Przygotowały:

Katarzyna Karpierz, nr indeksu: 276777

Klaudia Szczygieł, nr indeksu: 283557

Model Regresji

Opis danych

Dane wejściowe pochodzą ze strony http://rcarbonneau.com . Oryginalnie dane zostały zebrane w celu analizy trendu w lekkich samochodach oraz zużycia paliwa w modelach samochodowych z lat 1975-1991. Modele porównywano uwzględniając klasę wagową i wielkościową pojazdu, liczbę cylindrów oraz maksymalną prędkość. Zbiór zawiera 82 obserwacje dla każdej z 5 zmiennych.

W naszym projekcie uwzględnione zmienne przyjęły następujące nazwy:

fuel_cons – konsumpcja paliwa [mila/gallon]

cab_vol – pojemność kabiny [stopa sześcienna]

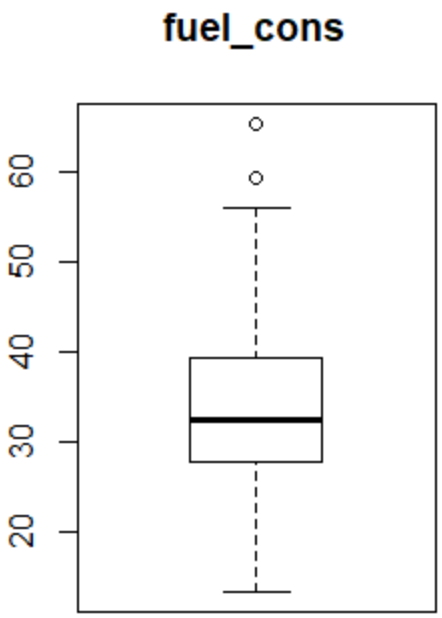
engine_pow – moc silnika

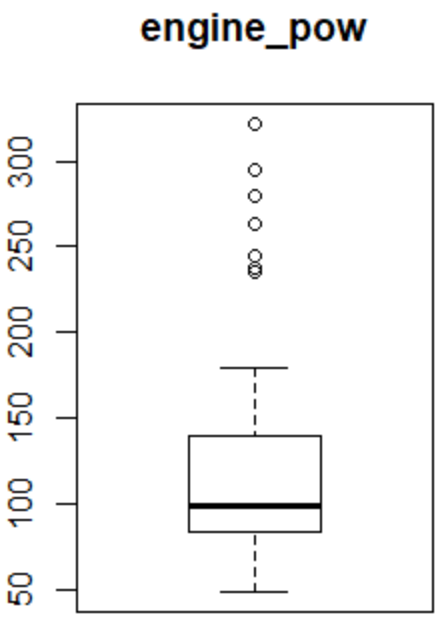
top_speed – maksymalna prędkość [mila/h]

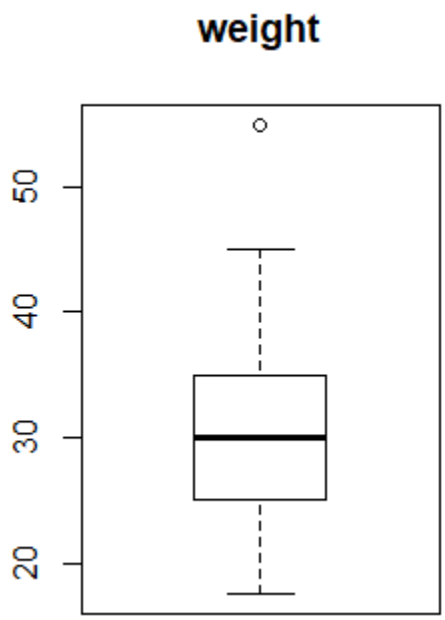
weight- waga [100 funtów]

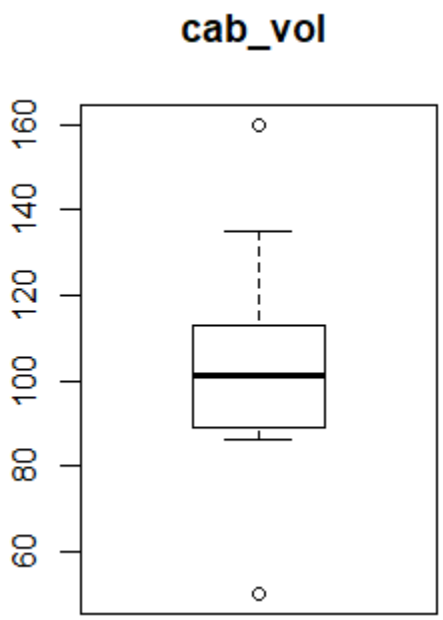
Przegląd danych

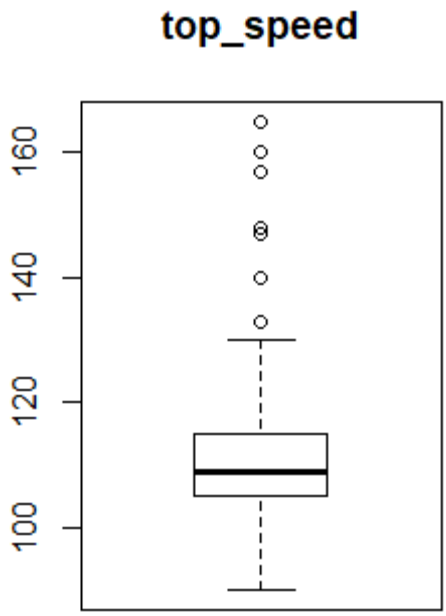
```
   fuel_cons        cab_vol        engine_pow      top_speed        weight
Min.   :13.20   Min.   : 50.0   Min.   : 49.0   Min.   : 90.0   Min.   :17.50
1st Qu.:27.77   1st Qu.: 89.5   1st Qu.: 84.0   1st Qu.:105.0   1st Qu.:25.00
Median :32.45   Median :101.0   Median : 99.0   Median :109.0   Median :30.00
Mean   :33.78   Mean   : 98.8   Mean   :117.1   Mean   :112.4   Mean   :30.91
3rd Qu.:39.30   3rd Qu.:113.0   3rd Qu.:140.0   3rd Qu.:114.8   3rd Qu.:35.00
Max.   :65.40   Max.   :160.0   Max.   :322.0   Max.   :165.0   Max.   :55.00
```


fuel_cons
20
30
40
50
60


engine_pow
50
100
150
200
250
300

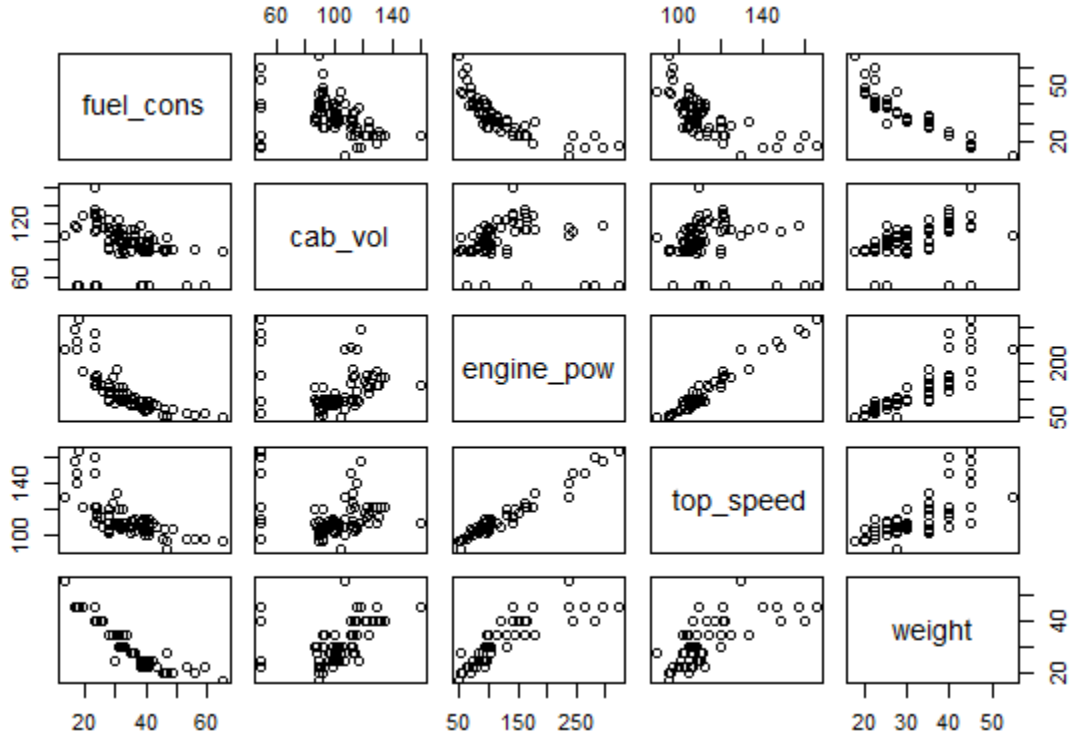

weight
20
30
40
50


cab_vol
60
80
100
120
140
160



Każda ze zmiennych zawiera przynajmniej jedną obserwację odstającą. Wnioskując po asymetrii wąsów możemy stwierdzić skośność rozkładów zmiennej ***cab_vol***.

Poniżej znajdują się wykresy zależności pomiędzy poszczególnymi zmiennymi. Przyglądając się wykresom możemy zauważyć, że prawdopodobnie istnieje liniowa zależność między parami zmiennych ***engine_pow*** i ***top_speed*** oraz ***fuel_cons*** i ***weight***.



Korelacje zmiennych.

	fuel_cons	cab_vol	engine_pow	top_speed	weight
fuel_cons	1.0000000	-0.36861368	-0.78985635	-0.68844623	-0.9050849
cab_vol	-0.3686137	1.00000000	0.07647905	-0.04306242	0.3849542
engine_pow	-0.7898564	0.07647905	1.00000000	0.96654517	0.8322202
top_speed	-0.6884462	-0.04306242	0.96654517	1.00000000	0.6785339
weight	-0.9050849	0.38495423	0.83222021	0.67853388	1.0000000

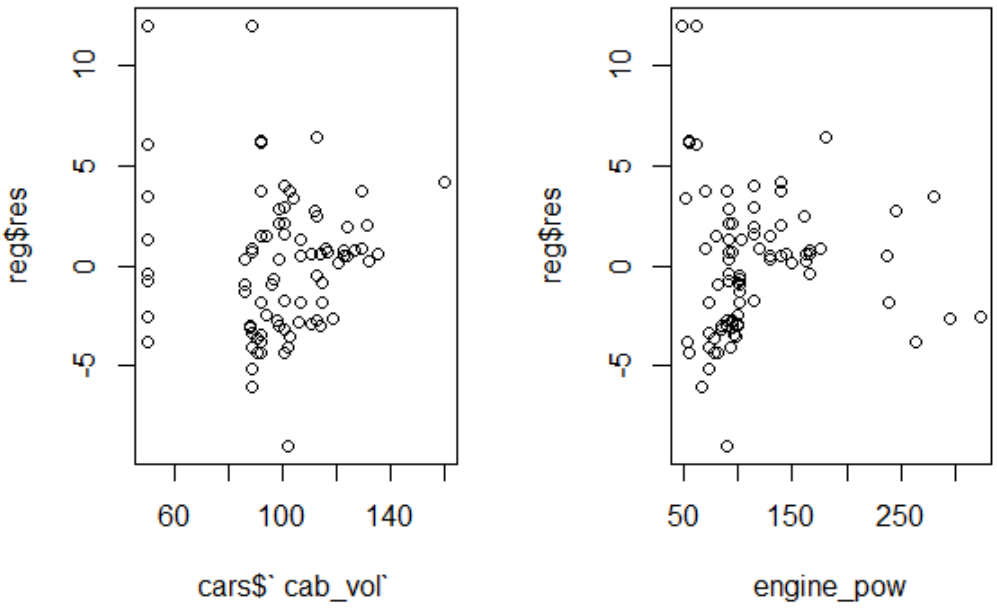
Większość zmiennych jest mocno ujemnie skorelowana ze zmienną ***fuel_cons***. Zatem przy ich pomocy możemy opisać parametr ***fuel_cons***, który będzie zmienną zależną.

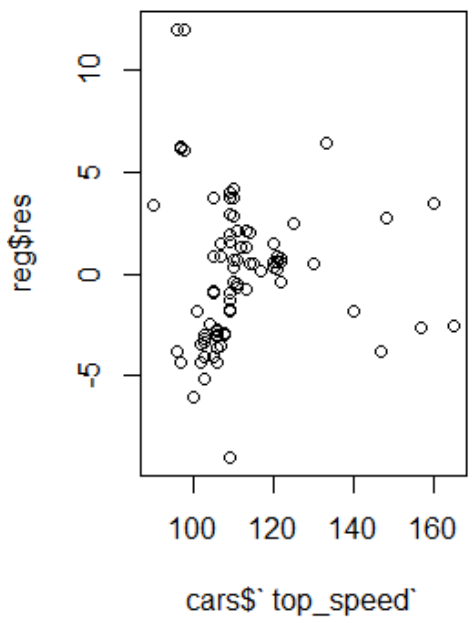
Dopasowanie modelu pełnego

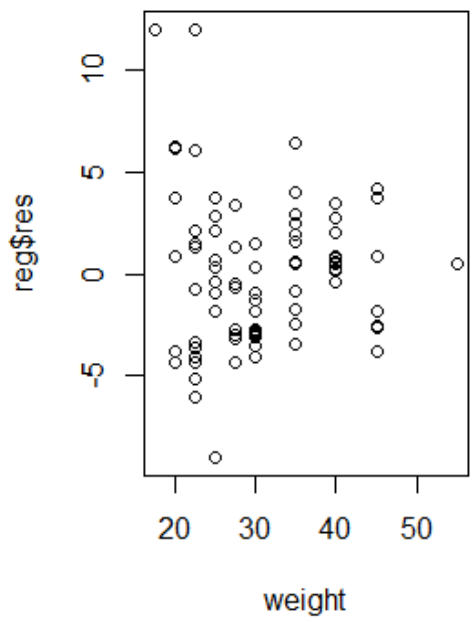
W pierwszej kolejności rozważamy model pełny. Widzimy, że współczynniki R^2 i R_a^2 są wysokie. Możemy z nich wnioskować, że model wyjaśnia ponad 85% zmienności dla ***fuel_cons***. Dla zmiennych ***engine_pow, top_speed*** oraz ***weight*** p-wartość jest odpowiednio niska, co świadczy, że są one zmiennymi istotnymi. Zmienna ***cab_vol*** jest zmienną nieistotną, co sugeruje nam rozważenie modelu bez tej zmiennej.

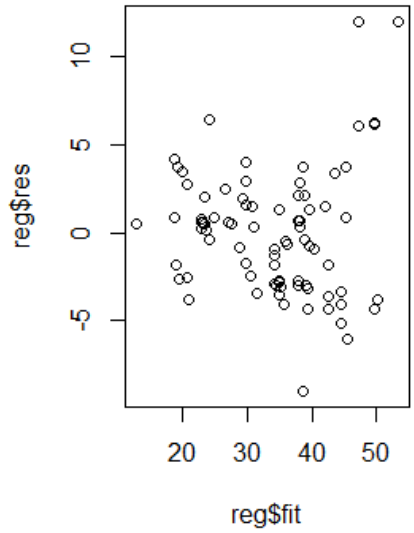
```
Call:
lm(formula = fuel_cons ~ ., data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  192.43775   23.53161   8.178 4.62e-12 ***
` cab_vol `   -0.01565    0.02283  -0.685    0.495
engine_pow     0.39221    0.08141   4.818 7.13e-06 ***
` top_speed`  -1.29482    0.24477  -5.290 1.11e-06 ***
weight        -1.85980    0.21336  -8.717 4.22e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-squared:  0.8733,    Adjusted R-squared:  0.8667
F-statistic: 132.7 on 4 and 77 DF,  p-value: < 2.2e-16
```

Wykresy reszt



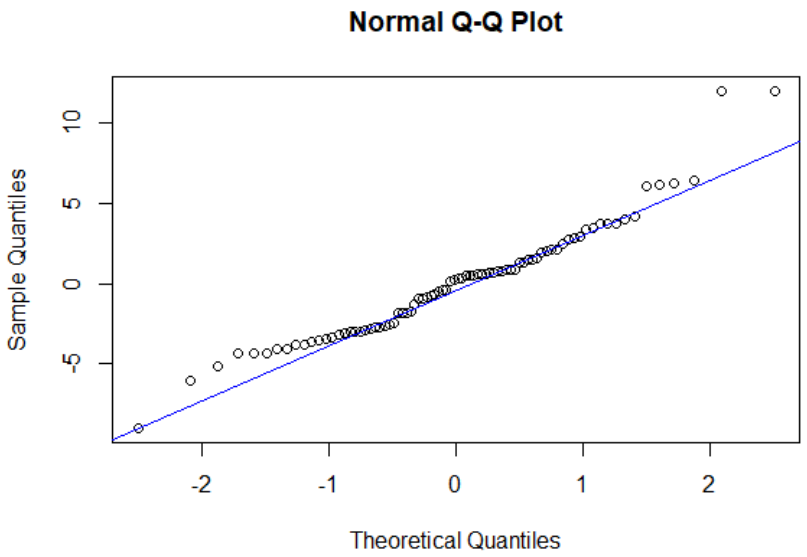


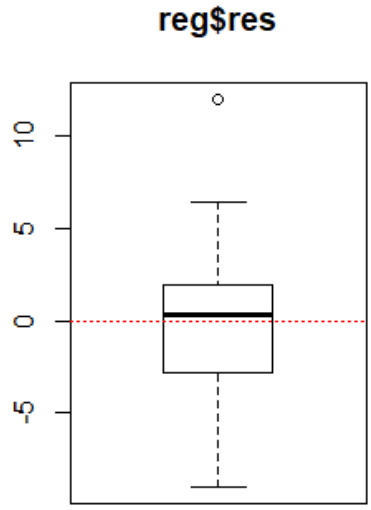




Wykresy przedstawiają zależność reszt od zmiennych niezależnych oraz od dopasowanych wartości.

Możemy zauważyć, że reszty układają się losowo, co świadczy od braku zależności.



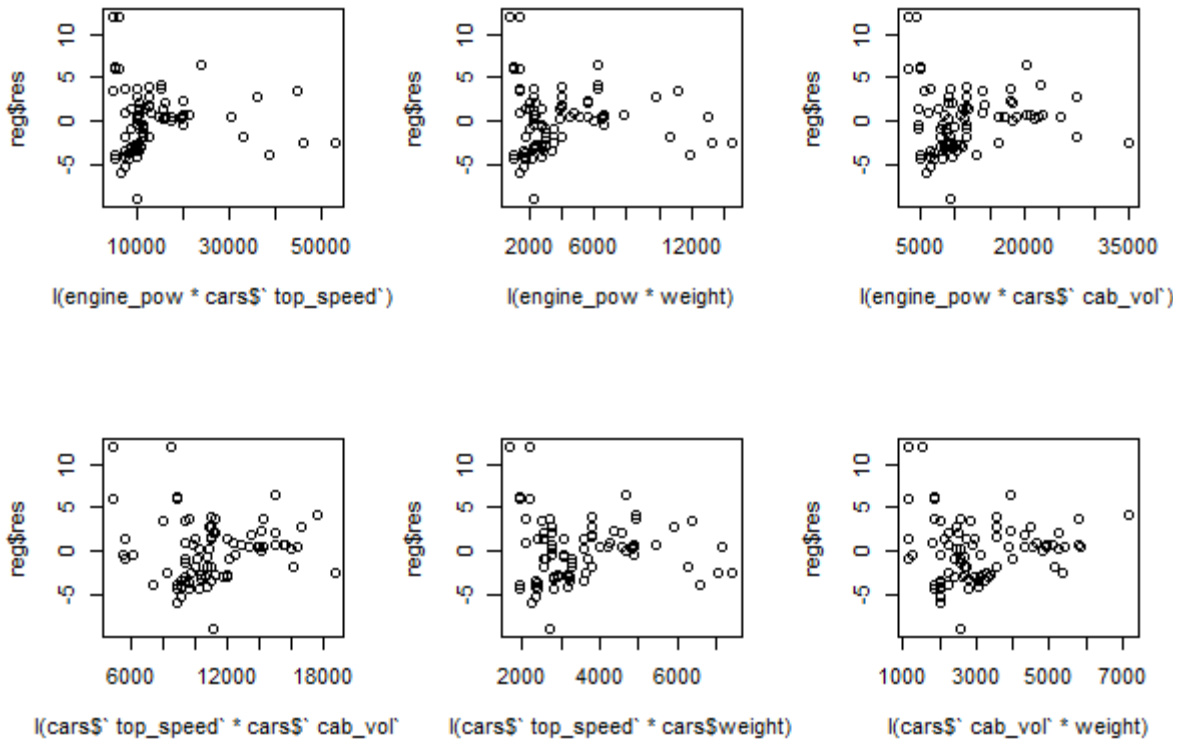


Powyżej widzimy, że ogony oddalają się od teoretycznych wartości, co sugeruje brak normalności reszt . W celu pogłębienia analizy przeprowadzimy test Shapiro-Wilka.

```
        Shapiro-Wilk normality test

data:  reg$res
W = 0.95053, p-value = 0.003196
```

Uzyskana p-wartość jest mniejsza od 0,05, zatem odrzucamy hipotezę o normalności reszt.

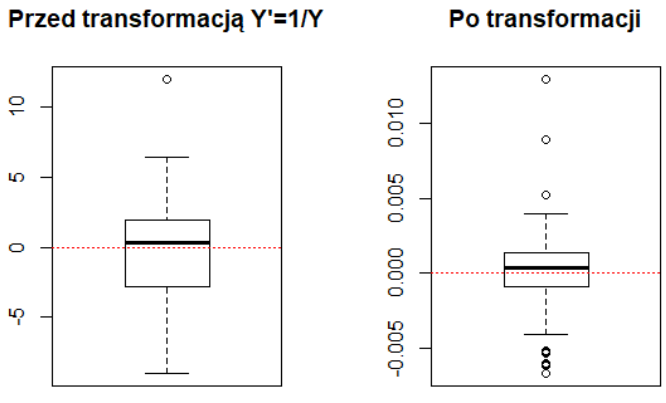
W kolejnym kroku sprawdzimy, czy współczynniki interakcji między zmiennymi mają wpływ na zachowanie się reszt.



Nie wykresach nie obserwujemy zależności pomiędzy resztami, a interakcjami. Z tego powodu nie będziemy rozważać modelu z interakcjami.

Transformacje zmiennych

Sprawdzimy jaki wpływ na zmienne mają transformacje zmiennej zależnej. Po przeprowadzeniu transformacji typu: ***Y'=ln(Y)***, ***Y'=exp(Y)***, ***Y'=sqrt(Y)***, ***Y'=(Y)^2***, ***Y'=1/Y*** najlepszy wynik otrzymaliśmy dla ***Y'=1/Y***.



```
Call:
lm(formula = fc ~ ., data = trans1)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0067174 -0.0009094  0.0003196  0.0013247  0.0129870

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.611e-02  2.027e-02   2.275  0.02570 *
X.cab_vol.   1.727e-06  1.966e-05   0.088  0.93022
engine_pow   2.213e-04  7.013e-05   3.156  0.00228 **
X.top_speed -4.997e-04  2.108e-04  -2.370  0.02028 *
weight       5.311e-04  1.838e-04   2.890  0.00501 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003146 on 77 degrees of freedom
Multiple R-squared:  0.9179,    Adjusted R-squared:  0.9136
F-statistic: 215.1 on 4 and 77 DF,  p-value: < 2.2e-16
```

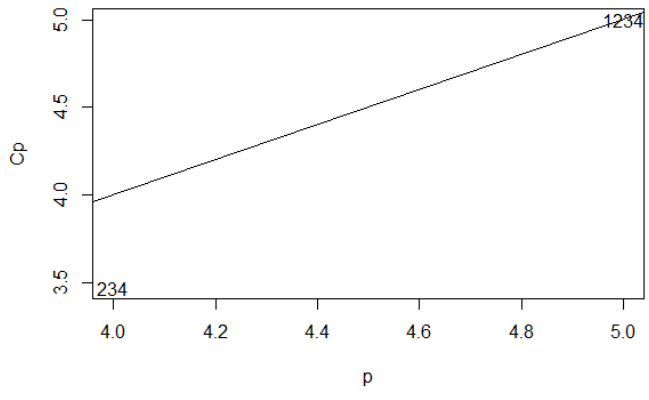
Przy transformacji ***Y'=1/Y*** możemy zaobserwować wzrost współczynników R^2 i R_a^2 do ponad 91%, co jest bardzo dobrym wynikiem. Widzimy również, ze wykres reszt jest bardziej symetryczny.

Redukcja ilości zmiennych niezależnych

W celu redukcji ilości zmiennych niezależnych skorzystamy z kryteriów Akaike i CpMallowa. Najpierw rozważymy model pełny.

```
Start:  AIC=217.3
fuel_cons ~ ` cab_vol ` + engine_pow + ` top_speed` + weight

               Df Sum of Sq    RSS    AIC
- ` cab_vol `   1      6.27 1033.7 215.80
<none>                     1027.4 217.30
- engine_pow    1    309.67 1337.0 236.90
- ` top_speed`  1    373.36 1400.7 240.72
- weight        1   1013.76 2041.2 271.59

Step:  AIC=215.8
fuel_cons ~ engine_pow + ` top_speed` + weight

               Df Sum of Sq    RSS    AIC
<none>                     1033.7 215.80
- engine_pow    1    349.37 1383.0 237.68
- ` top_speed`  1    396.97 1430.6 240.45
- weight        1   1322.87 2356.5 281.37

Call:
lm(formula = fuel_cons ~ engine_pow + ` top_speed` + weight,
    data = cars)

Coefficients:
 (Intercept)    engine_pow  ` top_speed`        weight
    194.1296        0.4052       -1.3200       -1.9221
```

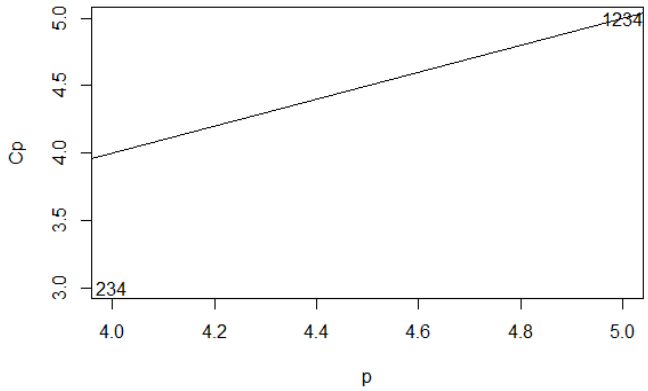
Kryterium Akaike (wyniki powyżej) jako najlepszy model dla danych nietransformowanych wybiera model bez zmiennej ***cab_vol***.



Zgodnie z wykresem, kryterium CpMallowa jako najlepszy wskazuje model ze zmiennymi ***engine-_power*** , ***top_speed*** oraz ***weight***, co odpowiada modelowi wybranemu przez kryterium Akaike.

Te same kryteria zastosujemy do modelu z transformacją.

```
               Df  Sum of Sq        RSS     AIC
- X.cab_vol.    1 7.6000e-08 0.00076235 -942.04
<none>                       0.00076227 -940.05
- X.top_speed  1 5.5616e-05 0.00081789 -936.27
- weight       1 8.2663e-05 0.00084493 -933.60
- engine_pow   1 9.8617e-05 0.00086089 -932.07

Step:  AIC=-942.04
fc ~ engine_pow + X.top_speed + weight

               Df  Sum of Sq        RSS     AIC
<none>                       0.00076235 -942.04
- X.top_speed  1 5.6267e-05 0.00081861 -938.20
- engine_pow   1 1.0291e-04 0.00086525 -933.66
- weight       1 1.0362e-04 0.00086597 -933.59

Call:
lm(formula = fc ~ engine_pow + X.top_speed + weight, data = trans1)

Coefficients:
(Intercept)   engine_pow  X.top_speed       weight
  0.0459214    0.0002199   -0.0004970    0.0005380
```



Dla modelu z transformacją oba kryteria wskazały te same rezultaty, co dla modelu pełnego.
W obu przypadkach najlepszym modelem okazał się model ze zmiennymi ***engine_power*** , ***top_speed*** oraz ***weight.***

Wybór modelu

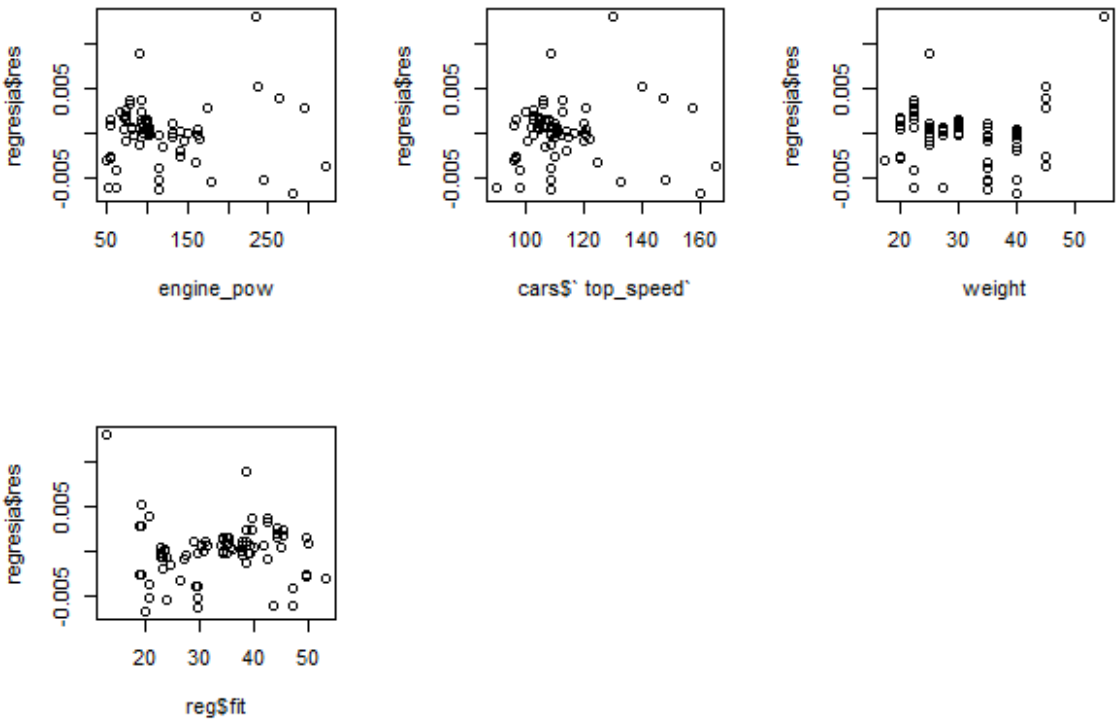
Po przeprowadzonej analizie wybieramy model transformowany:

$$\frac{1}{Y} = \beta_0 + \beta_1 \text{engine\_pow} + \beta_2 \text{ top\_speed} + \beta_3 \text{weight} + \varepsilon$$

Przeprowadzenie regresji dla tego modelu daje nam następujące wyniki.

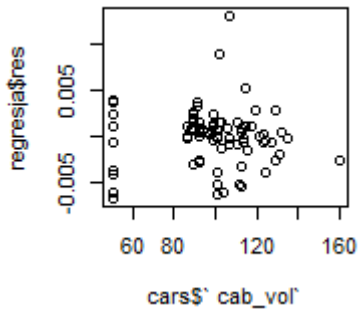
```
Call:
lm(formula = fc ~ ., data = model)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0067633 -0.0008944  0.0003524  0.0013145  0.0129569

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.592e-02  2.003e-02   2.293  0.02456 *
engine_pow   2.199e-04  6.777e-05   3.245  0.00173 **
X.top_speed -4.970e-04  2.071e-04  -2.399  0.01881 *
weight       5.380e-04  1.652e-04   3.256  0.00167 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003126 on 78 degrees of freedom
Multiple R-squared:  0.9179,    Adjusted R-squared:  0.9147
F-statistic: 290.5 on 3 and 78 DF,  p-value: < 2.2e-16
```

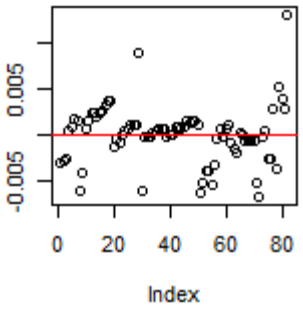
Wykresy reszt od zmiennych niezależnych oraz dopasowania dla wybranego modelu



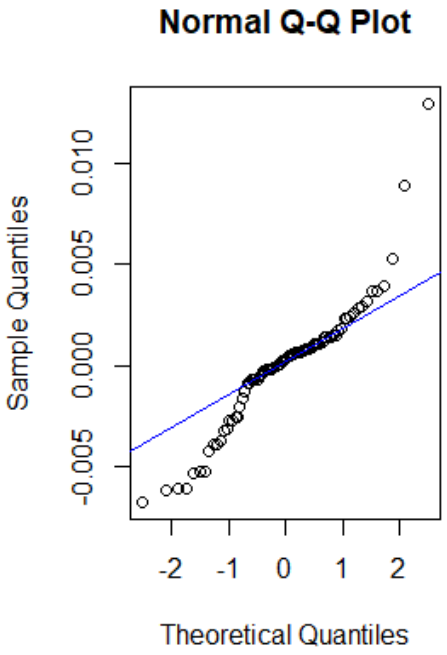
Na wykresach reszt regresji nie dostrzegamy żadnej zależności od zmiennych niezależnych.

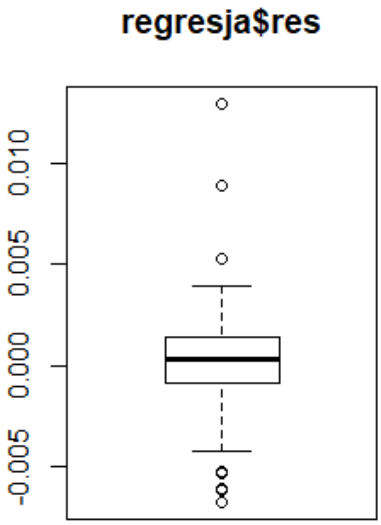
Sprawdzimy czy nie ma zależności reszt od zmiennej odrzuconej. Zgodnie z poniższym wykresem (po lewej), reszty układają się przypadkowo. Po prawej stronie znajduje się wykres reszt w zależności od czasu. Wartości na wykresie są rozłożone losowo, w okół zera.





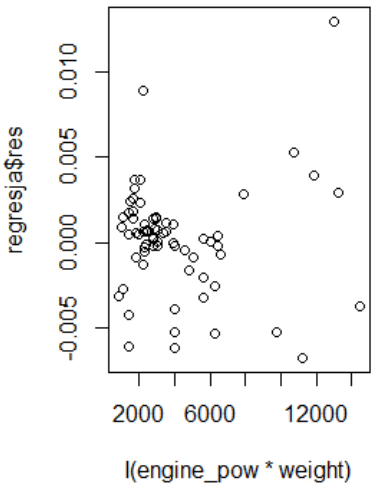

```
        Shapiro-Wilk normality test

data:  regresja$res
W = 0.90954, p-value = 2.548e-05
```

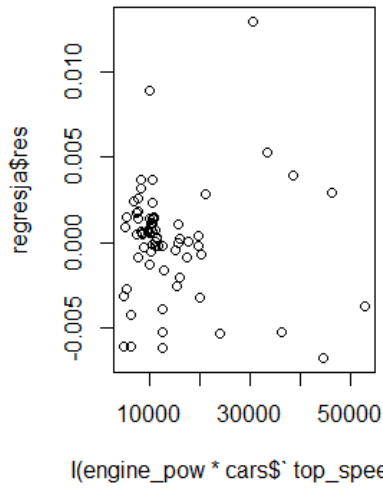


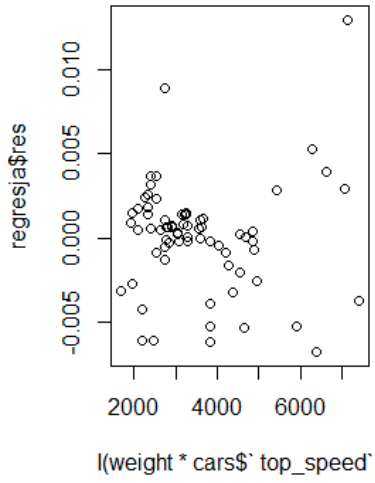


Wykres QQplot oraz mała p-wartość w teście Shapiro-Wilka wskazują na brak normalności reszt.

Poniżej znajdują się wykresy reszt od interakcji.







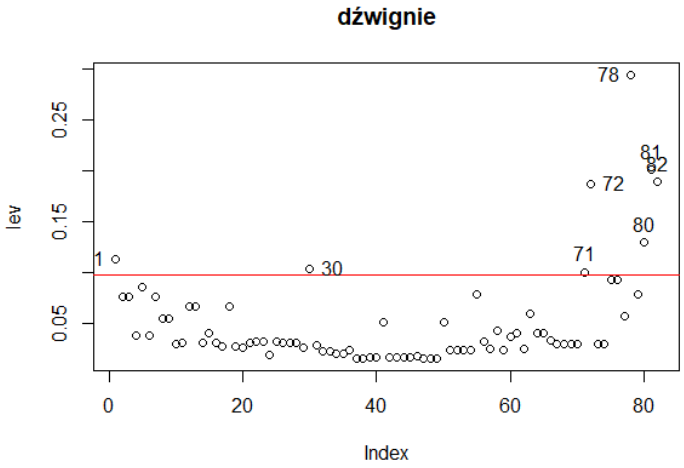
Współliniowość zmiennych

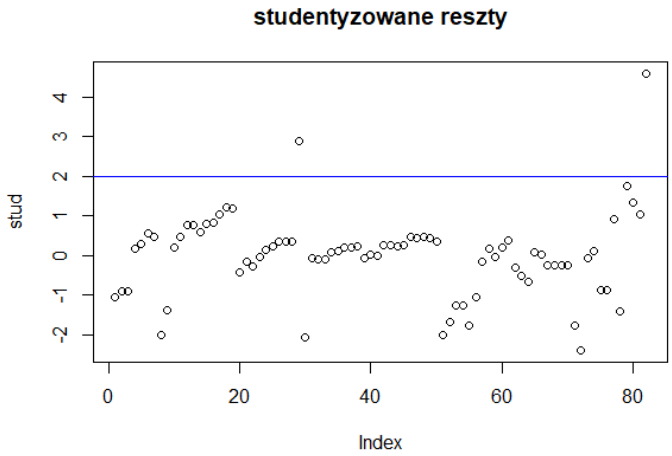
```
 vif(regresja)
engine_pow X.top_speed      weight
 122.97665    70.06073    14.99409
```

Otrzymaliśmy zaskakująco wysokie wartości, które wskazują na silną współliniowość zmiennych.

Obserwacje odstające

Aby sprawdzić jaki wpływ na przyjęty model mają obserwacje odstające wykonamy wykresy dźwigni dla zmiennych niezależnych oraz studentyzowanych reszt dla zmiennej zależnej.





Na wykresie po lewej stronie za odstające uznajemy te obserwacje, które znajdują się powyżej czerwonej linii. Wykres po prawej stronie przedstawia obserwacje wpływowe - te które znajdują się nad linią niebieską.

Wnioski

Po przeprowadzeniu gruntownej analizy regresji widzimy, że istnieje silna zależność pomiędzy zmiennymi *engine_pow, top_speed i weight*. Możemy wnioskować, że największy wpływ na zużycie paliwa mają właśnie pojemność kabiny, maksymalna prędkość oraz waga. Nasze przypuszczenia potwierdziła niska p- wartość dla *engine_pow, top_speed* oraz *weight*. Zmienną nieistotną okazała się być zmienna *cab_vol*, która odpowiada nam za pojemność kabiny.