Sylwester Dylik
Nr indeksu: 145820

Regresja liniowa

Do wykonania regresji liniowej wykorzystałem język R (Rstudio) oraz własną bazę danych. W projekcie postanowiłem przedstawić regresję liniową ceny paliwa, która miała swoje wzloty jak i upadki w czasach pandemii. Ceny paliw zostały pobrane jako średnie miesięczne ze strony autocentrum.pl, ceny dolara z archiwum giełdy zaś liczbę zarejestrowanych pojazdów pobrałem z archiwum CEPiK.

	A	B	C	D	E	F
1	Data	Pb95	Import	LPG	Car	USD
2	01.01.2019	4,74	1,72	2,22	145532	3,76
3	01.02.2019	4,69	1,64	2,14	150682	3,8
4	01.03.2019	4,8	1,83	2,13	180977	3,8
5	01.04.2019	5,12	1,77	2,16	185272	3,81
6	01.05.2019	5,23	1,65	2,2	177250	3,84
7	01.06.2019	5,19	1,61	2,12	161535	3,77
8	01.07.2019	5,06	1,54	2,04	183705	3,79
9	01.08.2019	5,02	1,57	1,98	165607	3,9
10	01.09.2019	4,9	1,65	1,94	146196	3,95
11	01.10.2019	4,92	1,73	1,95	171689	3,89
12	01.11.2019	4,92	1,83	2,08	145629	3,88
13	01.12.2019	4,96	1,94	2,34	154115	3,84
14	01.01.2020	4,92	2,02	2,33	147655	3,83
15	01.02.2020	4,84	2,1	2,19	149826	3,92
16	01.03.2020	4,45	1,78	1,96	106009	4,02
17	01.04.2020	3,96	1,65	1,72	67159	4,18
18	01.05.2020	3,9	1,88	1,6	111529	4,15
19	01.06.2020	4,12	1,73	1,87	149599	3,94
20	01.07.2020	4,27	1,8	1,94	173569	3,88
21	01.08.2020	4,39	1,83	1,92	144337	3,72
22	01.09.2020	4,42	1,88	1,95	160049	3,79
23	01.10.2020	4,37	1,74	1,99	152448	3,86
24	01.11.2020	3,37	1,69	2,07	133626	3,8
25	01.12.2020	4,45	1,72	2,08	160577	3,68

Data – data określająca miesiąc pobrania danych
Pb95 – cena benzyny – zmienna objaśniana
Import – import ropy do polski w (mln ton) – zmienna objaśniająca
LPG – cena LPG – zmienna objaśniająca
Car – liczba nowo zarejestrowanych samochodów – zmienna objaśniająca
USD – kurs dolara w zł – zmienna objaśniająca

Arkusz kalkulacyjny z bazą potrzebnych nam danych zapisałem w formacie .xlsx oraz umieściłem w folderze zawierającym projekt regresji liniowej.

Na początku należy zainstalować bibliotekę pozwalającą nam na wczytanie naszej bazy .

```
R 4.1.0 · C:/Users/sdyli/Desk
> library(readxl)
```

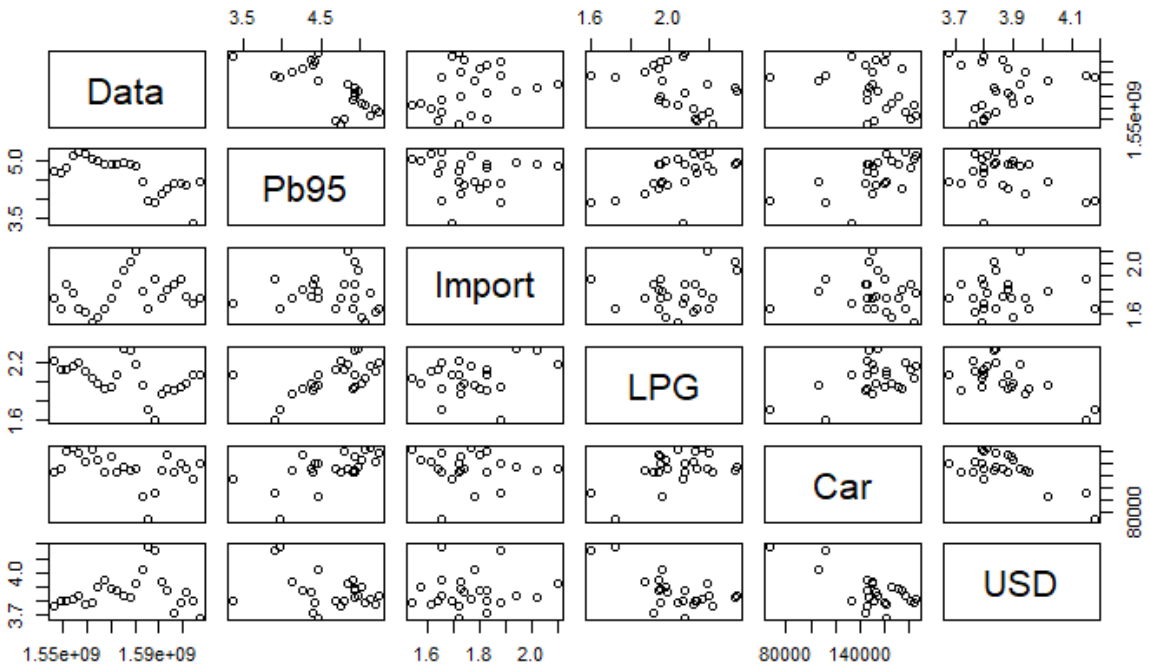
Następnie wczytujemy naszą bibliotekę zaznaczając lokalizację danych tylko na arkuszu 1 oraz wyświetlamy załadowane dane.

```
> sprzedaz <- read_excel("sprzedaz.xlsx", sheet = "Arkusz1")
> View(sprzedaz)
```

Sprawdzam poprawność załadowanych tabel czy poprawnie widzę nazwy kolumn w programie.

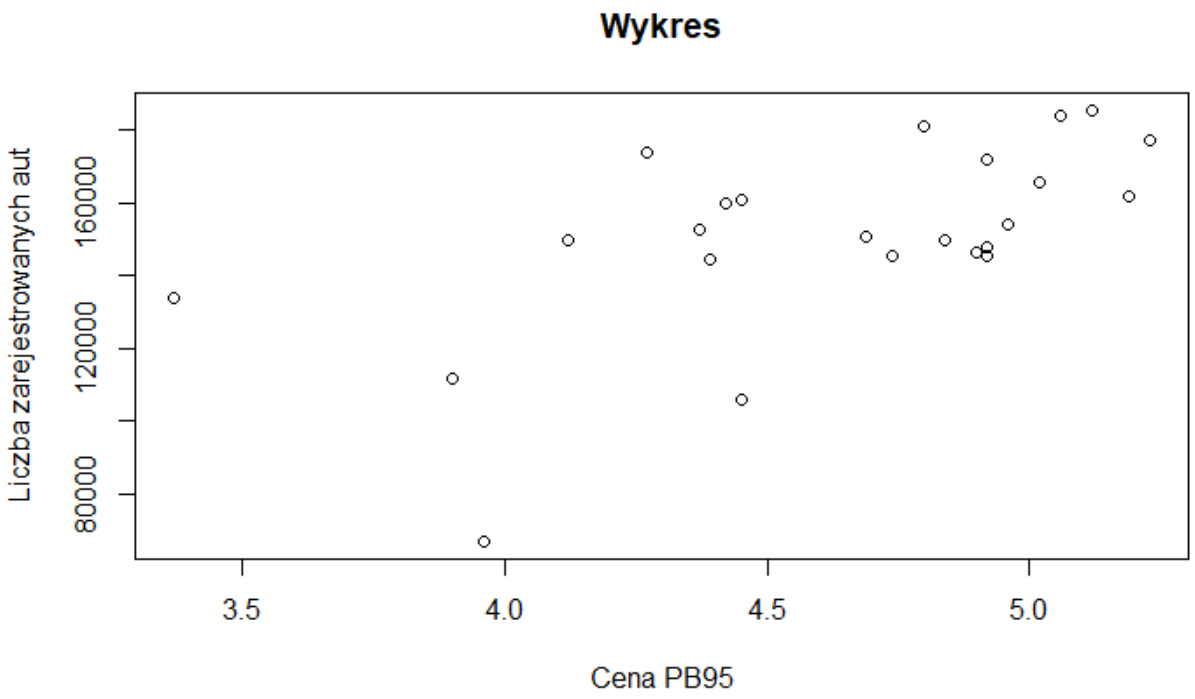
Sprawdzam współzależność danych od sibie:

```
> names(sprzedaz)
[1] "Data"   "Pb95"   "Import" "LPG"    "Car"    "USD"
> |
```




```
> cor(sprzedaz$Pb95,sprzedaz$Import)
[1] -0.05694634
> cor(sprzedaz$Pb95,sprzedaz$LPG)
[1] 0.5845759
> cor(sprzedaz$Pb95,sprzedaz$Car)
[1] 0.5997178
> cor(sprzedaz$Pb95,sprzedaz$USD)
[1] -0.3419371
```

Możemy zauważyć że największą współzależność naszej zmiennej objaśnianej PB95 ma zmienna Car. Wykres względem tych zmiennych:




```
> plot(sprzedaz$Pb95,sprzedaz$Car,main = "Wykres",xlab = "Cena PB95",ylab = "Liczba zarejestrowanych au
t")
```

Kolejnym krokiem naszej regresji liniowej jest wyznaczenie parametrów naszej linii jak i określenie wielkości błędu, w języku r możemy to bardzo prosto otrzymać:

```
> model_reg<-lm(sprzedaz$Pb95~sprzedaz$Car)
> summary(model_reg)

Call:
lm(formula = sprzedaz$Pb95 ~ sprzedaz$Car)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0735 -0.2716  0.1146  0.2971  0.4547

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.046e+00  4.559e-01   6.681 1.02e-06 ***
sprzedaz$Car 1.046e-05  2.975e-06   3.515  0.00195 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3802 on 22 degrees of freedom
Multiple R-squared:  0.3597,    Adjusted R-squared:  0.3306
F-statistic: 12.36 on 1 and 22 DF,  p-value: 0.001951
```

Przy wyznaczaniu parametrów naszej regresji wkradł się błąd w zapisie dotyczącym wyświetlenie nam parametrów, niestety zjadłem jedną literę.

Opis parametrów:
1) Median - Mediana
2) Estimate – Odchylenie standardowe
3) Std. Error – Oszacowanie błędu

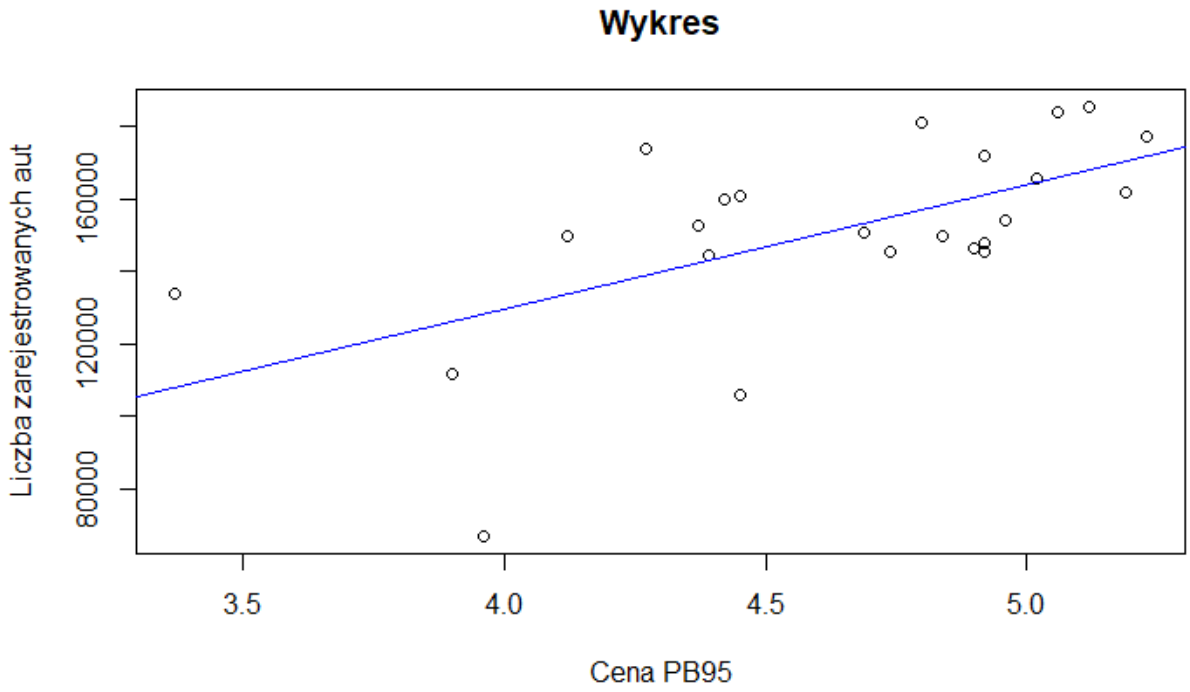
4) t value – sprawdza czy współczynnik jest różny od 0
5) Pr(>|t|) – badanie istotności modelu jako całości
6) Residual standard error – resztkowy błąd standardowy
7)Multiple R-squared – współczynnik
8)Adjusted R-squared – uwzględnienie liczby zmiennych w modelu
9)F-statistic – parametr mówiący nam czy regresja ma sens

Obliczamy przedział ufności :

```
> confint(model_reg)
                    2.5 %       97.5 %
(Intercept)  2.100556e+00 3.991646e+00
sprzedaz$Car 4.287864e-06 1.662693e-05
```

Teraz rysujemy naszą regresję liniową:

```
> model_reg<-lm(sprzedaz$Car~sprzedaz$Pb95)
> abline(model_reg,col = "blue" )
```



Analiza tabeli odchyleń:

```
> anova(model_reg)
Analysis of Variance Table

Response: sprzedaz$Car
              Df     Sum Sq    Mean Sq F value   Pr(>F)
sprzedaz$Pb95  1 5.8744e+09 5874414167  12.357 0.001951 **
Residuals     22 1.0459e+10  475398255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```