Politechnika Krakowska
Wydział Informatyki i Telekomunikacji

Sprawozdanie z przedmiotu:

Statystyka i Probabilistyka

Projekt nr 2

Temat:

Regresja Wielokrotna

Wykonał: **Rafał Gęgotek**
Kierunek: Informatyka
Stopień studiów: II stopnia
Specjalizacja: Data Science
Rok akademicki: 2020/2021

1. Cel projekty

Celem projektu jest zastosowanie regresji wielokrotnej dla zbadania zależności zmiennych objaśniających dla konkretnej zmiennej zależnej, a także poznanie podstawowych miar dopasowania modelu regresji i sposobu oceny otrzymanego modelu.

W ramach projektu należy znaleźć odpowiedni zestaw danych, który zostanie poddany analizie wykonanej na nim regresji wielokrotnej, w oparciu o techniki poznane na zajęciach projektowych i wyciągnięciu odpowiednich wniosków.

2. Zbiór danych

Badany zestaw danych dotyczą statystyk pojazdów samochodowych z roku 1985. Zbiór został upubliczniony przez Jeffrey C. Schlimmera i jako jedno ze źródeł podaję się pozycję: „Specyfikacje samochodów i ciężarówek importowanych modeli z roku 1985”. Dane były używane w pracach naukowych, dotyczących między innymi porównania skuteczności regresji liniowej i algorytmu uczenia się opartego na instancjach IBL.

Oryginalny zestaw danych zawiera 206 instancji i 26 atrybutów, natomiast zredukowany o wartości brakujące zawiera 160 instancji i 16 atrybutów. Zdjęcie poglądowe poniżej.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	symboling	normalized losses	wheel base	length	width	height	curb weight	engine size	bore	stroke	compression ratio	horsepower	peak-rpm	city mpg	highway mpg	class
2	2	164	99.8	176.6	66.2	54.3	2337	109	3.19	3.4	10	102	5500	24	30	13950
3	2	164	99.4	176.6	66.4	54.3	2824	136	3.19	3.4	8	115	5500	18	22	17450
4	1	158	105.8	192.7	71.4	55.7	2844	136	3.19	3.4	8.5	110	5500	19	25	17710
5	1	158	105.8	192.7	71.4	55.9	3086	131	3.13	3.4	8.3	140	5500	17	20	23875
6	2	192	101.2	176.8	64.8	54.3	2395	108	3.5	2.8	8.8	101	5800	23	29	16430
7	0	192	101.2	176.8	64.8	54.3	2395	108	3.5	2.8	8.8	101	5800	23	29	16925
8	0	188	101.2	176.8	64.8	54.3	2710	164	3.31	3.19	9	121	4250	21	28	20970
9	0	188	101.2	176.8	64.8	54.3	2765	164	3.31	3.19	9	121	4250	21	28	21105
10	2	121	88.4	141.1	60.3	53.2	1488	61	2.91	3.03	9.5	48	5100	47	53	5151
11	1	98	94.5	155.9	63.6	52	1874	90	3.03	3.11	9.6	70	5400	38	43	6295
12	0	81	94.5	158.8	63.6	52	1909	90	3.03	3.11	9.6	70	5400	38	43	6575
13	1	118	93.7	157.3	63.8	50.8	1876	90	2.97	3.23	9.41	68	5500	37	41	5572
14	1	118	93.7	157.3	63.8	50.8	1876	90	2.97	3.23	9.4	68	5500	31	38	6377
15	1	118	93.7	157.3	63.8	50.8	2128	98	3.03	3.39	7.6	102	5500	24	30	7957
16	1	148	93.7	157.3	63.8	50.6	1967	90	2.97	3.23	9.4	68	5500	31	38	6229
17	1	148	93.7	157.3	63.8	50.6	1989	90	2.97	3.23	9.4	68	5500	31	38	6692
18	1	148	93.7	157.3	63.8	50.6	1989	90	2.97	3.23	9.4	68	5500	31	38	7609
19	-1	110	103.3	174.6	64.6	59.8	2535	122	3.34	3.46	8.5	88	5000	24	30	8921
20	3	145	95.9	173.2	66.3	50.2	2811	156	3.6	3.9	7	145	5000	19	24	12964

Rysunek 1. Wycinek tabeli badanego zestawu danych

Zbiór danych zawiera liczną grupę parametrów, jednakże na cele projektu wykorzystana zostanie część z nich. Zmienną objaśnianą zaznaczoną kolorem szarym jest zużycie paliwa mierzone w milach na galon. Natomiast zmiennymi objaśniajacymi zaznaczonymi na rysunku nr 1 kolorem jasnoniebieskim są parametry:

- wheel base – rozstaw osi pojazdu,
- curb weight – masa własna pojazdu,
- engine size – pojemność skokowa silnika,
- stroke – skok tłoka w cylindrze
- compression ratio – stopień sprężania,
- horsepower - ilość koni mechanicznych pojazdu (moc silnika),

Głównym celem badanego zbioru jest zbadanie zależności pomiędzy zużyciem paliwa, a wybranymi parametrami. Należy sprawdzić czy model regresji wielokrotnej będzie dobrze odzwierciedlał zależność pomiędzy danymi, a jeżeli tak, to poddać analizie otrzymane wyniki, aby stwierdzić w jak dużym stopniu danę są ze sobą powiązane.

3. Optymalny wybór zmiennych niezależnych

W ramach tego punktu należy podjąć decyzję o słuszności wyboru zmiennych objaśniających. W przypadku gdyby model miał lepszą bądź taką samą skuteczność dzięki zastosowaniu mniejszej liczby parametrów, to zmienną nieistotne można usunąć z tego modelu.

W tym celu pierwszym krokiem jest analiza dwóch początkowych problemów dopasowania modelu regresji. Mianowicie czy **zbiór jest zbyt mały**, a dokładniej czy danych jest więcej od 6-krotności liczby zmiennych niezależnych. W tym przypadku zbiór zawiera 160 instancji co jest większe niż 90. Kolejnym problemem jest **autokorelacja danych**, również dla badanego przykładu nie ma szeregów czasowych, więc można wyeliminować te problemy.

Kolejnym krokiem jest **zbadanie korelacji zmiennych niezależnych**, w tym celu można posłużyć się metodą korelacji *Pearsona*.

Na podstawie otrzymanych wyników z rysunku nr 2 można wywnioskować, iż waga pojazdu jest silnie skorelowana z rozkładem osi, pojemnością silnika oraz jego mocą, Dwa ostatnie wymienione są również ze sobą silnie skorelowane.

```
> cor(data[2:7], method="pearson")
                  curb_weight wheel_base engine_size    stroke compression_ratio horsepower
curb_weight         1.0000000  0.8101815   0.8886261 0.1738444         0.2247240  0.7900954
wheel_base          0.8101815  1.0000000   0.6492056 0.1674487         0.2914314  0.5169475
engine_size         0.8886261  0.6492056   1.0000000 0.2996831         0.1410967  0.8120726
stroke              0.1738444  0.1674487   0.2996831 1.0000000         0.2435868  0.1488038
compression_ratio   0.2247240  0.2914314   0.1410967 0.2435868         1.0000000 -0.1623052
horsepower          0.7900954  0.5169475   0.8120726 0.1488038        -0.1623052  1.0000000
```

Rysunek 2. Macierz korelacji zmiennych niezależnych

Następnym etapem jest zbadanie **R^2**, w tym przypadku pokazanym na rysunku nr 2 współczynnik ten wynosi 0.8256, natomiast jego **test istotności** podaje wartość *p* mniejszą od 2.2e-16, dzięki czemu można zdecydowanie odrzucić Hipotezę Zerową, iż „R kwadrat nie różni się istotnie od zera".

Dodatkowo **R^2 dostosowany** wynosi 0.8187, a więc nie różni się o więcej niż 5% od R^2. Dlatego też można stwierdzić, iż model jest istotny, oraz zmienne objaśniające bardzo dobrze odzwierciedlają zmienną zależną

```
> model <- lm(city_mpg ~ wheel_base + curb_weight + engine_size + stroke + compression_ratio + horsepower)
> summary(model)

Call:
lm(formula = city_mpg ~ wheel_base + curb_weight + engine_size + 
    stroke + compression_ratio + horsepower)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1634 -1.2901 -0.5776  0.8356 13.5396 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       54.336210   5.878889   9.243  < 2e-16 ***
wheel_base        -0.138742   0.073908  -1.877   0.0624 .  
curb_weight       -0.007154   0.001466  -4.878 2.67e-06 ***
engine_size        0.027364   0.017380   1.574   0.1175    
stroke             0.608516   0.786412   0.774   0.4403    
compression_ratio  0.548978   0.068915   7.966 3.58e-13 ***
horsepower        -0.077056   0.014862  -5.185 6.80e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.596 on 152 degrees of freedom
  (239 observations deleted due to missingness)
Multiple R-squared:  0.8256,	Adjusted R-squared:  0.8187 
F-statistic: 119.9 on 6 and 152 DF,  p-value: < 2.2e-16
```

Rysunek 3. Podsumowanie modelu regresji dla 6 zmiennych objaśniających

Ponadto z tego wykresu możemy jeszcze odczytać informacje, o jakości zmiennych objaśniających. Mianowicie dla każdej zmiennej obok istotności z *testu p* jest dołączona pewna liczba gwiazdek, która określa, jak bardzo dany parametr jest istotny. Dzięki temu możemy stwierdzić, że zmienne *pojemność silnika* oraz *skok tłoka* nie są istotne (parametry dla tych zmiennych nie będą się istotnie różnić od zera).

Biorąc pod uwagę fakt istotności zmiennej *stroke* możemy spróbować usunąć ją z modelu i sprawdzić jak wpłynie to na wynik regresji.

```
> model <- lm(city_mpg ~ wheel_base + curb_weight + engine_size + compression_ratio + horsepower)
> summary(model)

Call:
lm(formula = city_mpg ~ wheel_base + curb_weight + engine_size +
    compression_ratio + horsepower)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2339 -1.2193 -0.5898  0.7545 13.6505

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       55.670907   5.612777   9.919  < 2e-16 ***
wheel_base        -0.132149   0.073319  -1.802   0.0735 .
curb_weight       -0.007462   0.001410  -5.294 4.10e-07 ***
engine_size        0.031885   0.016346   1.951   0.0529 .
compression_ratio  0.562191   0.066678   8.431 2.37e-14 ***
horsepower        -0.076313   0.014811  -5.152 7.83e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 153 degrees of freedom
  (239 observations deleted due to missingness)
Multiple R-squared:  0.8249,	Adjusted R-squared:  0.8192
F-statistic: 144.1 on 5 and 153 DF,  p-value: < 2.2e-16
```

Rysunek 4. Podsumowanie modelu regresji dla 5 zmiennych objaśniających

Jak widać na powyższym rysunku wyrzucenie tej zmiennej nie spowodowało spadku R^2 oraz R^2 dostosowanego, natomiast polepszyła się istotność parametru pojemności silnika. W dalszym ciągu można spróbować usunąć kolejną najmniej znaczącą zmienną objaśniającą, w tym przypadku jest to długość rozstawu osi.

```
> model <- lm(city_mpg ~  curb_weight + engine_size + compression_ratio + horsepower)
> summary(model)

Call:
lm(formula = city_mpg ~ curb_weight + engine_size + compression_ratio +
    horsepower)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5410 -1.3624 -0.5302  0.8032 14.2700

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       45.767428   1.153643  39.672  < 2e-16 ***
curb_weight       -0.009147   0.001062  -8.612 7.93e-15 ***
engine_size        0.036299   0.016280   2.230   0.0272 *
compression_ratio  0.560550   0.067157   8.347 3.75e-14 ***
horsepower        -0.070495   0.014561  -4.841 3.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.612 on 154 degrees of freedom
  (239 observations deleted due to missingness)
Multiple R-squared:  0.8212,	Adjusted R-squared:  0.8165
F-statistic: 176.8 on 4 and 154 DF,  p-value: < 2.2e-16
```

Rysunek 5. Podsumowanie modelu regresji dla 4 zmiennych objaśniających

Dla tego przykładu można zaobserwować bardzo niewielki spadek R^2 oraz R^2 dostosowany, natomiast ponownie zmiana polepszyła istotność parametru pojemności silnika, w którym wartość testu p zmalała poniżej 5%.

Pomimo, iż zmienne zależne wskazują teraz na dobry poziom objaśniania zmiennej niezależnej, to nadal można starać się usunąć z modelu zmienne, które wcześniej wskazywały na korelacje pomiędzy sobą, a mianowicie wagę pojazdu i moc silnika wyrażaną w koniach mechanicznych.

```
> model <- lm(city_mpg ~ wheel_base + curb_weight +
+ engine_size + stroke + compression_ratio)
> summary(model)

Call:
lm(formula = city_mpg ~ wheel_base + curb_weight + engine_size +
    stroke + compression_ratio)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3305 -1.5290 -0.4170  0.6186 14.3046

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       49.0710334  6.2611575   7.837 7.25e-13 ***
wheel_base        -0.0531031  0.0778943  -0.682    0.496
curb_weight       -0.0105885  0.0014145  -7.485 5.25e-12 ***
engine_size        0.0003528  0.0179280   0.020    0.984
stroke             0.3452168  0.8485551   0.407    0.685
compression_ratio  0.7449480  0.0623089  11.956  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.807 on 153 degrees of freedom
Multiple R-squared:  0.7947,	Adjusted R-squared:  0.788
F-statistic: 118.5 on 5 and 153 DF,  p-value: < 2.2e-16
```

```
> model <- lm(city_mpg ~ wheel_base + engine_size +
+ stroke + compression_ratio + horsepower)
> summary(model)

Call:
lm(formula = city_mpg ~ wheel_base + engine_size + stroke + co
    horsepower)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9667 -1.4199 -0.5356  0.5738 13.2932

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       67.00302    5.65376  11.851  < 2e-16 ***
wheel_base        -0.37877    0.05912  -6.407 1.74e-09 ***
engine_size       -0.02092    0.01531  -1.366   0.1740
stroke             1.64933    0.81136   2.033   0.0438 *
compression_ratio  0.43486    0.06949   6.258 3.73e-09 ***
horsepower        -0.10981    0.01421  -7.726 1.36e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.783 on 153 degrees of freedom
Multiple R-squared:  0.7983,	Adjusted R-squared:  0.7917
F-statistic: 121.1 on 5 and 153 DF,  p-value: < 2.2e-16
```

Rysunek 6. Podsumowanie modelów regresji dla 5 zmiennych objaśniających po usunięciu z lewej mocy silnika, z prawej wagi pojazdu

Jednakże dla pokazanych powyżej przykładów usuniecie tych zmiennych w podobny sposób negatywnie wpłynęło na wynik regresji dla badanego modelu.

Kolejnym krokiem jest zbadanie współliniowości zmiennych objaśniających, w tym celu należy wykorzystać **miarę VIF**. Pokazane poniżej wynik działania tej metody w programie R wskazują na znacząco większy od wartości 5 (wartość graniczna) parametr wagi pojazdu, a także lekko przekraczający tą wartość parametr pojemności skokowej

```
> model <- lm(city_mpg ~ wheel_base + curb_weight + engine_size + stroke + compression_ratio + horsepower)
> vif(mod=model)
       wheel_base       curb_weight       engine_size            stroke compression_ratio        horsepower
         3.419052         11.707407          6.569485          1.260620          1.684156          4.885673
> |
```

Rysunek 7. Kryterium VIF dla 6 zmiennych objaśniających

Po usunięciu z modelu najbardziej odstającego od normy parametru pod względem tego współczynnika, ponowna analiza wskazuje na poprawę dla większości parametrów, tak że żadna zmienna niezależna nie przekracza wartości granicznej. Można zauważyć powiązanie działania tej metody z macierzą korelacji, która ten konkretny parametr wskazywała jako najbardziej skorelowany z innymi.

```
> model <- lm(city_mpg ~ wheel_base + engine_size + stroke + compression_ratio + horsepower)
> vif(mod=model)
    wheel_base       engine_size            stroke compression_ratio        horsepower 
      1.903784          4.439067          1.167836          1.490112          3.888512 
> 
```

Rysunek 8. Kryterium VIF dla 5 zmiennych objaśniających

Kolejnym etapem, aby potwierdzić możliwość usunięcia zmiennych niezależnych z modelu bez znaczącego pogorszenia skuteczności regresji jest **kryterium C(p)**. Przedstawiony poniżej według metody **postępowego dołączania**, wskazuje, iż dla 5 z 6 zmiennych model jest równie istotny. Najsłabiej wypada parametr pojemności silnika którego wartość C(p) wynosi 5.5987, jednakże nadal jest to powyżej 5 co pozwala go zachować w modelu regresji.

```
> model <- lm(city_mpg ~ wheel_base + curb_weight + engine_size + stroke + compression_ratio + horsepower)
> ols_step_forward_p(model)

                                  Selection Summary                                    
--------------------------------------------------------------------------------------
        Variable                     Adj.                                                 
Step        Entered        R-Square    R-Square     C(p)         AIC          RMSE     
--------------------------------------------------------------------------------------
   1    horsepower           0.7009      0.6990    105.6152    839.1821    3.3450    
   2    wheel_base           0.7308      0.7273     81.6155    824.4717    3.1839    
   3    compression_ratio    0.7917      0.7877     30.5277    785.6818    2.8096    
   4    curb_weight          0.8205      0.8159      7.3935    763.9856    2.6163    
   5    engine_size          0.8249      0.8192      5.5987    762.0800    2.5928    
--------------------------------------------------------------------------------------
```

Rysunek 9. Test Ols Step Forward zastosowany dla zmiennych objaśniających

Podobnym test jednakże polegającym na **eliminacji**, pokazuje dokładnie, która zmienna powinna zostać usunięta z modelu. W tym przypadku jest to *stroke* (skok tłoka w cylindrze).

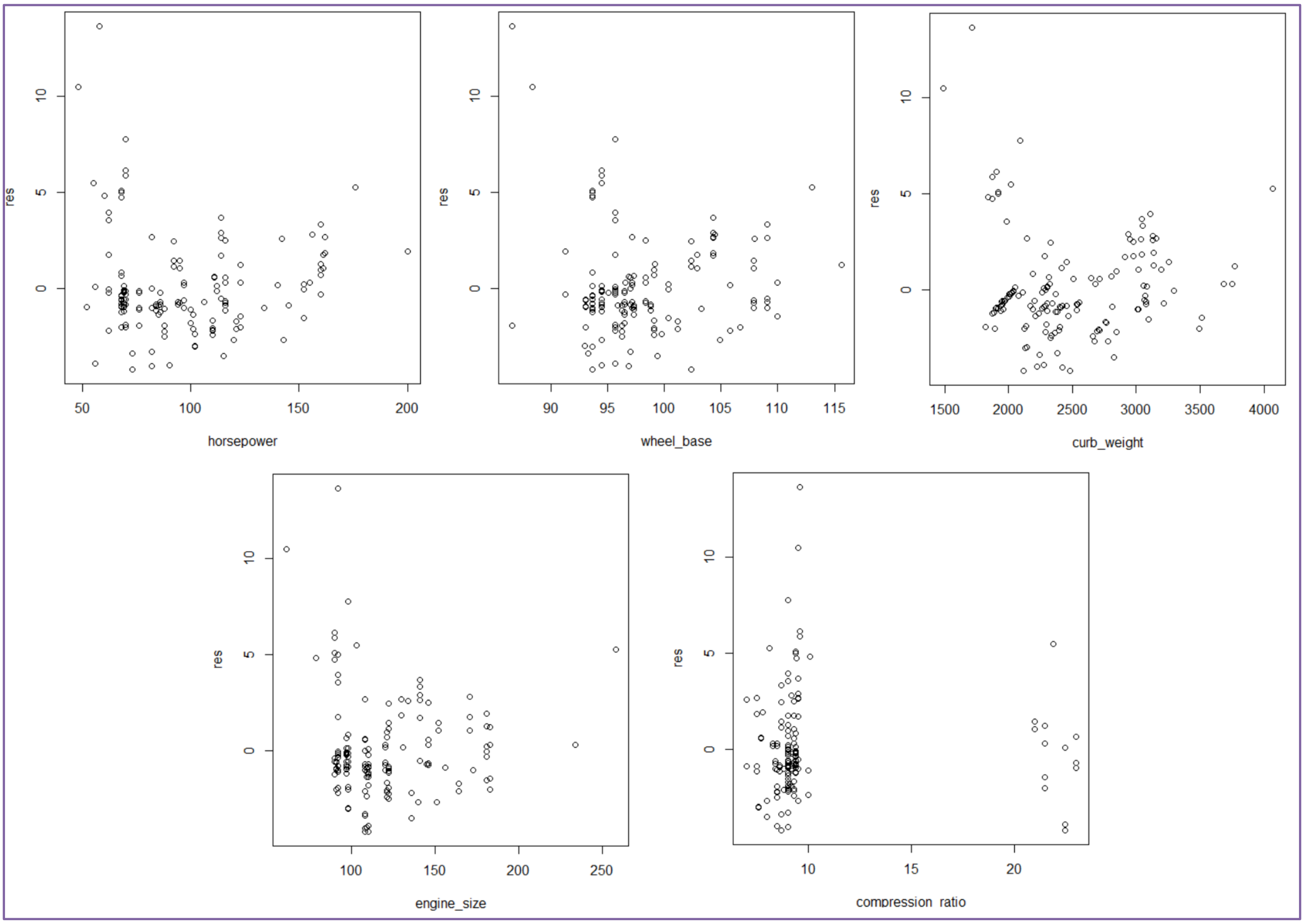
```
> model <- lm(city_mpg ~ wheel_base + curb_weight + engine_size + stroke + compression_ratio + horsepower)
> ols_step_backward_p(model)


                           Elimination Summary                            
-------------------------------------------------------------------------
        Variable                  Adj.                                        
Step    Removed     R-Square    R-Square     C(p)       AIC        RMSE     
-------------------------------------------------------------------------
   1    stroke        0.8249      0.8192    5.5987    762.0800    2.5928    
-------------------------------------------------------------------------
```

Rysunek 10. Test Ols Step Backward zastosowany dla zmiennych objaśniających

4. Analiza wrażliwości

W tym etapie w modelu uwzględniane będzie 5 parametrów, na które wskazywało kryterium C(p), z wyłączeniem *skoku tłoka*.

W pierwszym etapie zmienne zostały poddane **analizie homoskedastyczności**. W tym celu dla każdego parametru został stworzony wykres obrazujący jego zależność na tle standardowych składników resztowych dla pełnego modelu 5 zmiennych. Na podstawie wyników można stwierdzić, iż zachowana jest stałość wariancji reszt dla wszystkich zmiennych objaśniających.



Rysunek 11. Wykresy standardowych składników resztowych dla poszczególnych zmiennych niezależnych

Kolejnym etapem diagnostycznym w tej części, który możemy wykonać jest **test na autokorelację reszt**. Poniżej zostały zobrazowane wynik testu reszt Boxa dla autokorelacji reszt rzędu 1, 2, 3 oraz 4. Co oznacza że dla zastosowania operacji z parametrem rzędu 2, obliczony zostanie test autokorelacji pomiędzy bieżącą resztą, a resztą odległą o dwie obserwacje.

Jak można wywnioskować dla każdego przypadku nie można odrzucić Hipotezy Zerowej świadczącej o „braku autokorelacji standardowych składników resztowych", gdyż dla każdego podanego rzędu autokorelacji wartość p jest większa niż 5%.

```
> Box.test(res,lag=1)

        Box-Pierce test

data:  res
X-squared = 3.0876, df = 1, p-value = 0.07889

> Box.test(res,lag=2)

        Box-Pierce test

data:  res
X-squared = 4.8325, df = 2, p-value = 0.08926

> Box.test(res,lag=3)

        Box-Pierce test

data:  res
X-squared = 4.9423, df = 3, p-value = 0.1761

> Box.test(res,lag=4)

        Box-Pierce test

data:  res
X-squared = 5.0366, df = 4, p-value = 0.2836
```

Rysunek 12. Wynik autokorelacji reszt Boxa

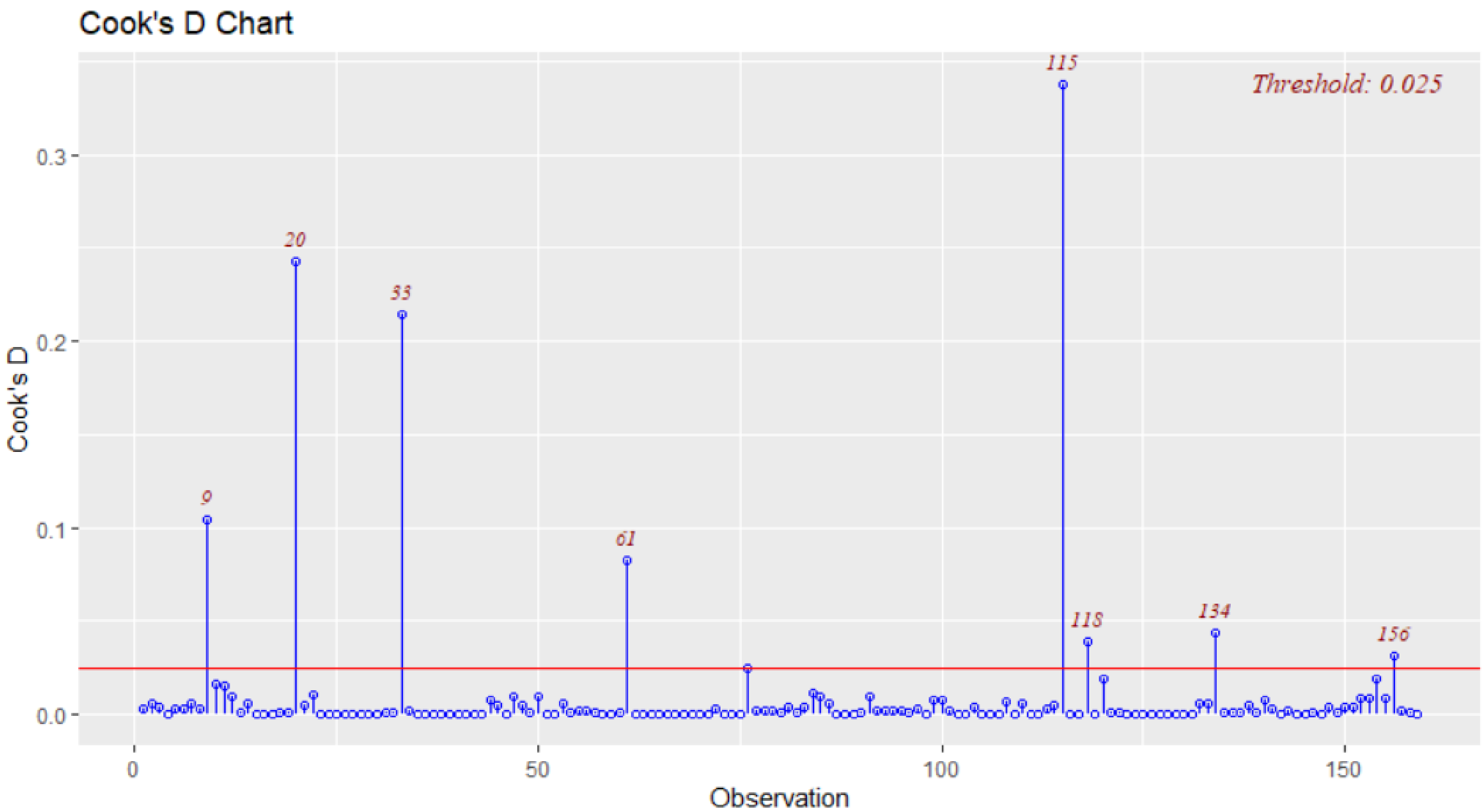
Standardowe składniki resztowe można poddać jeszcze **testowi normalności,** zobrazowany na przykładzie wywołania testu Shapiro-Wilka. Na podstawie którego możemy zdecydowanie odrzucić Hipotezę Zerową odnośnie Reszt. Obliczona wartość p pokazana na zdjęciu poniżej jest znacznie mniejsza niż 5%.

```
> res <- model$residuals
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.85567, p-value = 3.32e-11

> 
```

Rysunek 13. Wynik testu Shapiro-Wilka dla składników resztowych

W następnym etapie zostanie przeprowadzana **walidacja krzyzowa**. Jako pierwszy przykład ją opisujący została przedstawionat **Miara Cook'a**, polegajaca na usuwaniu po jednej obserwacji i obliczeniu dla każdego takiego zestawu predykcji modelu.

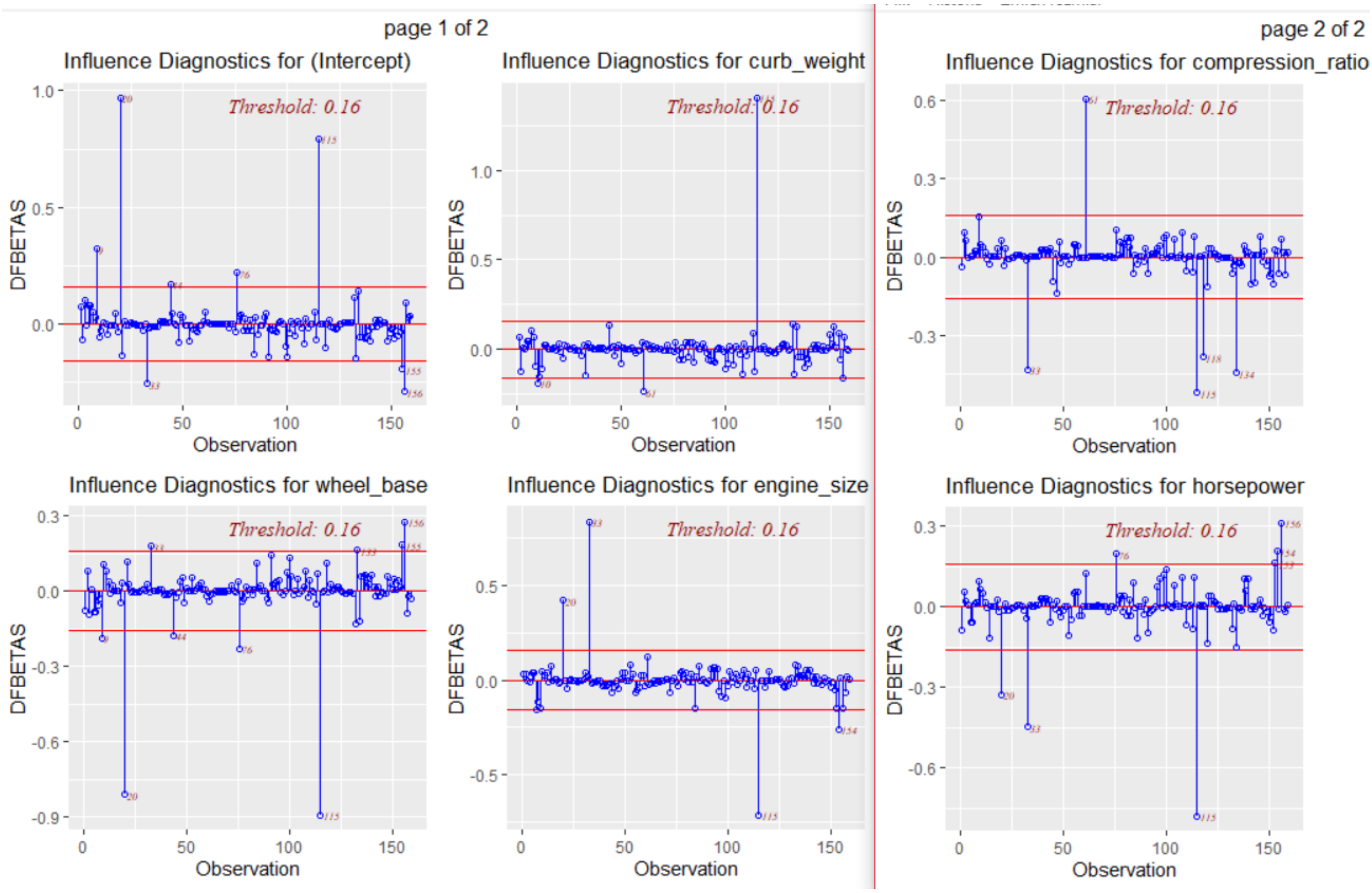
W wyniku dzialania metody obrazującej tą miarę możemy zobaczyć wykres, na podstawie którego można zidentyfikować obserwacje, zakwalifkowane jako odstające. Łącznie takich wartości dla zbioru 160 elementów jest 8, są to elemnty które przekraczją wartość progowa równą 0.025.



Rysunek 14. Wykres obrazujący wyniki dla miary Cook'a

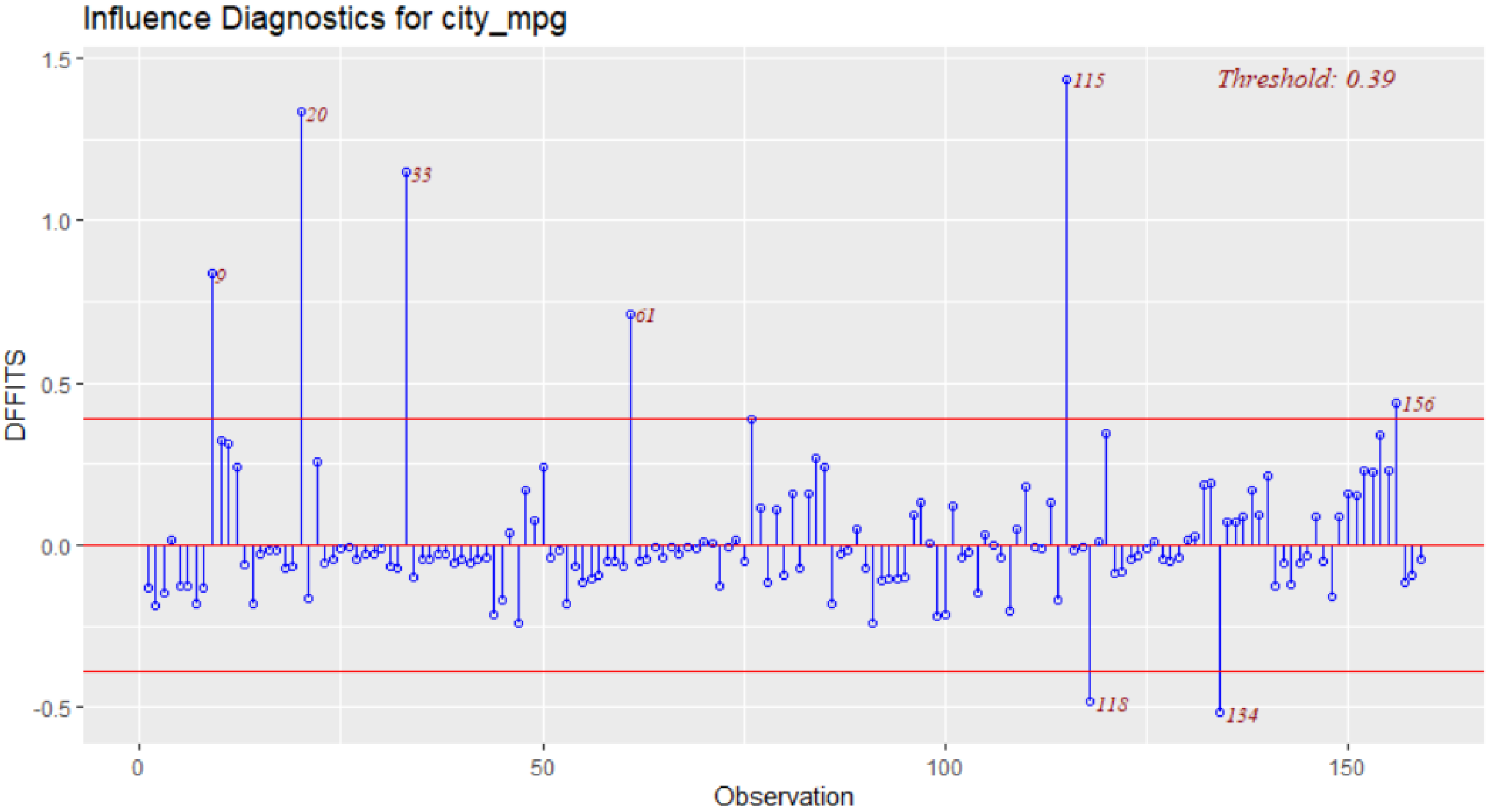
Kolejną miara jest **miara DFBetas**, która analizuje dopasowanie wszystkich zmiennych niezależnych. W odróżnieniu od poprzedniej miary, która ma formę kwadratową, a więc zawsze wartość jest dodatnia, ta miara może być ujemna, dlatego też w przedstawionym poniżej wyniku jej działania, możemy zauważyć dwa poziomy graniczne.

W tym przypadku dla łącznej ilości pokazanych wykresów (po jednej dla każdej zmiennej niezależnej oraz wyrazu wolnego) możemy naliczyć 13 różnych wartości odstających od normy (wartość progowa 0.16). Po mimo iż wartości odstających jest więcej niż w poprzednim przykładzie, to nadal jest ich mniej niż 10% całego zbioru, co mieści się w normie dotyczącej możliwości usuwania takich wartości w celu poprawy działania modelu.



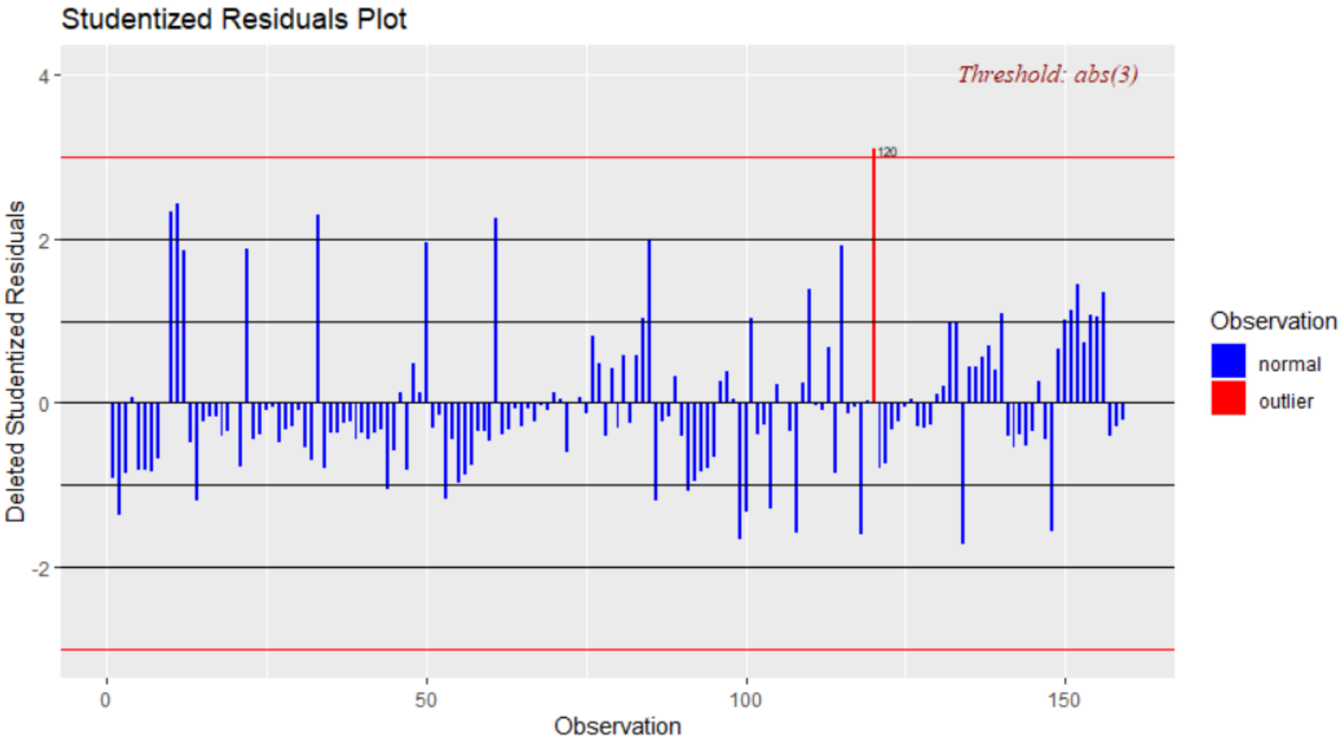
Rysunek 15. Wykresy obrazujące wyniki dla miary DFBetas

Następna jest **miara DFFITS**, podobnie w jej przypadku mamy dwa poziomy progowe jednakże wykres wizualizujący jej działanie jest jeden, gdyż miara ta analizuje dopasowanie modelu. Ponownie można zaobserwować podobne jak poprzednio wartości odstające, których łącznie jest 8. Dla tej miary wartość progowa wynosiła 0.39.



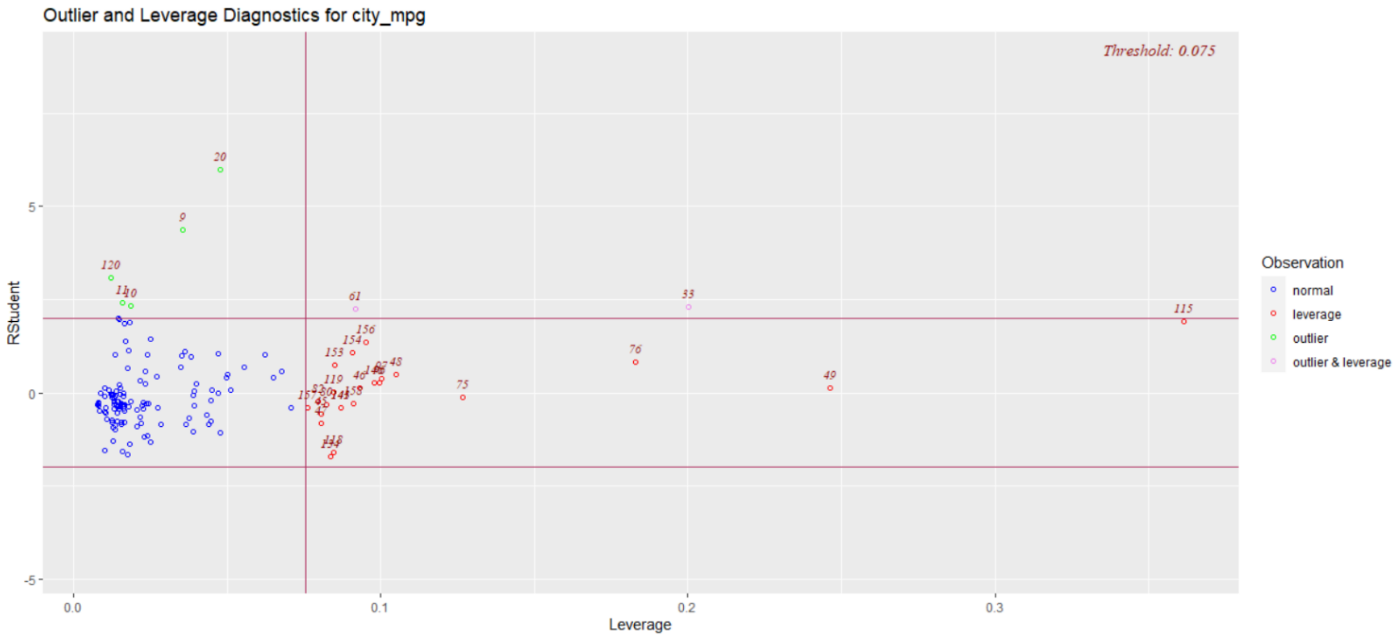
Rysunek 16. Wykresy obrazujące wyniki dla miary DFFITS

Ostatnią pokazaną miarą walidacji krzyżowej będą **studentyzowane reszty**. W tym przypadku możemy zaobserwować jedną wartość odstającą od norm w zakresie <-3;3>, będącą zidentyfikowaną jako 120 element zbioru.



Rysunek 17. Wykresy obrazujące wyniki dla miary studentyzowanych reszt

Dodatkowo w środowisku R możemy przedstawić również klasyfikację dla punktów normalnych, punktów będących obserwacją odstającymi, bycie punktem który jest ‘dźwigniowy’, a takie obserwacje które są po części dwoma poprzednimi. W takim przypadku można stwierdzić, że obserwacji które ‘sprawiają kłopoty’ jest więcej niż 10%.



Rysunek 18. Wykresy klasyfikacji miary studentyzowanych reszt

Tak jak wspomniano już wcześniej, jeżeli wartości odstający jest mniej niż 10% zbioru to można spróbować usunąć te pozycje i sprawdzić jak to wpłynie na działanie modelu. W każdym z zaprezentowanych miar wartości takie się pokrywały. Po usunięciu ich z modelu ponownie sprawdzono skuteczność modelu.

```
> summary(model)

Call:
lm(formula = city_mpg ~ wheel_base + curb_weight + engine_size +
    compression_ratio + horsepower)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6521 -1.0836 -0.3363  0.6833  8.0717

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       44.644376   5.323892   8.386 4.87e-14 ***
wheel_base         0.005508   0.070022   0.079   0.9374
curb_weight       -0.009334   0.001413  -6.608 7.54e-10 ***
engine_size        0.036225   0.015256   2.374   0.0189 *
compression_ratio  0.608034   0.061020   9.964  < 2e-16 ***
horsepower        -0.067418   0.013984  -4.821 3.67e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.994 on 140 degrees of freedom
Multiple R-squared:  0.8611,    Adjusted R-squared:  0.8561
F-statistic: 173.5 on 5 and 140 DF,  p-value: < 2.2e-16
```

Rysunek 19. Podsumowanie modelu regresji dla 5 zmiennych objaśniających, po usunięciu wartości odstających

Jak widać zmiana wpłynęła pozytywnie, pomimo iż współczynnik R^2 był dosyć wysoki, to jeszcze wzrósł o 4 punktu do wartości 0.8611, podobnie jak R^2 dostosowany.

Ponadto kryterium C(p) dla pokazanej poniżej metody stopniowego dołączania pokazuje, iż model jest równie skuteczny po usunięciu z niego parametru dotyczącego rozstawu osi, co potwierdza ponowna analiza modelu dla 4 zmiennych niezależnych.

```
> ols_step_forward_p(model)

                                 Selection Summary
-----------------------------------------------------------------------------------
        Variable                           Adj.
Step        Entered          R-Square    R-Square      C(p)          AIC        RMSE
-----------------------------------------------------------------------------------
   1    horsepower             0.7338      0.7319    126.2744    710.7291    2.7221
   2    compression_ratio      0.7573      0.7539    104.5912    699.2355    2.6083
   3    curb_weight            0.8552      0.8522      7.8815    625.7836    2.0214
   4    engine_size            0.8611      0.8571      4.0062    621.7819    1.9873
-----------------------------------------------------------------------------------
```

Rysunek 20. Test Ols Step Forward zastosowany dla zmiennych objaśniających, po usunięciu wartości odstających

Poniżej rysunek prezentujące ostateczny model regresji, po uwzględnieniu wszystkich wcześniej zanalizowanych kroków.

```
> model <- lm(city_mpg ~  curb_weight + engine_size + compression_ratio + horsepower)
> summary(model)

Call:
lm(formula = city_mpg ~ curb_weight + engine_size + compression_ratio + 
    horsepower)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6612 -1.0895 -0.3366  0.6793  8.0720 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       45.0564371  0.9444975  47.704  < 2e-16 ***
curb_weight       -0.0092543  0.0009755  -9.487  < 2e-16 ***
engine_size        0.0359431  0.0147763   2.432   0.0162 *  
compression_ratio  0.6078231  0.0607461  10.006  < 2e-16 ***
horsepower        -0.0677525  0.0132746  -5.104 1.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.987 on 141 degrees of freedom
Multiple R-squared:  0.8611,    Adjusted R-squared:  0.8571 
F-statistic: 218.4 on 4 and 141 DF,  p-value: < 2.2e-16
```

Rysunek 21. Podsumowanie modelu regresji dla 4 zmiennych objaśniających, po usunięciu wartości odstających i zredukowaniu o zmienna nieistotną

5. Wnioski

Dzięki przeprowadzonym zajęciom laboratoryjnym mogliśmy bliżej poznać zasadę funkcjonowania regresji wielokrotnej oraz sposoby analizy tego modelu.

Na podstawie otrzymanych wyników dla zastosowanego zestawu danych, możemy wyciągnąć wniosek, iż istnieje duża zależność pomiędzy zużyciem paliwa, a mocą pojazdu, jego wagą oraz stopniem sprężania. W mniejszym stopniu również wpływa na to pojemność skokowa silnika. Natomiast skok tłoka jako zmienna objaśniająca jest nieistotna, podobnie jak rozstaw osi, którego brak w modelu, po usunięciu wartości odstających nie pogarsza jego skuteczności.

Końcowy wzór na regresje liniową prezentuje się następująco:

Y = 44.644376 – 0.009254*x1 + 0.35943*x2 + 0.607823*x3 – 0.067752*x4

Gdzie:

x1 – waga pojazdu	x3 – stopień sprężania
x2 – pojemność silnika	x4 – moc silnika
Y – zużycie paliwa	

Biorą pod uwagę współczynnik R^2 możemy stwierdzić, że model dobrze odwzorowuje zmiany zużycia paliwa dla zastosowanych zmiennych niezależnych. Natomiast uwzględnić trzeba, iż część tych zmiennych jest w pewnym stopniu ze sobą skorelowana.

Podsumowując im większa waga pojazdu oraz moc jego silnika tym na mniej mil starczy jeden galon paliwa. Natomiast im większy stopień sprężania i pojemność silnika tym zużycie paliwa spada.

Wyciągnięte wnioski nie są zaskoczeniem, gdyż często samochody charakteryzujące się większymi osiągami silnika, zużywają więcej paliwa. Często też ta kwestia idzie w parze z masą własną pojazdu, dlatego te dwa parametry, są ze sobą skorelowane, tak samo jak pojemność silnika, która ma bezpośredni wpływ na ilości koni mechanicznych jednostki napędowej. Jednakże uwzględnienie tych parametrów jako zmienne niezależne ma pozytywny wpływ na działanie modelu regresji.

Nie jest również zdziwieniem, że rozstaw osi pojazdu nie wpływa w zauważalny sposób na zużycie paliwa. Ponadto uwzględnienie w modelu wagi pojazdu, która jest lepszym tego wyznacznikiem, może eliminować ją jako zmienną niezależną, ze względną na korelację pomiędzy nimi, jak miało to miejsce po eliminacji wartości odstających.