R: Eksploracja danych i regresja liniowa

2018-04-13

JDSZ1

Monika Kucal

Plan analizy

- Eksploracja zbioru mtcars
- Wizualizacja danych
  - Poszukiwanie zależności - wykresy i korelacje
  - Rozkłady zmiennych - skośność
- Estymacja modeli regresji liniowej
  - Modele z jedną zmienną objaśniającą
  - Modele z dwiema zmiennymi objaśniającymi
  - Wybór modeli na podstawie istotności zmiennych objaśniających i współczynnika determinacji R^2
  - Prezentacja graficzna i interpretacja
  - Zastosowanie metody gradientu prostego
- Analiza wykonana w R z wykorzystaniem bibliotek: tidyverse, corrplot, e1071, plot3D

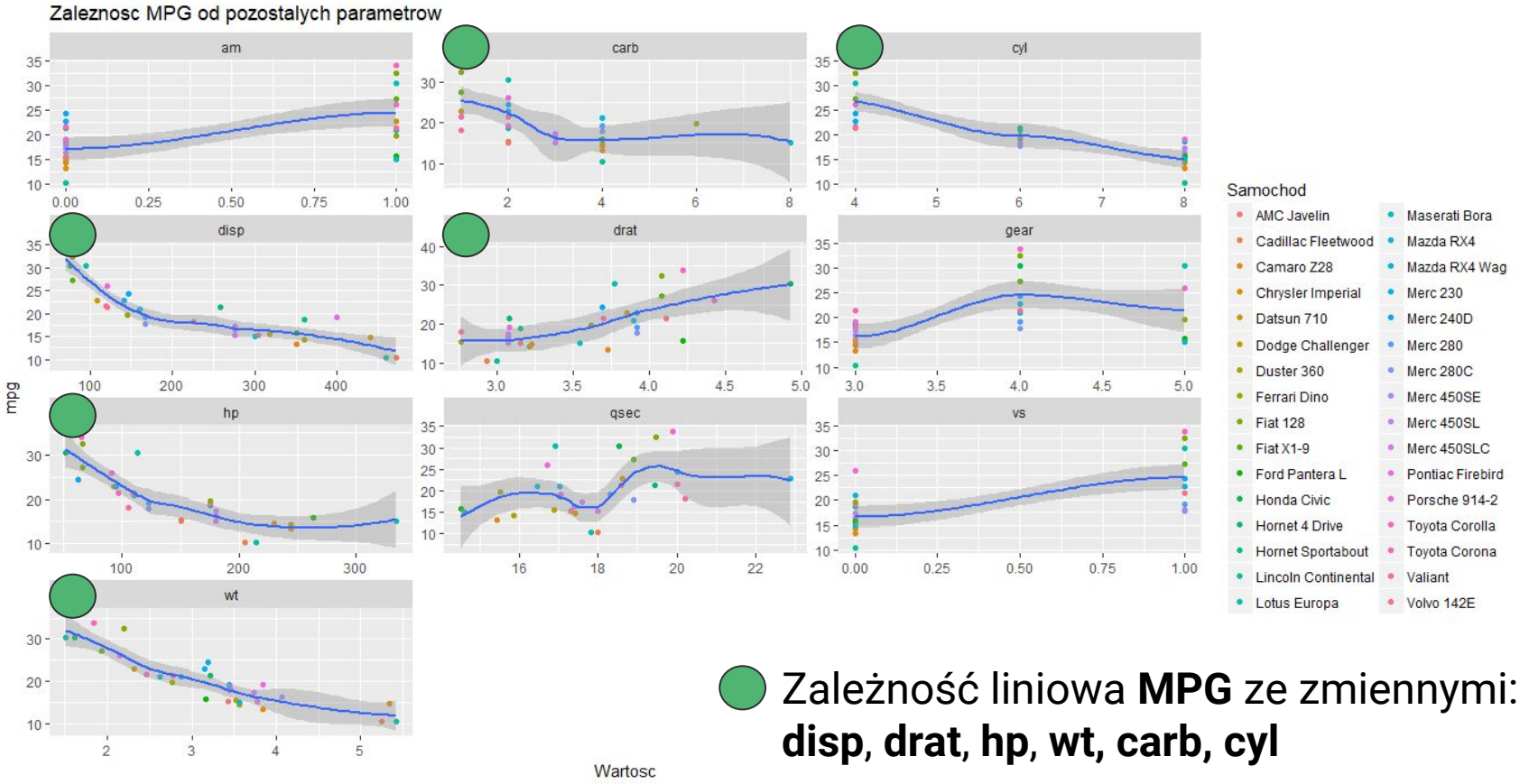
Eksploracja zbioru mtcars

- 11 zmiennych:
  - **mpg** - Spalanie paliwa
  - **cyl** - Liczba cylindrów
  - **disp** - Objętość skokowa cylindra
  - **hp** - Liczba koni mechanicznych
  - **drat -** Przełożenie osi tylnej
  - **wt** - Masa
  - **qsec** - Czas przejazdu ¼ mili
  - **vs** - Typ silnika V/S
  - **am** - Automatyczna/Manualna skrzynia biegów
  - **gear** - Liczba biegów
  - **carb** - Liczba gaźników

```
      mpg             cyl             disp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8  
 Median :19.20   Median :6.000   Median :196.3  
 Mean   :20.09   Mean   :6.188   Mean   :230.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0  
       hp             drat             wt       
 Min.   : 52.0   Min.   :2.760   Min.   :1.513  
 1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581  
 Median :123.0   Median :3.695   Median :3.325  
 Mean   :146.7   Mean   :3.597   Mean   :3.217  
 3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610  
 Max.   :335.0   Max.   :4.930   Max.   :5.424  
      qsec             vs              am       
 Min.   :14.50   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :17.71   Median :0.0000   Median :0.0000  
 Mean   :17.85   Mean   :0.4375   Mean   :0.4062  
 3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :22.90   Max.   :1.0000   Max.   :1.0000  
      gear            carb      
 Min.   :3.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:2.000  
 Median :4.000   Median :2.000  
 Mean   :3.688   Mean   :2.812  
 3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :8.000  
```

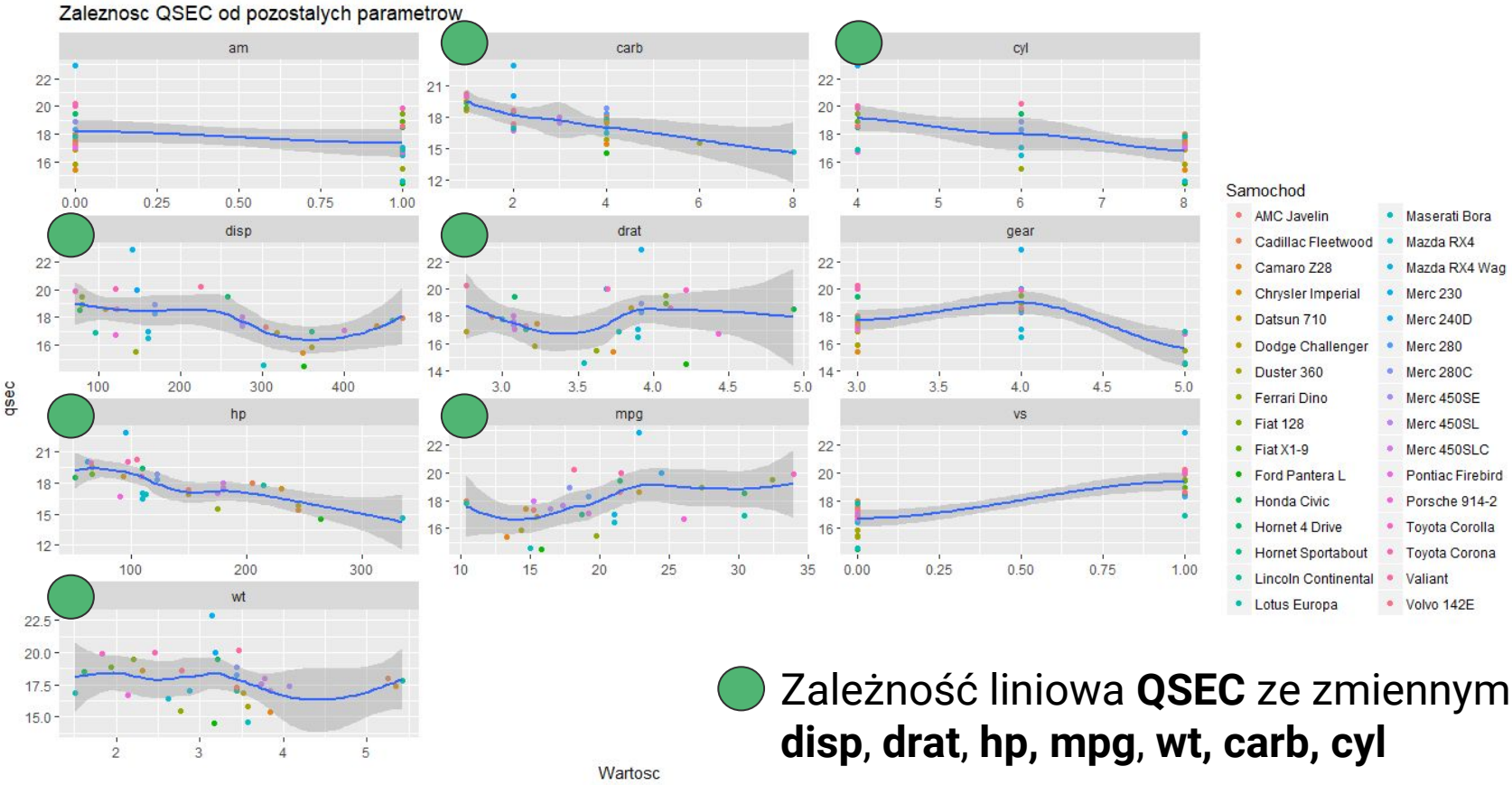
- Zmienne objaśniające, które mogą zależeć od pozostałych parametrów
  - **mtg** - Spalanie samochodu,
  - **qsec** - Czas przejazdu ¼ mili

Wizualizacja zmiennych - MPG vs. zmienne objaśniające



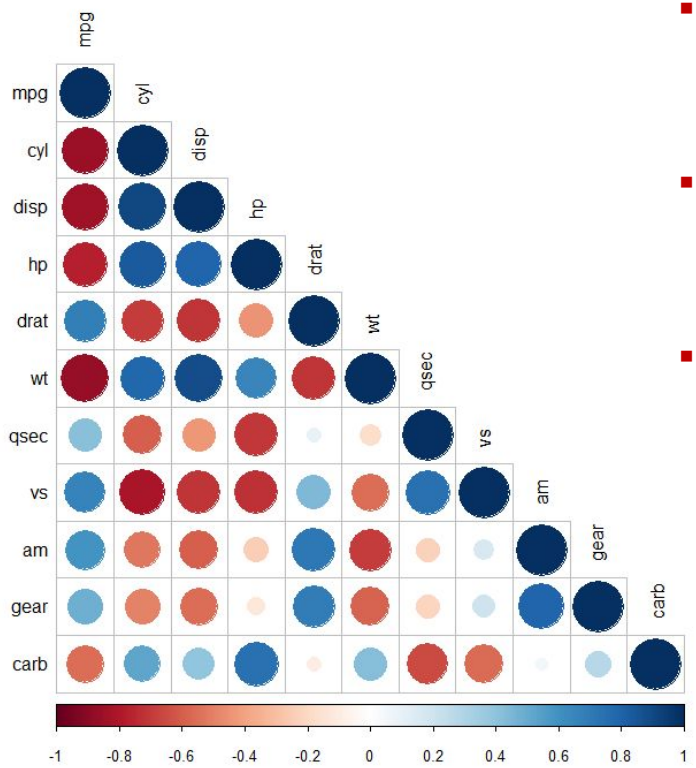
- Zależność liniowa **MPG** ze zmiennymi: **disp**, **drat**, **hp**, **wt**, **carb**, **cyl**

Wizualizacja zmiennych - QSEC vs. zmienne objaśniające



Zależność liniowa **QSEC** ze zmiennymi:
disp, **drat**, **hp**, **mpg**, **wt**, **carb**, **cyl**

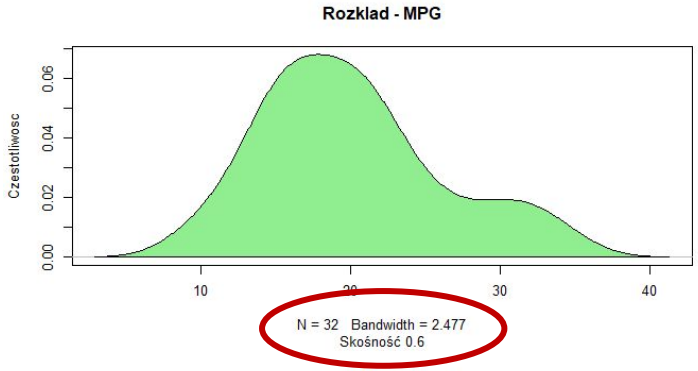
Wizualizacja zmiennych - Korelacje

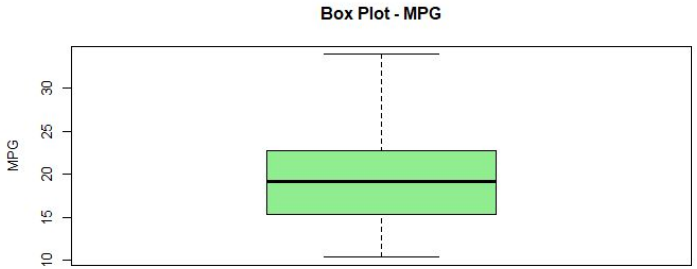


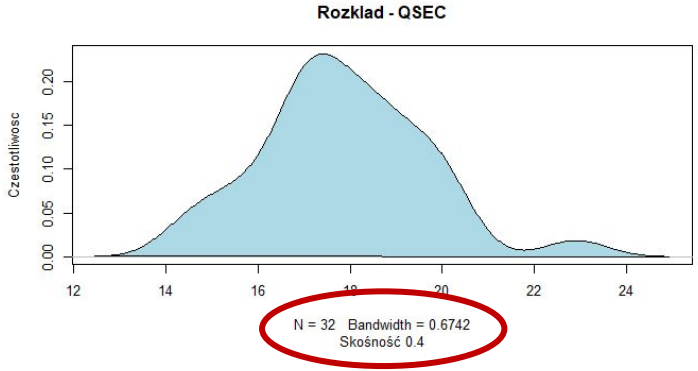
- Korelacje ze zmienną objaśnianą **mpg**:
  - silna korelacja ujemna: **wt**, **cyl**, **disp**, **hp**
  - silna korelacja dodatnia: **drat**, **vs**
- Korelacje ze zmienną objaśnianą **qsec**:
  - silna korelacja ujemna: **hp, carb**
  - silna korelacja dodatnia: **vs**
- W przypadku modelu z wieloma zmiennymi objaśniającymi silnie skorelowane zmienne nie powinny być wybierane jednocześnie jako zmienne objaśniające.

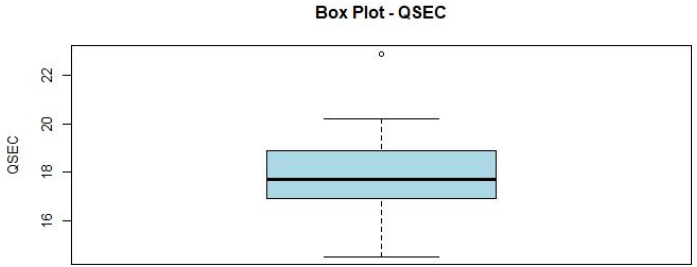
Korelacja ujemna - wraz ze wzrostem wartości zmiennej objaśniającej maleje wartość zmiennej objaśnianej.
Korelacja dodatnia - wraz ze wzrostem wartości zmiennej objaśniającej rośnie wartość zmiennej objaśnianej.

Wizualizacja zmiennych - Rozkłady





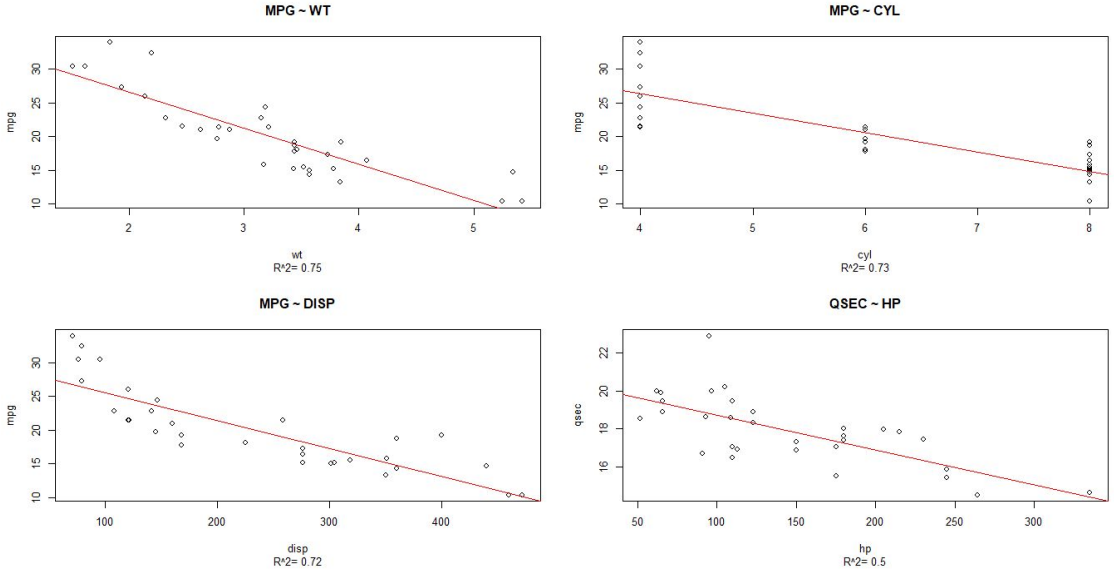




Interpretacja

Rozkłady obu zmiennych są prawoskośne (Skośność>0).

Modele regresji liniowej - jedna zmienna objaśniająca



Zaprezentowano modele, których współczynnik determinacji R^2 jest najwyższy - model wyjaśnił najwięcej zmienności oraz zmienne objaśniające są istotne.

Interpretacja

Wzrost masy samochodu WT, wzrost liczby cylindrów CYL, wzrost objętości skokowej cylindra DISP powodują spadek spalania MPG.

Wzrost liczby koni mechanicznych HP powoduje spadek czasu przejazdu ¼ mili QSEC.

Predykcja

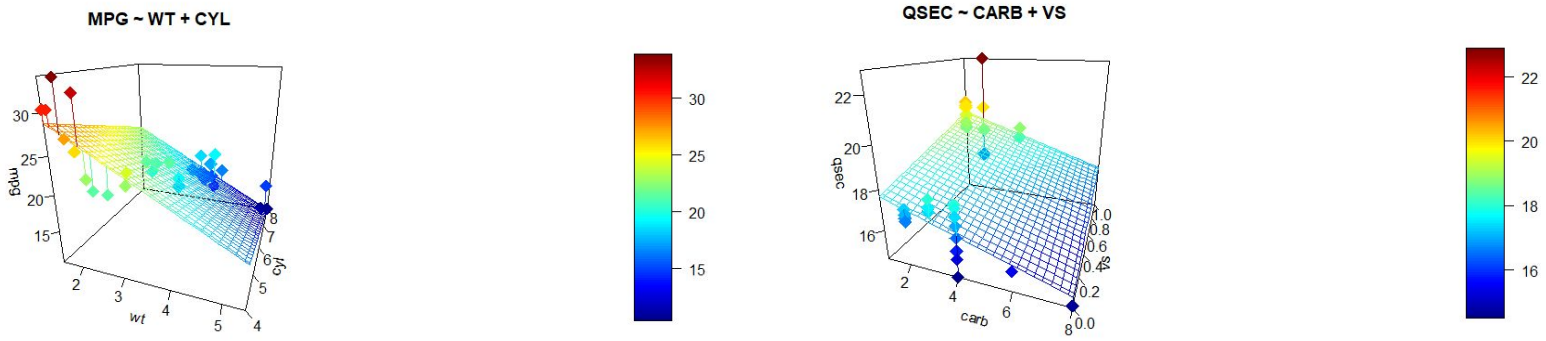
Jeśli masa samochodu WT=2.5, to spalanie MPG=23.92.

Jeśli liczba cylindrów CYL=8, to spalanie MPG=14.88.

Jeśli objętość skokowa cylindra DISP=350, to spalanie MPG=15.17.

Jeśli liczba koni mechanicznych HP=300, to czas przejazdu ¼ mili QSEC=15.02.

Modele regresji liniowej - dwie zmienne objaśniające



R^2=0.83

R^2=0.63

Zaprezentowano modele, których współczynnik determinacji R^2 jest najwyższy - model wyjaśnił najwięcej zmienności oraz wszystkie zmienne są istotne.

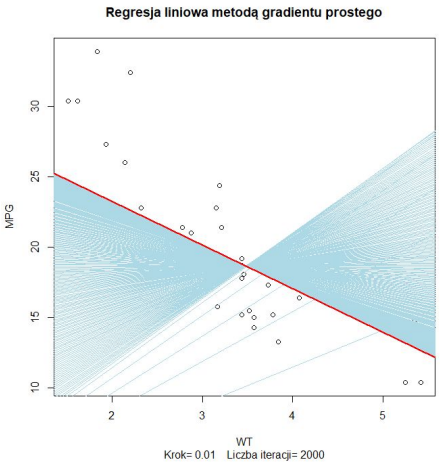
Interpretacja

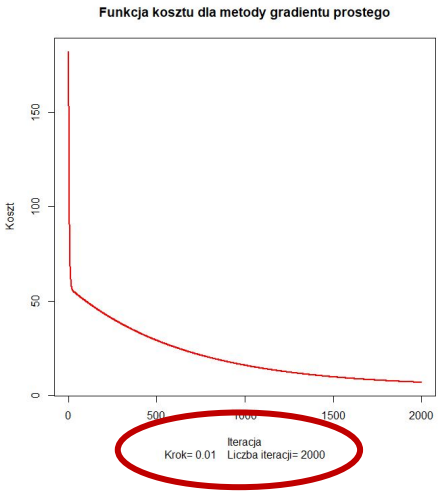
Im większa masa samochodu WT i liczba cylindrów CYL tym mniejsze spalanie MPG.
Im większa liczba gaźników CARB i typ silnika V, tym mniejszy czas przejazdu ¼ mili QSEC.

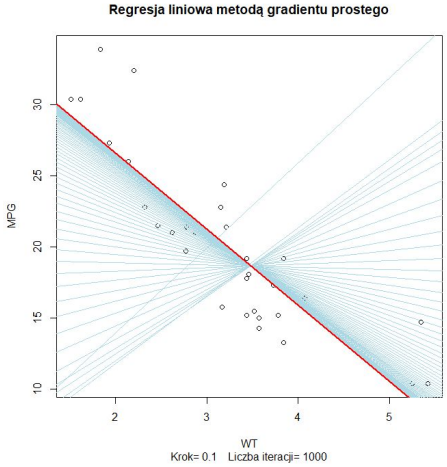
Predykcja

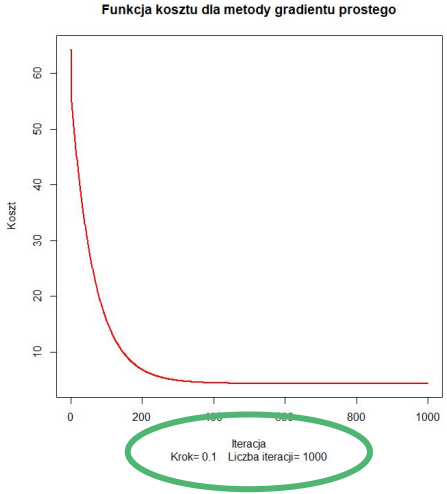
Jeśli masa samochodu WT=2.5 i liczba cylindrów CYL=8, to spalanie MPG=19.65.
Jeśli liczba gaźników CARB=2 i typ silnika V, to czas przejazdu ¼ mili QSEC=17.31.

Modele regresji liniowej - metoda gradientu prostego









Zaprezentowano zastosowanie metody gradientu prostego dla wybranego modelu regresji liniowej MPG ~ WT.

- Na podstawie funkcji kosztu dla **2000 iteracji z krokiem 0.01** stwierdzono, że otrzymana funkcja liniowa nie jest optymalna, ponieważ wykres funkcji kosztu nie wypłaszczył się całkowicie. W kolejnych etapach należy zwiększyć liczbę iteracji lub/i zwiększyć krok.
- Następnie analizowano wykresy funkcji kosztu dla zmieniających się parametrów (liczba iteracji i krok).
- Ostatecznie dla **1000 iteracji z krokiem 0.1** otrzymano zadowalający kształt funkcji kosztu i dobre dopasowanie funkcji regresji do danych.