to tylko pozorne czy wymaga transformacji. Po wypróbowaniu transformacji logarytmicznej na zmiennej zależnej nie zauważamy zmian we wcześniej wspomnianych wariancjach, więc przeanalizujemy funkcję step.

```
> step(reg2)
Start:  AIC=400.68
consumption ~ tax * income * drivers - tax:income:drivers

                 Df Sum of Sq    RSS    AIC
- income:drivers  1    2680.4 153986 399.52
<none>                       151306 400.68
- tax:income      1    7653.0 158959 401.05
- tax:drivers     1   27157.8 178464 406.61

Step:  AIC=399.52
consumption ~ tax + income + drivers + tax:income + tax:drivers

              Df Sum of Sq    RSS    AIC
<none>                     153986 399.52
- tax:income   1      7072 161058 399.68
- tax:drivers  1     37275 191261 407.93

Call:
lm(formula = consumption ~ tax + income + drivers + tax:income +
    tax:drivers, data = petrol)

Coefficients:
(Intercept)          tax       income      drivers   tax:income  tax:drivers
 -1.947e+03    2.739e+02   -2.755e-01    6.854e+03    2.461e-02   -7.155e+02
```

Przed wurzuceniem zmiennej *income*drivers* z modelu zgodnie z sugestią kryterium Akaike, wypróbowaliśmy model ze wszystkimi interakcjami włącznie z potrójną.

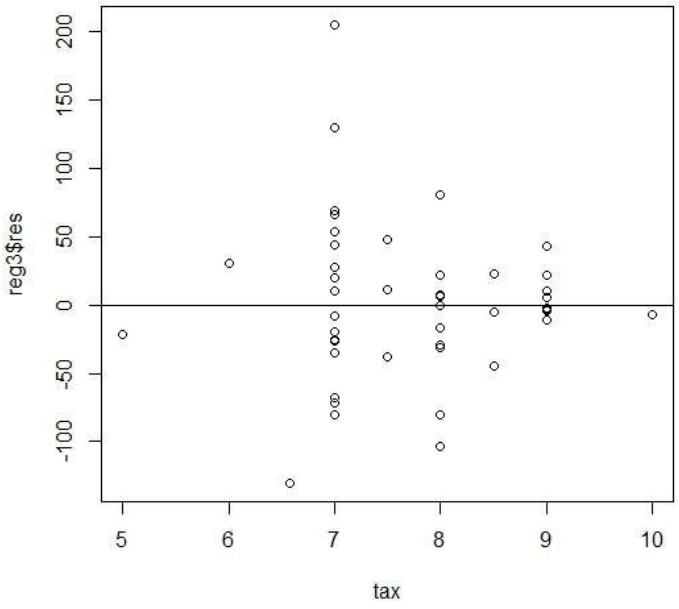
```
> reg4<-lm(consumption~tax*income*drivers,petrol)
> summary(reg4)

Call:
lm(formula = consumption ~ tax * income * drivers, data = petrol)

Residuals:
     Min       1Q   Median       3Q      Max
-103.607  -29.382   -1.881   21.309  217.791

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5213.4564  6564.6111   0.794    0.432
tax                -578.1528   847.4673  -0.682    0.499
income               -1.8660     1.5019  -1.242    0.221
drivers           -5610.1348 11442.8848  -0.490    0.627
tax:income            0.2139     0.1954   1.095    0.280
tax:drivers         766.5270  1481.4657   0.517    0.608
income:drivers        2.7555     2.5912   1.063    0.294
tax:income:drivers   -0.3275     0.3386  -0.967    0.339

Residual standard error: 60.8 on 40 degrees of freedom
Multiple R-squared:  0.7487,    Adjusted R-squared:  0.7047
F-statistic: 17.03 on 7 and 40 DF,  p-value: 3.373e-10
```

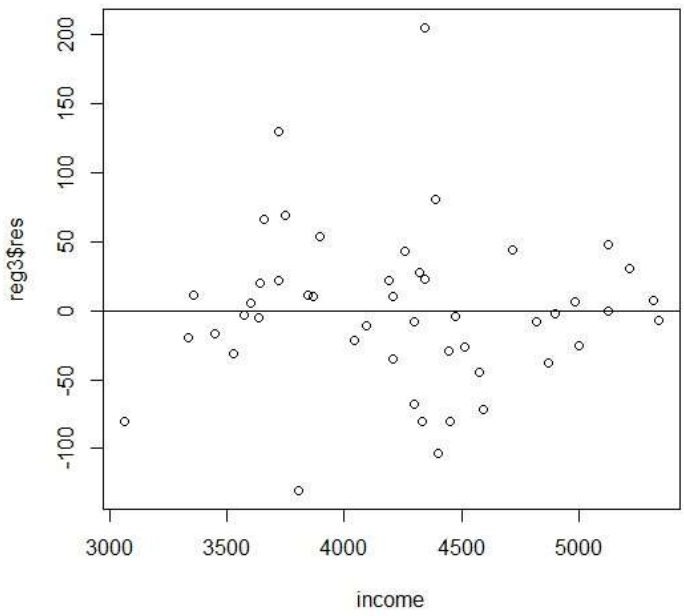
```
> step(reg4)
Start:  AIC=401.57
consumption ~ tax * income * drivers

                     Df Sum of Sq    RSS    AIC
- tax:income:drivers  1    3458.3 151306 400.68
<none>                             147847 401.57

Step:  AIC=400.68
consumption ~ tax + income + drivers + tax:income + tax:drivers +
    income:drivers

                 Df Sum of Sq    RSS    AIC
- income:drivers  1    2680.4 153986 399.52
<none>                        151306 400.68
- tax:income      1    7653.0 158959 401.05
- tax:drivers     1   27157.8 178464 406.61

Step:  AIC=399.52
consumption ~ tax + income + drivers + tax:income + tax:drivers

             Df Sum of Sq    RSS    AIC
<none>                    153986 399.52
- tax:income  1      7072 161058 399.68
- tax:drivers 1     37275 191261 407.93

Call:
lm(formula = consumption ~ tax + income + drivers + tax:income +
    tax:drivers, data = petrol)

Coefficients:
(Intercept)          tax       income      drivers   tax:income  tax:drivers
 -1.947e+03    2.739e+02   -2.755e-01    6.854e+03    2.461e-02   -7.155e+02
```

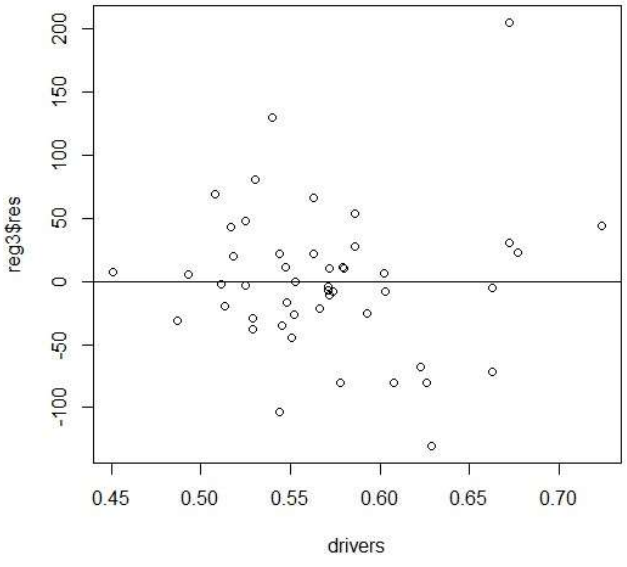
Funkcja step zastosowana na tym modelu zaproponowała nam ten sam model co poprzednio wykonana funkcja step.

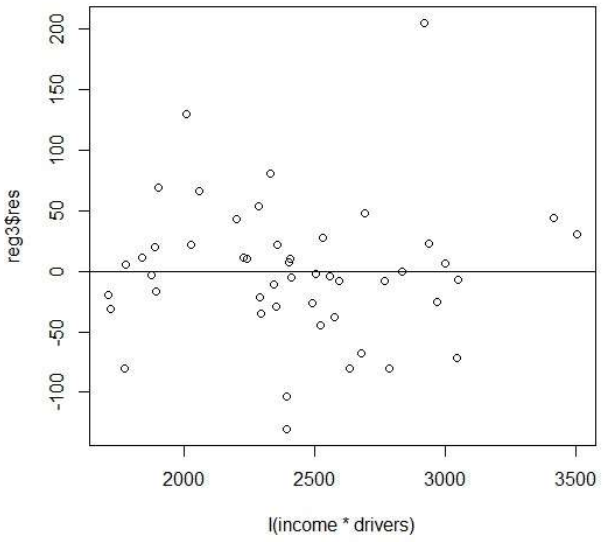
Przejdźmy zatem do analizy poniższego modelu.

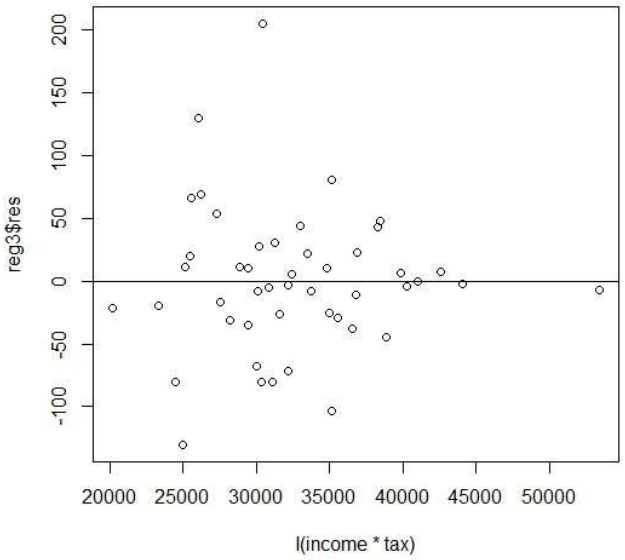
```
> reg3<-lm(consumption~tax+income+drivers+tax:income+tax:drivers,petrol)
> summary(reg3)

Call:
lm(formula = consumption ~ tax + income + drivers + tax:income +
    tax:drivers, data = petrol)

Residuals:
     Min       1Q   Median       3Q      Max
-130.204  -26.885   -2.576   22.547  205.156

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.947e+03  9.071e+02  -2.146 0.037699 *
tax          2.739e+02  1.202e+02   2.280 0.027758 *
income      -2.755e-01  1.422e-01  -1.938 0.059370 .
drivers      6.854e+03  1.726e+03   3.971 0.000275 ***
tax:income   2.461e-02  1.772e-02   1.389 0.172198
tax:drivers -7.155e+02  2.244e+02  -3.189 0.002702 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.55 on 42 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7071
F-statistic:  23.7 on 5 and 42 DF,  p-value: 3.052e-11
```

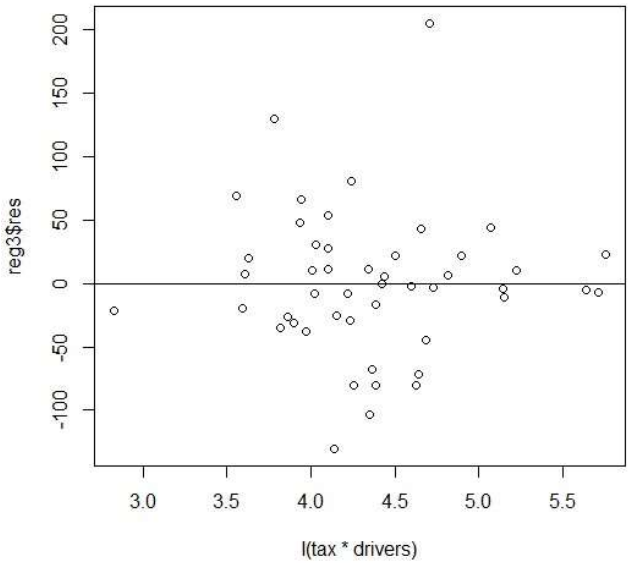






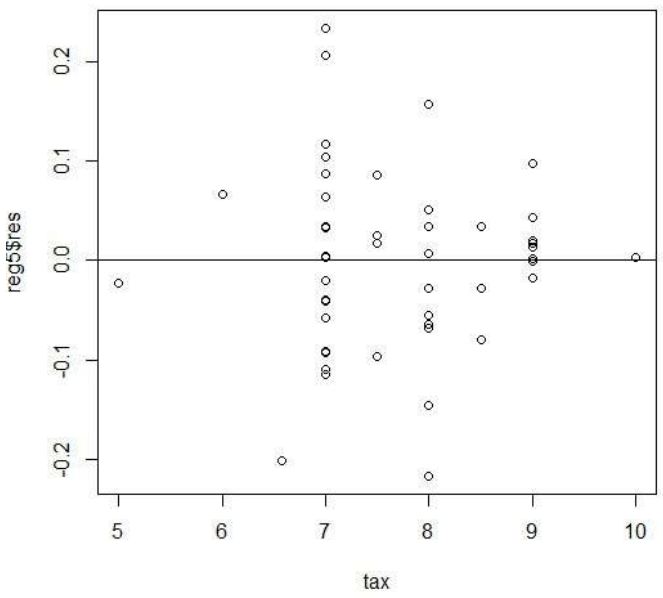


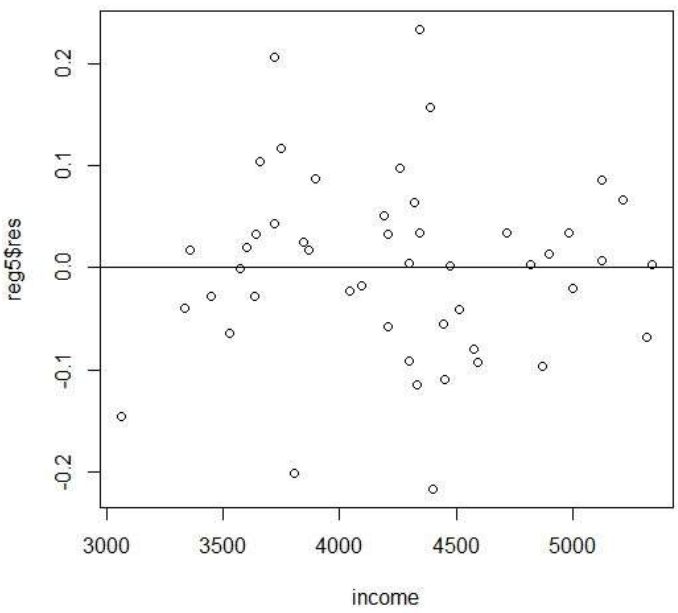


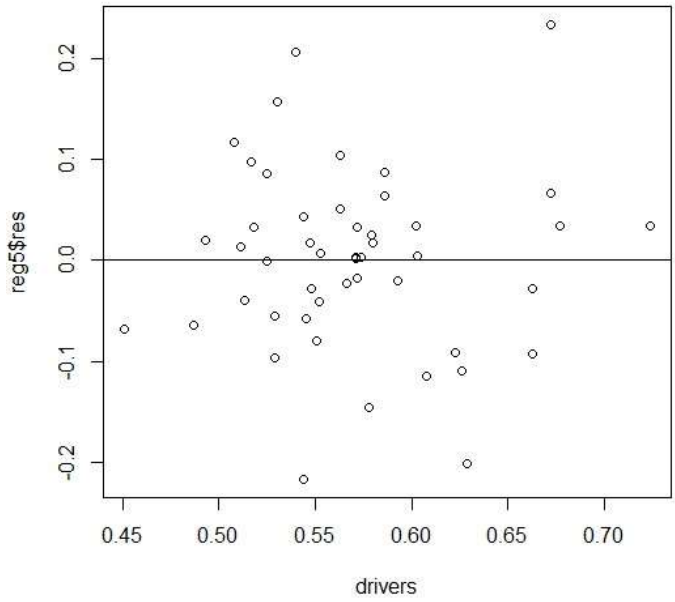


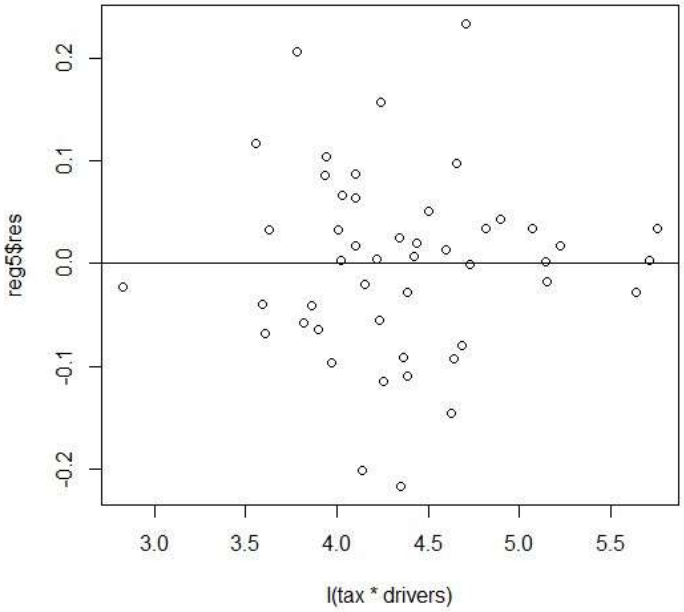
Niektóre wykresy reszt nadal budzą niepokój co do stałości wariancji więc ponownie decydujemy się na transformację logarytmiczną zmiennej zależnej.

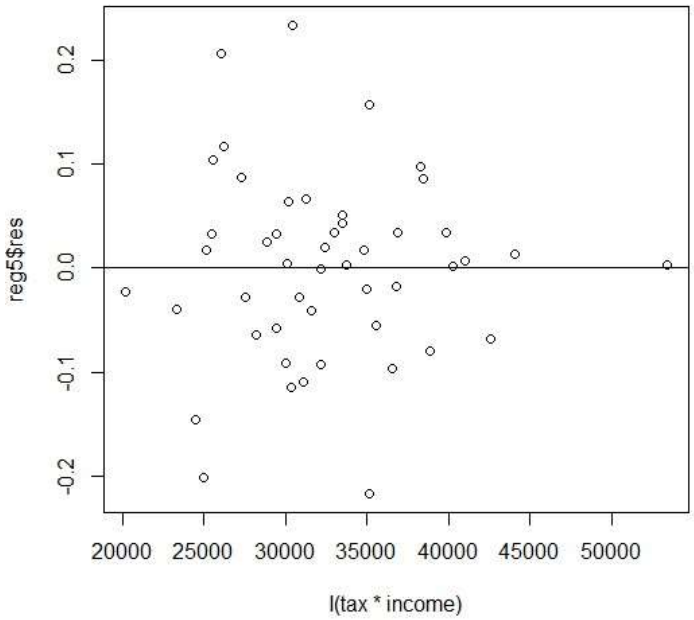
```
> summary(reg5)

Call:
lm(formula = log(consumption) ~ tax + drivers + income + tax:drivers +
    tax:income, data = petrol)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21628 -0.05609  0.00413  0.03688  0.23285

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.799e+00  1.421e+00   1.969 0.055594 .
tax          3.788e-01  1.883e-01   2.012 0.050705 .
drivers      1.013e+01  2.705e+00   3.744 0.000545 ***
income      -4.417e-04  2.228e-04  -1.982 0.053996 .
tax:drivers -1.019e+00  3.516e-01  -2.897 0.005964 **
tax:income   3.613e-05  2.777e-05   1.301 0.200281
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09489 on 42 degrees of freedom
Multiple R-squared:  0.7696,    Adjusted R-squared:  0.7422
F-statistic: 28.06 on 5 and 42 DF,  p-value: 2.225e-12
```


reg5$res
0.2
0.1
0.0
-0.1
-0.2
5
6
7
8
9
10
tax


reg5$res
0.2
0.1
0.0
-0.1
-0.2
3000
3500
4000
4500
5000
income


reg5$res
0.2
0.1
0.0
-0.1
-0.2
0.45
0.50
0.55
0.60
0.65
0.70
drivers


reg5$res
0.2
0.1
0.0
-0.1
-0.2
3.0
3.5
4.0
4.5
5.0
5.5
I(tax * drivers)


reg5$res
0.2
0.1
0.0
-0.1
-0.2
20000
25000
30000
35000
40000
45000
50000
I(tax * income)

W modelu wzrosło R^2 oraz zauważamy widoczną poprawę na wykresach reszt. Nieistotność interakcji *tax*income* skłoniła nas jednak do analizy modelu bez niej. Zasugerowała to również funkcja step.

```
> step(reg5)
Start:  AIC=-220.5
log(consumption) ~ tax + income + drivers + tax:income + tax:drivers

              Df Sum of Sq     RSS     AIC
- tax:income   1  0.015244 0.39338 -220.60
<none>                     0.37814 -220.50
- tax:drivers  1  0.075549 0.45369 -213.75

Step:  AIC=-220.6
log(consumption) ~ tax + income + drivers + tax:drivers

              Df Sum of Sq     RSS     AIC
<none>                     0.39338 -220.60
- tax:drivers  1   0.06048 0.45387 -215.74
- income       1   0.33350 0.72689 -193.13

Call:
lm(formula = log(consumption) ~ tax + income + drivers + tax:drivers,
    data = petrol)

Coefficients:
(Intercept)          tax       income      drivers  tax:drivers
  2.4153321    0.4279640   -0.0001536    8.6384755   -0.8333148
```

Spójrzmy zatem na podsumowanie modelu bez tej interakcji.

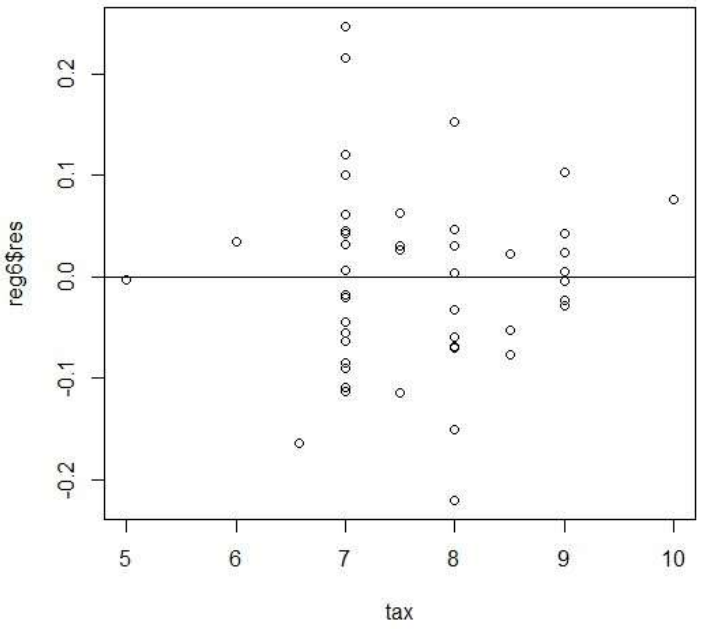
```
> reg6<-lm(log(consumption)~tax+income+drivers+tax:drivers,petrol)
> summary(reg6)

Call:
lm(formula = log(consumption) ~ tax + income + drivers + tax:drivers,
    data = petrol)

Residuals:
      Min        1Q    Median        3Q       Max
-0.219766 -0.060090  0.004644  0.043857  0.246578

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.415e+00  1.402e+00   1.723  0.09206 .
tax          4.280e-01  1.859e-01   2.302  0.02627 *
income      -1.536e-04  2.545e-05  -6.038 3.22e-07 ***
drivers      8.638e+00  2.471e+00   3.496  0.00111 **
tax:drivers -8.333e-01  3.241e-01  -2.571  0.01368 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09565 on 43 degrees of freedom
Multiple R-squared:  0.7603,    Adjusted R-squared:  0.738
F-statistic:  34.1 on 4 and 43 DF,  p-value: 7.967e-13
```

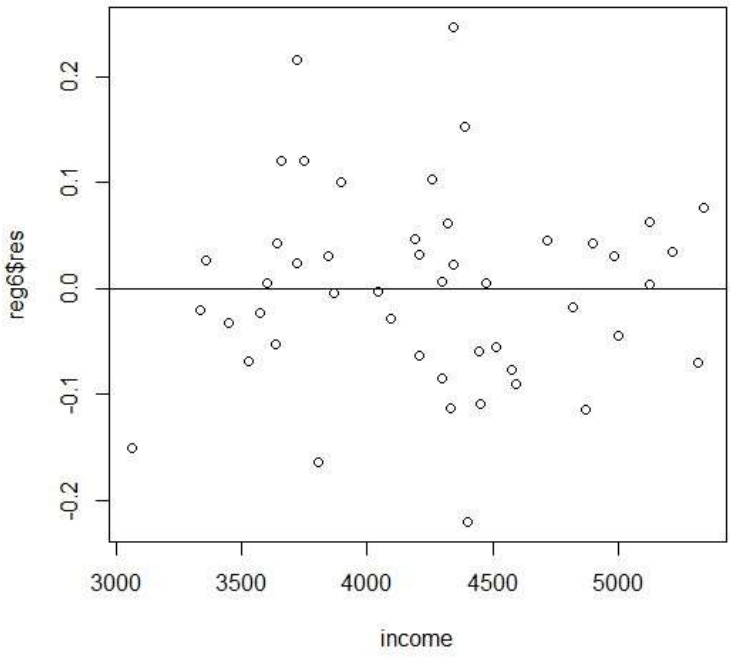
Poprawiła się istotność zmiennych, a R^2 jest na podobnym poziomie co poprzednio.

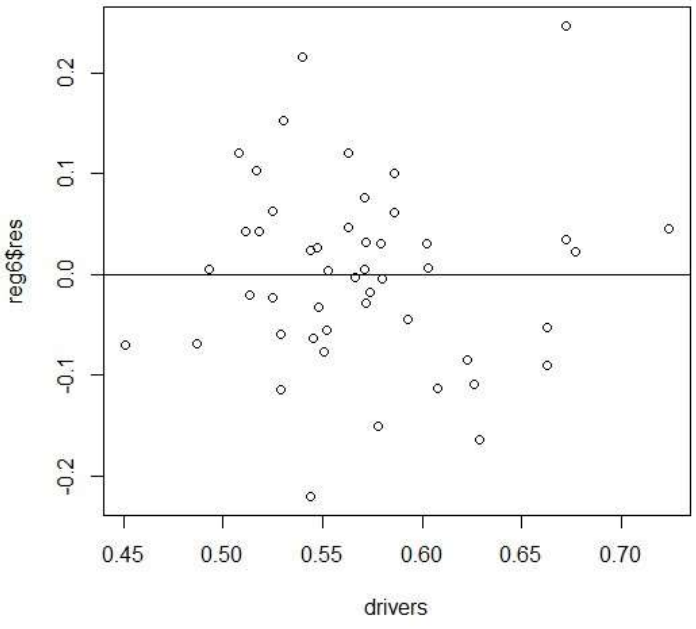
```
> anova(reg6,reg5)
Analysis of Variance Table

Model 1: log(consumption) ~ tax + drivers + income + tax:drivers
Model 2: log(consumption) ~ tax + drivers + income + tax:drivers + tax:income
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     43 0.39338
2     42 0.37814  1  0.015244 1.6931 0.2003
```

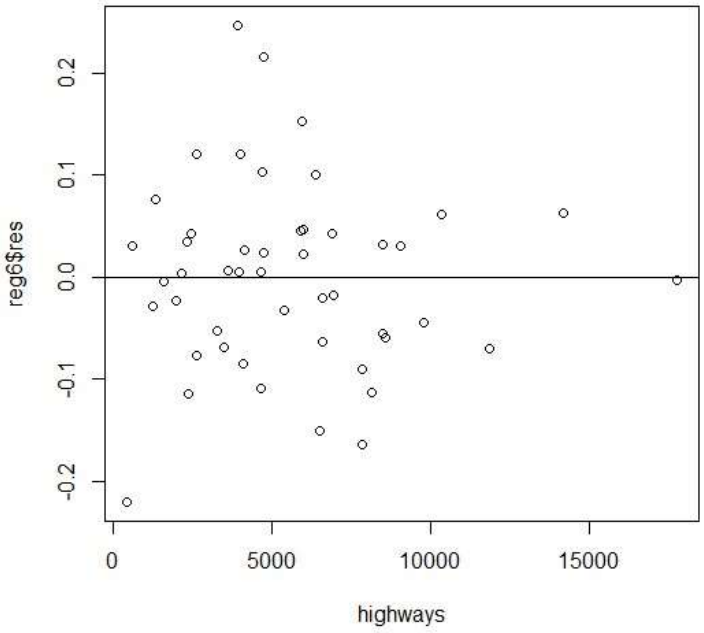
Do podjęcia ostatecznej decyzji sprawdziliśmy analizę wariancji dla obydwu modeli. Wartość p > 0.05 zasugerowała, że modele te nie różnią się istotnie, więc wybieramy model z mniejszą ilością zmiennych.

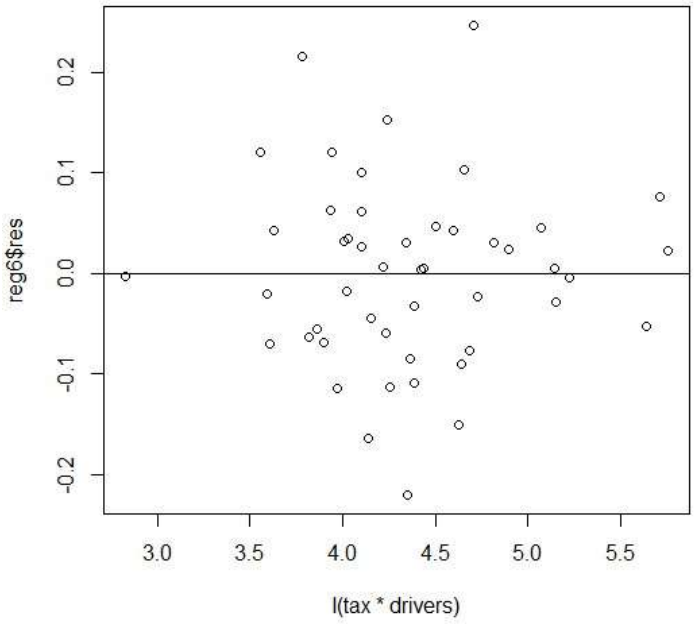
4. Diagnostyka wybranego modelu

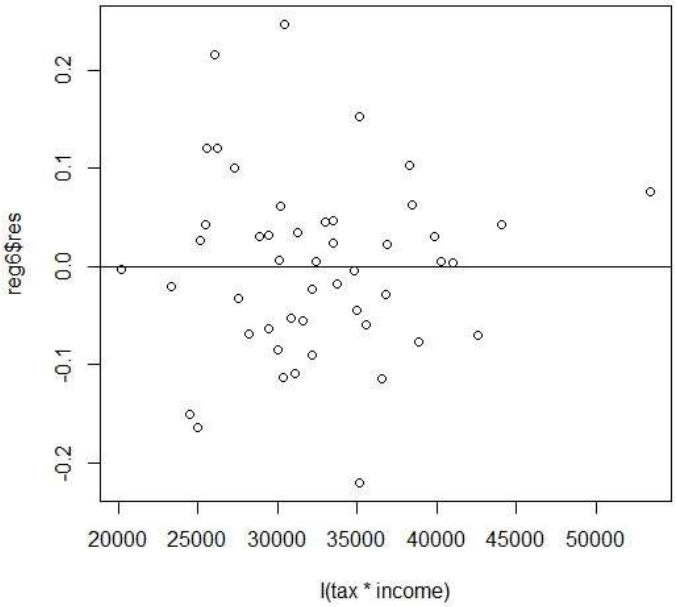


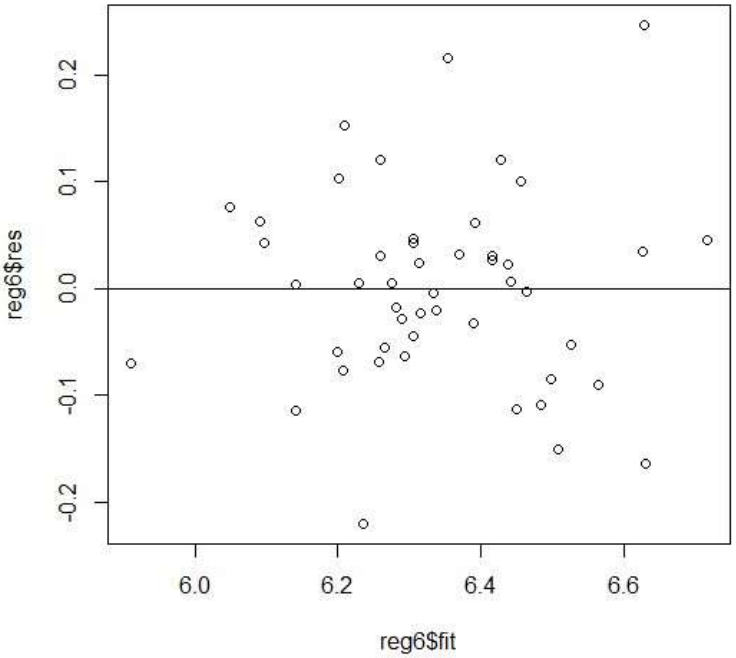




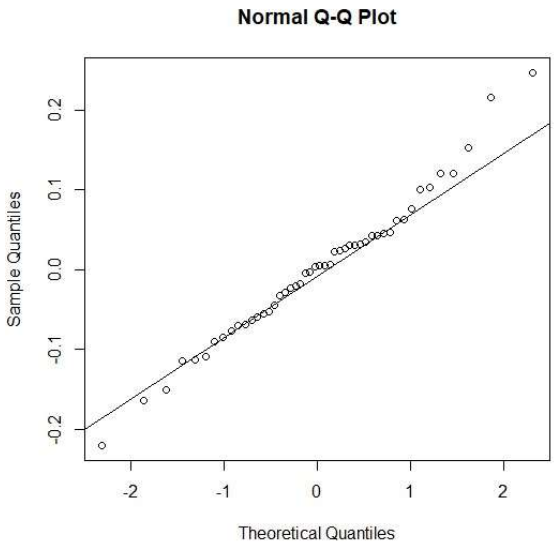


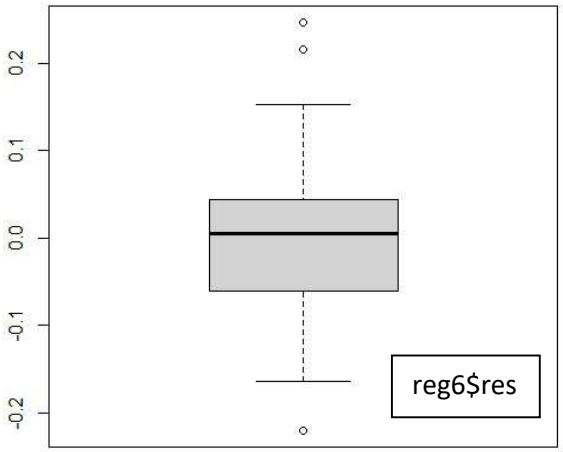






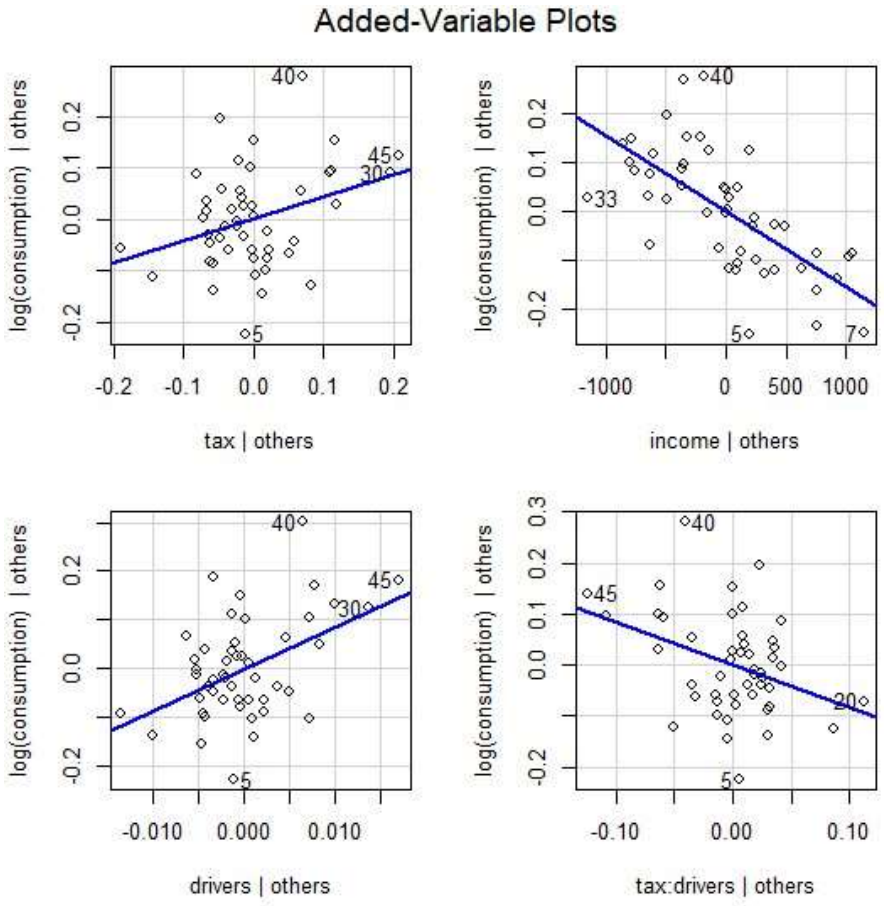
Wykresy reszt nie wykazują żadnych niepokojących zależności, a wariancje wydają się być stałe.






```
> shapiro.test(reg6$res)

        Shapiro-Wilk normality test

data:  reg6$res
W = 0.98429, p-value = 0.7618
```

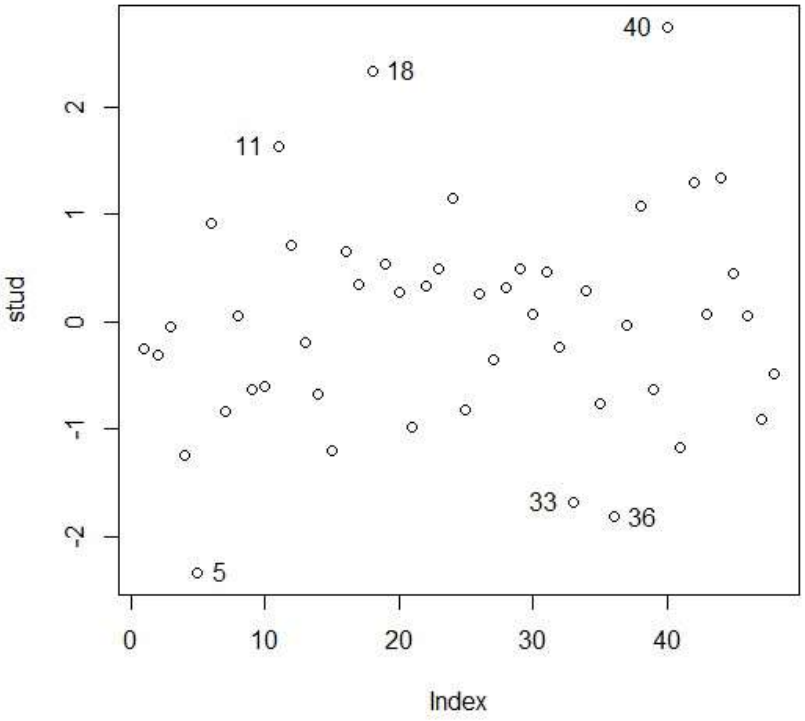
Wysoka wartość p w teście Shapiro-Wilka potwierdza hipotezę o normalności reszt.



Wykresy regresji częściowej nie wykazują zależności liniowej ani też żadnej innej.

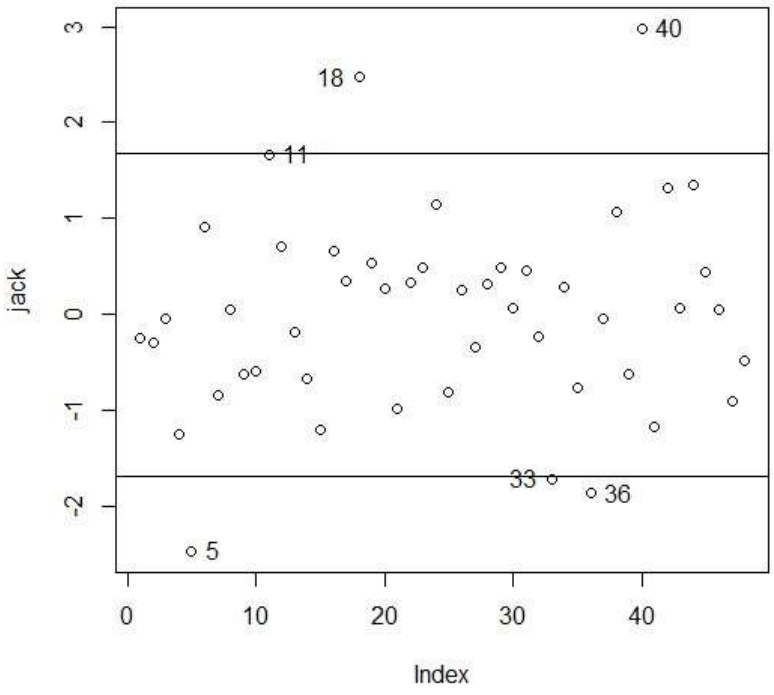
Przejdźmy zatem do analizy obserwacji odstających i wpływowych w wybranym modelu.

```
> stud<-rstandard(reg6)
> plot(stud)
> identify(stud)
[1]  5 11 18 33 36 40
```



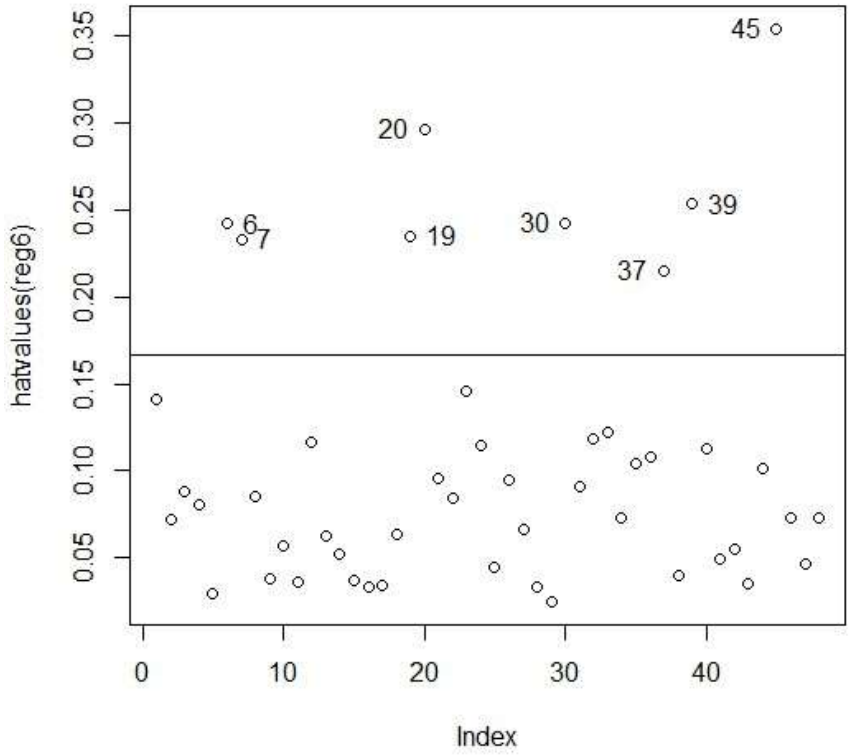
Wykres reszt standaryzowanych pokazuje powyższe nieco odstające obserwacje ze względu na zmienną zależną.

```
> jack<-rstudent(reg6)
> dim(petrol)
[1] 48  5
> qt(0.95,43)
[1] 1.681071
> plot(jack)
> abline(h=1.681071)
> abline(h=-1.681071)
> identify(jack)
[1]  5 11 18 33 36 40
```



Wykres reszt studentyzowanych proponuje dokładnie te same obserwacje jako odstające.

```
> plot(hatvalues(reg6))
> abline(h=8/48)
> identify(hatvalues(reg6))
```



Wykres wartości dźwigni dla kolejnych obserwacji wskazuje obserwacje odstające ze względu na zmienne niezależne nieco odmienne niż ze względu na zmienną zależną.

```
> inf<-influence.measures(reg6)
> summary(inf)
Potentially influential observations of
        lm(formula = log(consumption) ~ tax + income + drivers + tax:drivers,      data = p$

   dfb.1_ dfb.tax dfb.incm dfb.drvr dfb.tx:d dffit   cov.r   cook.d hat
5  -0.06   0.05   -0.13     0.07    -0.05    -0.43    0.59_*  0.03   0.03
7   0.03  -0.04   -0.29    -0.01     0.05    -0.46    1.35_*  0.04   0.23
18  0.28  -0.23   -0.23    -0.23     0.20     0.64    0.61_*  0.07   0.06
19 -0.16   0.13    0.00     0.16    -0.13     0.30    1.42_*  0.02   0.23
20  0.11  -0.12    0.02    -0.11     0.13     0.18    1.58_*  0.01   0.30
30 -0.03   0.03   -0.02     0.03    -0.03     0.04    1.48_*  0.00   0.24
37 -0.01   0.01    0.00     0.00     0.00    -0.02    1.43_*  0.00   0.21
39 -0.18   0.20    0.09     0.19    -0.21    -0.36    1.44_*  0.03   0.25
40 -0.51   0.43   -0.15     0.52    -0.44     1.06_*  0.49_*  0.19   0.11
45 -0.24   0.23    0.06     0.24    -0.24     0.33    1.70_*  0.02   0.35_*
> pf(0.19,4,44)
[1] 0.05763109
> pf(0.02,4,44)
[1] 0.0008124426
```

Dla zdecydowanej większości podejrzanych obserwacji powyższe podsumowanie nie daje podstaw do dalszego rozważania ich jako wpływowe.

Dla obserwacji 45 mamy wysoką wartość dźwigni, co widzieliśmy na wcześniejszym wykresie wartości dźwigni. Jednakże, ani wartości DFBETAS nie przekraczają progu 2/sqrt(48), ani wartość DFFITS nie jest niepokojąca. Decydujemy się więc nie wyrzucać jej z modelu, zwłaszcza, że nie mamy wielu obserwacji.

Jedyną obserwacją dla której wartość DFFITS na moduł przekracza wartość 1 jest obserwacja 40. Dla niej też odległość Cooka jest największa spośród wszystkich obserwacji potencjalnie wpływowych, jednak po sprawdzeniu kwantyla rozkładu F dla tejże wartości nie stwierdzamy, żeby ta obserwacja była wpływowa. Mimo to zbadaliśmy różnicę pomiędzy modelem z nią oraz bez niej.

```
> reg7<-lm(log(consumption)~tax+income+drivers+tax:drivers,petrol,subset=(row.names(petrol)!="40"))
> summary(reg6)

Call:
lm(formula = log(consumption) ~ tax + income + drivers + tax:drivers,
    data = petrol)

Residuals:
      Min        1Q    Median        3Q       Max
-0.219766 -0.060090  0.004644  0.043857  0.246578

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.415e+00  1.402e+00   1.723  0.09206 .
tax          4.280e-01  1.859e-01   2.302  0.02627 *
income      -1.536e-04  2.545e-05  -6.038 3.22e-07 ***
drivers      8.638e+00  2.471e+00   3.496  0.00111 **
tax:drivers -8.333e-01  3.241e-01  -2.571  0.01368 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09565 on 43 degrees of freedom
Multiple R-squared:  0.7603,    Adjusted R-squared:  0.738
F-statistic:  34.1 on 4 and 43 DF,  p-value: 7.967e-13

> summary(reg7)

Call:
lm(formula = log(consumption) ~ tax + income + drivers + tax:drivers,
    data = petrol, subset = (row.names(petrol) != "40"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.22040 -0.06189  0.00237  0.04644  0.21441

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.066e+00  1.307e+00   2.346  0.02380 *
tax          3.540e-01  1.728e-01   2.049  0.04676 *
income      -1.501e-04  2.343e-05  -6.406 1.03e-07 ***
drivers      7.451e+00  2.307e+00   3.230  0.00241 **
tax:drivers -7.019e-01  3.013e-01  -2.330  0.02469 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08795 on 42 degrees of freedom
Multiple R-squared:  0.7592,    Adjusted R-squared:  0.7362
F-statistic:  33.1 on 4 and 42 DF,  p-value: 1.761e-12
```

Polepszyła się nieco istotność wyrazu wolnego, natomiast wartość R^2 nieco spadła. Wykresy reszt nie różniły się znacząco, więc ostatecznie pozostawiamy tę obserwację oraz wszystkie inne w modelu.

5. Podsumowanie

Powyższa analiza regresji skłania nas do spodziewanego wniosku, iż najbardziej istotnym czynnikiem wpływającym na konsumpcję paliwa w Stanach Zjednoczonych jest dochód na osobę. Podobnie niezaskakujący jest istotny wpływ odsetka osób posiadających prawo jazdy oraz wielkości podatku od benzyny. Nieco zaskakującym faktem może być usunięcie z modelu zmiennej opisującej łączną długość autostrad. Niemniej jednak postulujemy, iż nie jest to spowodowane tym, że czynnik ten nie ma wpływu na zużycie paliwa, a faktem, że zmienna ta może być wyjaśniana przez pozostałe zmienne w modelu.