Politechnika Krakowska
Wydział Informatyki i Telekomunikacji

Sprawozdanie z przedmiotu:

Statystyka i Probabilistyka

Projekt nr 1
Temat:

Regresja Prosta

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

1. Cel projekty

Celem projektu jest zapoznanie się z zasadą działania regresji liniowej, a także poznanie podstawowych miar dopasowania modelu regresji i sposobu oceny otrzymanego modelu w oparciu o analizę reszt.

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

2. Zbiór danych

Badany zestaw danych dotyczą zużycia paliwa w cyklu miejskim w milach na galon. Dane stanowią statystyki zebrane w 1993 roku i są upublicznione z biblioteki StatLib, która jest utrzymywana na Carnegie Mellon University.

Łącznie zbiór danych liczy 398 pozycji i zawiera 8 kategorii, kolejno:
- mpg - zużycie paliwa w milach na galon,
- cylinders - liczba cylindrów silnika pojazdu,
- displacement - pojemność skokowa silnika,
- horsepower - ilość koni mechanicznych pojazdu
- weight - waga pojazdu,
- acceleration - przyśpieszenie pojazdu od 0 do 60 mil na godzinę,
- model year - rok produkcji,
- origin - miejsce wyprodukowania (1-Ameryka, 2-Europa, 3-Japonia),
- car_name - marka i model pojazdu

	A	B	C	D	E	F	G	H	I
1	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
2	18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
3	15	8	350	165	3693	11.5	70	1	buick skylark 320
4	18	8	318	150	3436	11	70	1	plymouth satellite
5	16	8	304	150	3433	12	70	1	amc rebel sst
6	17	8	302	140	3449	10.5	70	1	ford torino
7	15	8	429	198	4341	10	70	1	ford galaxie 500
8	14	8	454	220	4354	9	70	1	chevrolet impala
9	14	8	440	215	4312	8.5	70	1	plymouth fury iii
10	14	8	455	225	4425	10	70	1	pontiac catalina

Rysunek 1 Wycinek tabeli badanego zestawu danych

Głównym celem badanego zbioru jest zbadanie zależności pomiędzy zużyciem paliwa, a wartością przyśpieszenia pojazdu (od 0 do 60 mph). Należy sprawdzić czy model regresji liniowej 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. Model Regresji prostej i jego diagnostyka w programie Excel

Regresja liniowa obliczona dla wskazanych parametrów wskazuję, iż dane są od siebie zależne. Funkcja liniowa określana jest poniższym wzorem:

mpg = 4.96979 * 1.191204X
*gdzie X to przyśpieszenie pojazdu

Na podstawie tego wzoru można wywnioskować, że średnio dla pojazdu, którego przyspieszenie od 0 do 60 mil na godzinę jest większe o 1 sekundę, jego wartość zużycia paliwa wzrasta o ok. 1.191204 galon/mila.

Współczynnik **R kwadrat** wynosi 0.176642, natomiast **test istotności** tego współczynnika jest równy 1.82309E-18 W związku z tym można definitywnie odrzucić Hipotezę Zerową, iż „R kwadrat nie różni się istotnie od zera", biorąc pod uwagę wartości p równą 5%.

PODSUMOWANIE - WYJŚCIE

Statystyki regresji	
Wielokrotność R	0.420288912
R kwadrat	0.17664277
Dopasowany R kwadrat	0.174563585
Błąd standardowy	7.101097755
Obserwacje	398

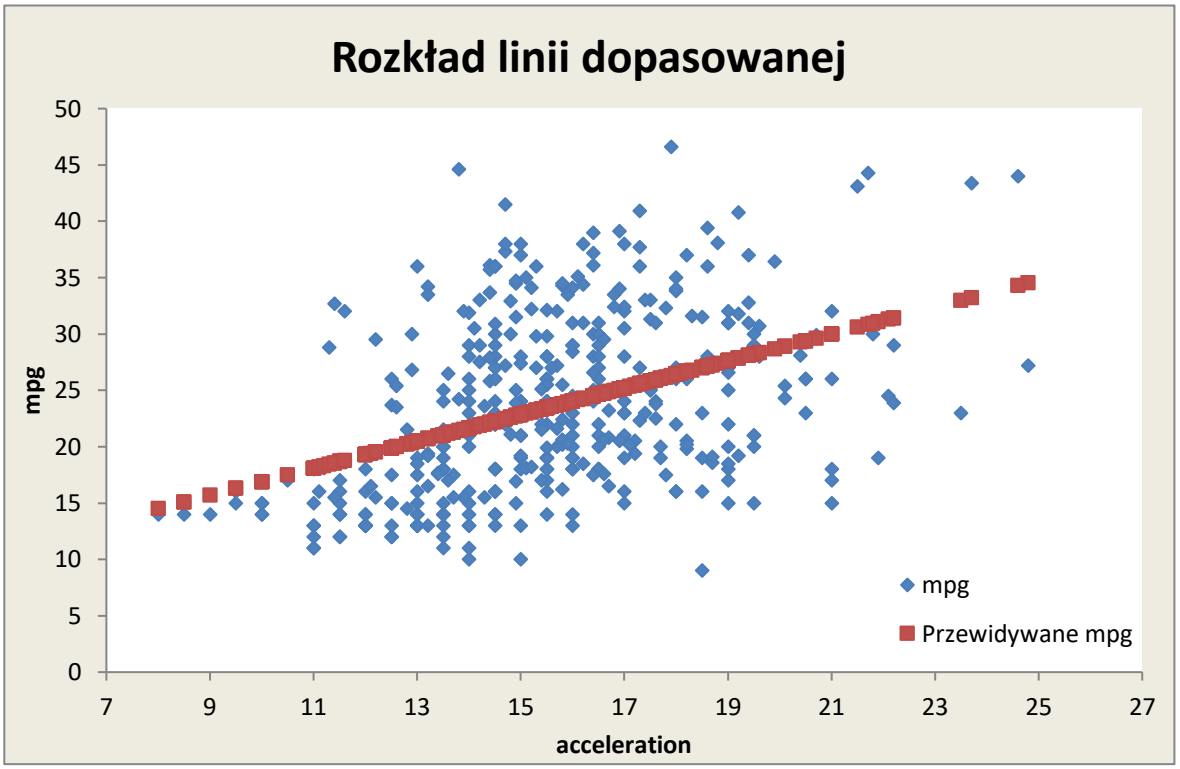
ANALIZA WARIANCJI

	df	*SS*	*MS*	*F*	*Istotność F*
Regresja	1	4284.042103	4284.0421	84.9577	1.82E-18
Resztkowy	396	19968.53337	50.425589		
Razem	397	24252.57548			

	Współczynniki	*Błąd standardowy*	*t Stat*	*Wartość-p*	*Dolne 95%*	*Górne 95%*	*Dolne 95.0%*	*Górne 95.0%*
Przecięcie	4.969793004	2.043207898	2.4323482	0.0154435	0.952902	8.9866838	0.95290224	8.9866838
acceleration	1.191204529	0.129236433	9.2172501	1.823E-18	0.937129	1.4452798	0.93712924	1.4452798

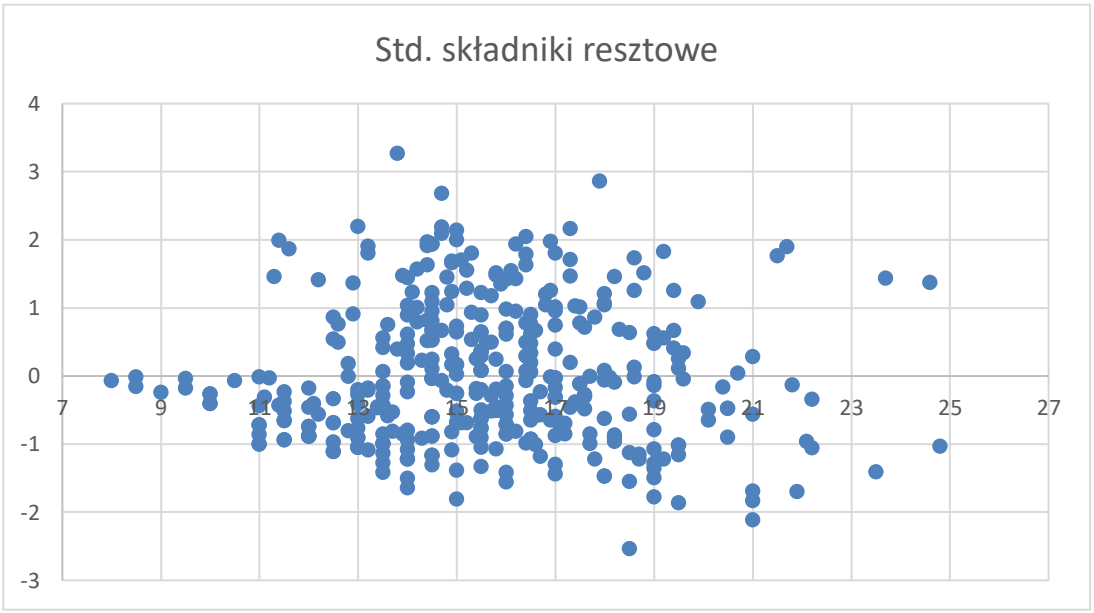
Rysunek 2 Statystyki regresji liniowej w programie Excel

Poniżej wykres obrazujący zależność zużycia paliwa i przyśpieszenia pojazdu wraz z rozkładem linii dopasowanej modelu regresji liniowej.



Rysunek 3 Wykres regresji liniowej w programie Excel

Natomiast rozkład standardowych składników resztowych względem zmiennej niezależnej obrazuje kolejny rysunek, dzięki któremu można dużo więcej stwierdzić o jakości modelu regresji liniowej.



Rysunek 4 Wykres standardowych składnik resztowych i zmiennej niezależnej w programie Excel

Biorąc pod uwagę diagnostykę reszt, należy w pierwszej kolejności dokonać **analizę normalności**. Analiza ta w tym przypadku jednak ma tylko wartość pomocniczą, gdyż ilość pozycji w danych testowych jest większa niż 30, a dokładniej jest to 398.

Jednakże na podstawie wyników z wykresu nr 4 możemy stwierdzić po wskazaniach skośności i kurtozy, że standardowe składniki resztowe, są rozkładem normalnym prawoskośnym i platokurtycznym.

Kolejnym etapem jest **analiza homoskedastyczności**, a więc stałości wariancji reszt. Tak więc na podstawie rys. 4 możemy stwierdzić brak wyraźnych zmian wariancji reszt funkcji zmiany zmiennej niezależnej.

Dodatkowo w ramach tego samego wykresu możemy dokonać **analizy autokorelacji**, a wiec poszukiwanie zależności funkcyjnej reszt od zmiennej niezależnej. Również i w tym wypadku możemy zdecydowanie stwierdzić, iż nie ma żadnych zależności funkcyjnych.

Kolejnym krokiem w ramach diagnostyki reszt jest tzw. **„Poszukiwanie Kink-Konga”**, a więc poszukiwanie dużych obserwacji, które przyciągają do siebie model. W tym celu należy się posłużyć analizą opisową, z której możemy odczytać, iż minimum standardowych reszt wynosi -2.539, natomiast maksimum tych reszt to 3.27. W związku z czym możemy stwierdzić występowanie obserwacji odstających, gdyż wyniki te nie mieszczą się w przedziale od -3 do 3.

acceleration		*Std. składniki resztowe*	
Średnia	15.56809045	Średnia	-3.36219E-15
Błąd standardowy	0.138230456	Błąd standardowy	0.050125471
Mediana	15.5	Mediana	-0.175159707
Tryb	14.5	Tryb	-0.883265093
Odchylenie standardov	2.75768893	Odchylenie standardowe	1
Wariancja próbki	7.604848234	Wariancja próbki	1
Kurtoza	0.419496883	Kurtoza	-0.31883459
Skośność	0.278776845	Skośność	0.499257308
Zakres	16.8	Zakres	5.809052124
Minimum	8	Minimum	-2.539015697
Maksimum	24.8	Maksimum	3.270036427
Suma	6196.1	Suma	-1.33815E-12
Licznik	398	Licznik	398

Rysunek 4 Statystyka opisowa zmiennej niezależnej oraz standardowych składników resztowych

4. Model Regresji prostej i jego diagnostyka przy użyciu języka R

Podobny sposób diagnostyki modelu jaki został przeprowadzony w programie Excel, możemy przeprowadzić z wykorzystaniem języka R. W ramach tych rozważań skorzystano z programu RStudio.

Na początku dane przy pomocy pakietu *gdata* zostały wczytane do konkretnych zmiennych programu i następnie przeprowadzono na nich konkretne operacje.

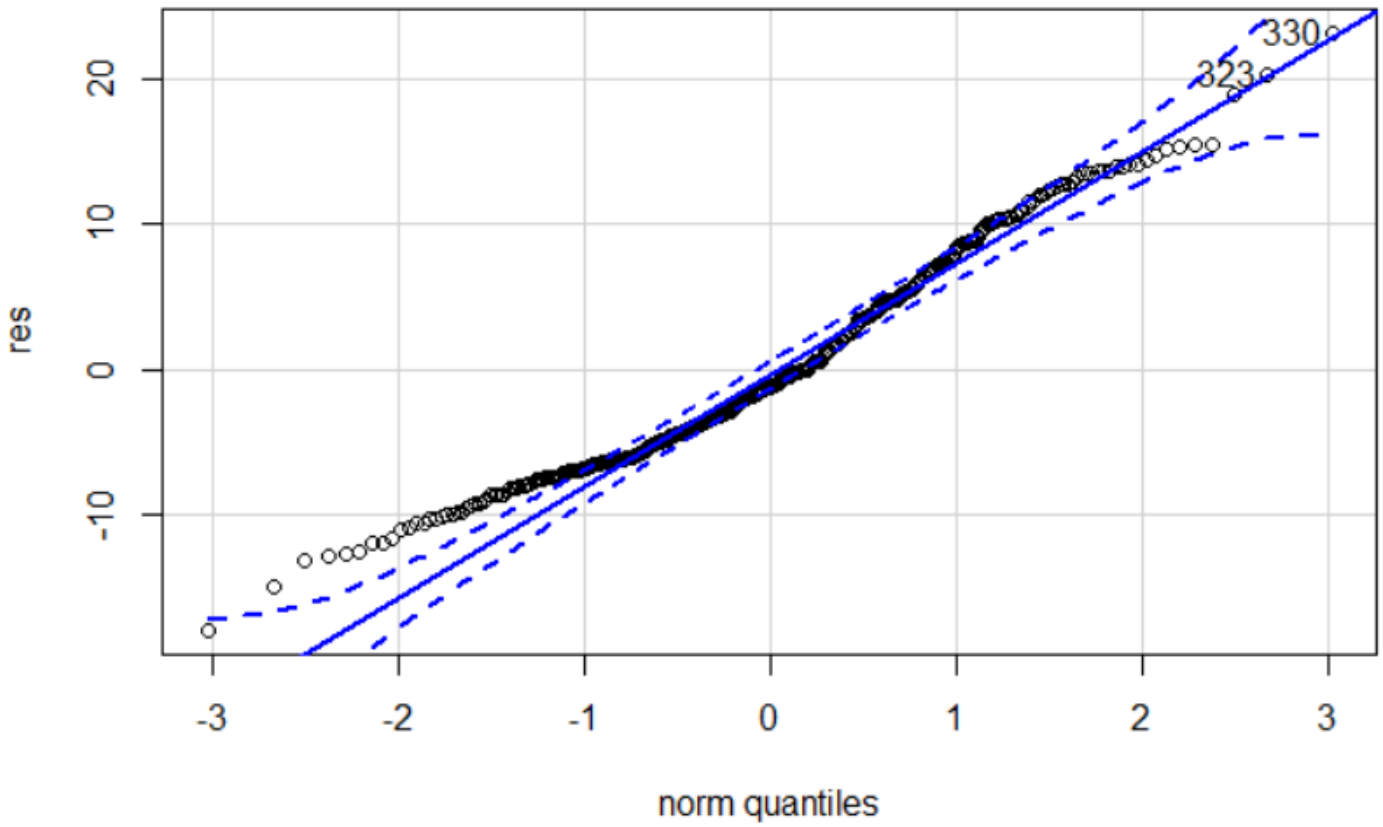
Poniżej znajdują się wynik podsumowujący regresję prostą, z której również możemy odczytać wartość współczynnika **R kwadrat** równa 0.1746, a także wartość współczynnika nachylenia prostej 1.29, oraz punktu przecięcia osi oy (mpg) 4.96, a więc wartości pokrywające się z tym co otrzymano przy pomocy programu Excel. Jedyną różnicą jest wynik **testu istotności** R kwadrat który ze względu na niską wartość został podany z przybliżeniem (p < 2.2e-16).

```
> lm1 <- lm(formula=mpg ~ acceleration)
> summary(lm1)

Call:
lm(formula = mpg ~ acceleration)

Residuals:
    Min      1Q  Median      3Q     Max
-18.007  -5.636  -1.242   4.758  23.192

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.9698     2.0432   2.432   0.0154 *
acceleration   1.1912     0.1292   9.217   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.101 on 396 degrees of freedom
Multiple R-squared:  0.1766,    Adjusted R-squared:  0.1746
F-statistic: 84.96 on 1 and 396 DF,  p-value: < 2.2e-16
```

Rysunek 5 Statystyki podsumowujące regresję prostą

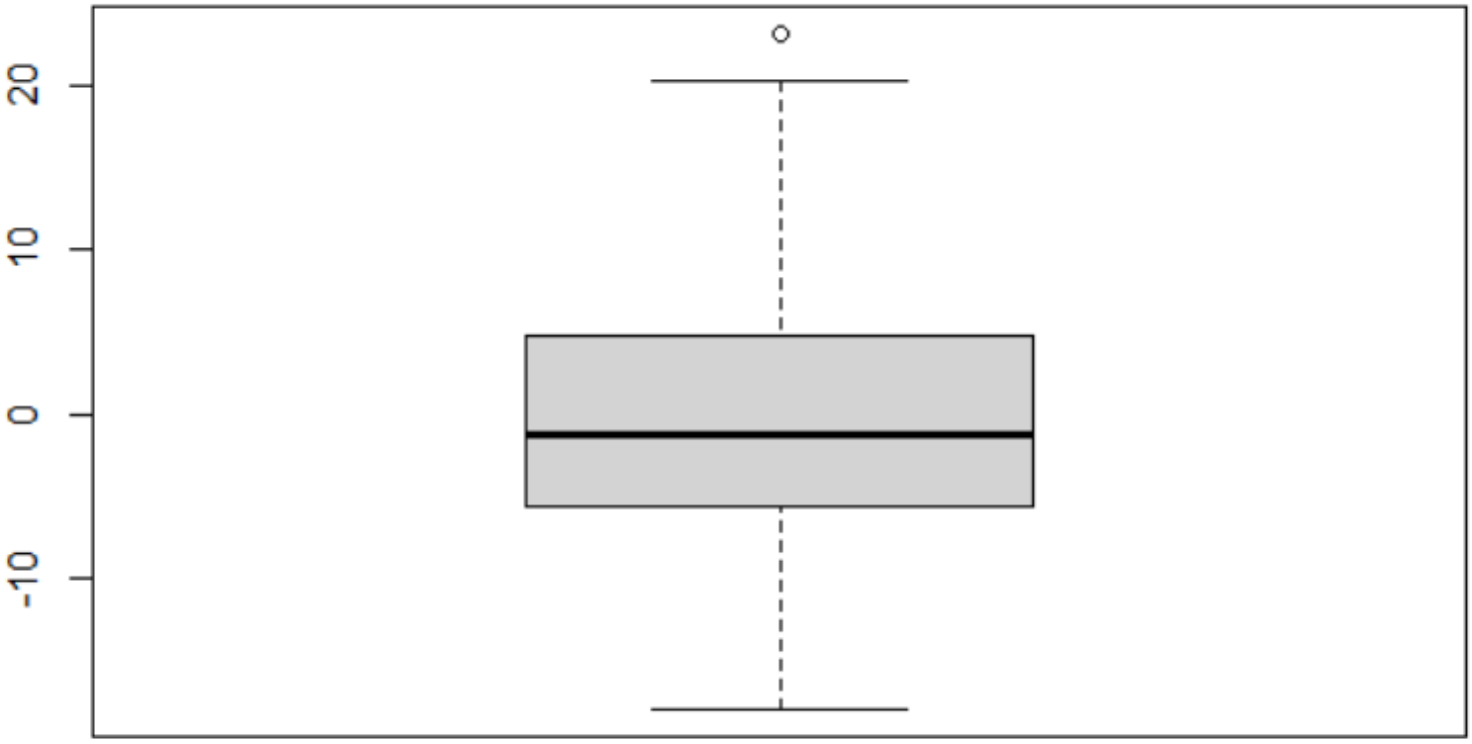
Kolejnym etap diagnostyki modelu jest analiza **wykresu kwantylowego** dla standardowych składników resztowych, przedstawiony na rysunku poniżej.

Można zauważyć między innymi dwa punkty jako obserwacje podejrzane, tj. punkt 323 i 330, a więc punkty świadczące o byciu obserwacjami odstającymi. Ponadto na wykresie qqPlot , możemy zaobserwować, że dla wartości najniższych wychodzą poza zakres tolerancji.



Rysunek 6 Wykres kwantylowy standardowych składników resztowych

Odnośnie wcześniej wspomnianych wartości podejrzanych o bycie odstającymi program R udostępnia możliwość graficznego przedstawienia składników resztowych w formie **wykresu pudełkowego** przy pomocy polecenia boxPlot, z którego możemy odczytać, iż jest jedna obserwacja powyżej górnego płotka. Wykres pudełkowy obrazuje rysunek nr 7.



Rysunek 7 Wykres pudełkowy składników resztowych

Następnym etapem diagnostycznym jest **test 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%.

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.97087, p-value = 3.873e-07
```

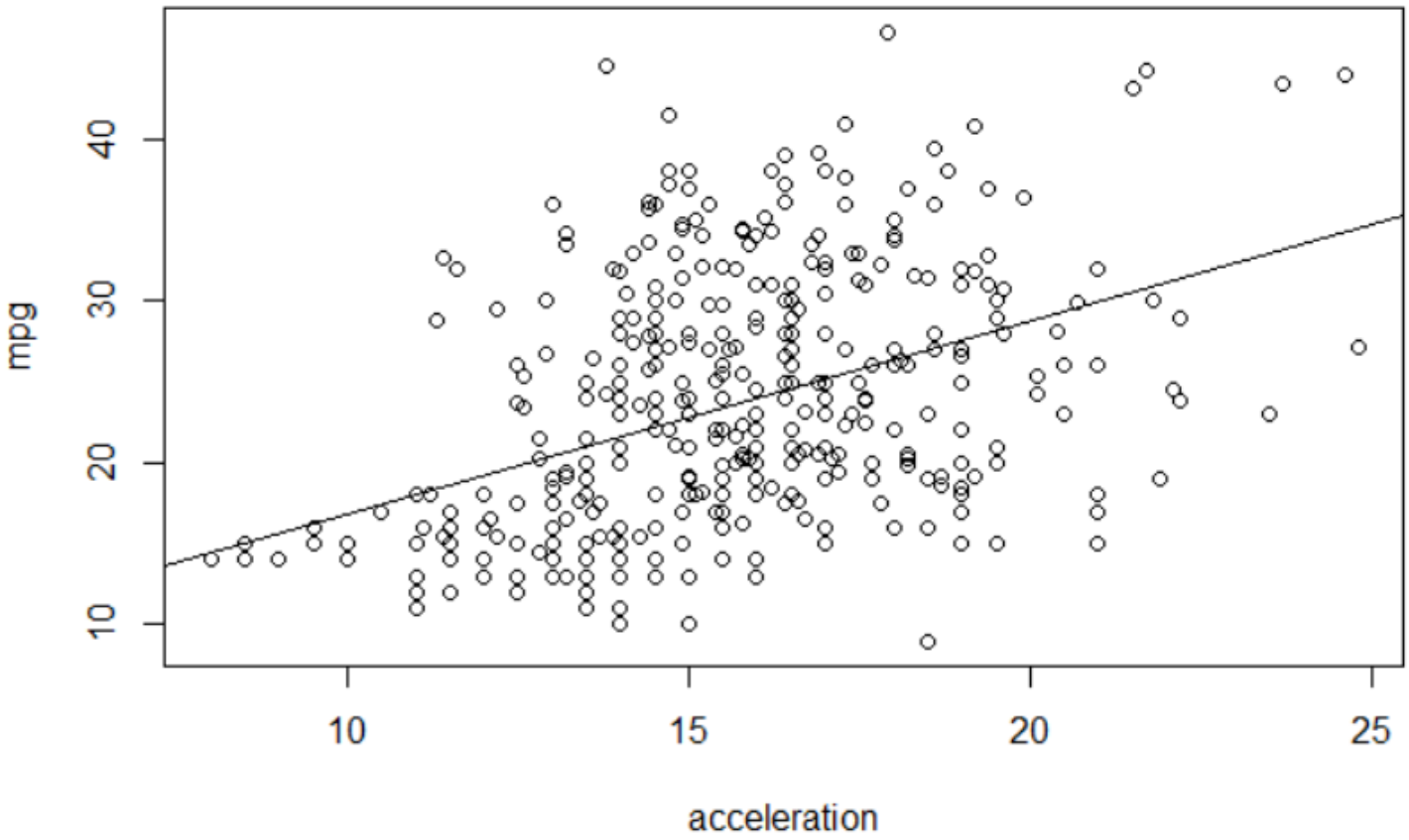
Rysunek 8: Wynik testu Shapiro-Wilka dla składników resztowych

Kolejnym testem, który możemy wykonać przy pomocy języka R jest **test na autokorelację**. 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 zdecydowanie można odrzucić Hipotezę Zerową świadczącą o „braku autokorelacji standardowych składników resztowych", gdyż dla każdego podanego rzędu autokorelacji wartość p jest znacznie mniejsza niż 5%.

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

        Box-Pierce test

data:  res
X-squared = 173.99, df = 1, p-value < 2.2e-16

> Box.test(res,lag=2)

        Box-Pierce test

data:  res
X-squared = 291.91, df = 2, p-value < 2.2e-16

> Box.test(res,lag=3)

        Box-Pierce test

data:  res
X-squared = 378.71, df = 3, p-value < 2.2e-16

> Box.test(res,lag=4)

        Box-Pierce test

data:  res
X-squared = 436.01, df = 4, p-value < 2.2e-16
```

Rysunek 9: Wynik testów autokorelacji reszt Boxa

Następnym etapem jest zobrazowanie zależności graficznej pomiędzy zmienną zależna i niezależna wraz z rozkładem linii dopasowania, rys nr 10, który to podobnie jak miało to miejsce w programie Excel obrazuje **model regresji prostej**.



Rysunek 10: Wykres modelu regresji w programie RStudio

Ostatnim etap jest **identyfikacji obserwacji odstających**, który to bierze pod uwagę punkt zobrazowany wcześniej na wykresie qqPlot i boxPlot o identyfikatorze 330.
Na podstawie tego testu możemy stwierdzić, że biorąc pod uwagę największą studentyzowaną resztę, która nie mieści się w zakresie w optymalnym ($3.31 \notin \langle -3,3 \rangle$), obserwacja ta jest obserwacją odstającą.

```
> outlierTest(lm1)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
330  3.31267          0.0010093       0.4017
>
```

Rysunek 11: Test Bonferroniego dla identyfikacji obserwacji odstających

5. Wnioski

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

Na podstawie otrzymanych wyników dla zastosowanego zestawu danych, możemy wyciągnąć wniosek, iż istnieje zależność liniowa pomiędzy zużyciem paliwa, a przyspieszeniem pojazdu. Biorą pod uwagę współczynnik R kwadrat możemy stwierdzić, że model nie jest jednak doskonały i słabo jest przystosowany do danych, z pośród których jedna obserwacja jest odstająca (Honda Civic 1500 gl – spalanie 44.6 mpg dla przyśpieszenia 13.8 mph).

Podsumowując zużycie paliwa jest mniejsze w pojazdach których wartość przyśpieszenia (od 0 do 60 mph) jest większa, a więc dla tych, które maja dłuższy czas osiągnięcia optymalnej prędkości. Ma to związek z tym, iż pojazdy które osiągają lepsze prędkości, a wiec także badane w tym projekcie przyśpieszenie, mają zarazem większą moc (więcej koni mechanicznych), co ma wpływ na większe zużycie paliwa, tak więc jeden galon starczy na mniejszą ilość mil.