Sprawozdanie 3

Jakub Markowiak
album 255705

3 maja 2021

Spis treści

1 Krótki opis zagadnienia

W tym sprawozdaniu zajmiemy się badaniem własności modeli regresji liniowej. Najpierw przygotujemy prosty model regresji liniowej dla odpowiednich danych i spróbujemy wykorzystać go do prognozowania przyszłych wartości. Następnie dla danych `Cars.93` sporządzimy modele regresji liniowej dla jednej, dwóch i trzech zmiennych objaśniających, a następnie spróbujemy wyłonić ten najlepiej dopasowany.

2 Opis eksperymentów/analiz

Przeprowadzimy następujące analizy i eksperymenty:

1. Badanie własności prostego modelu regresji liniowej,
2. Porównanie modeli regresji liniowej dla danych `Cars.93`.

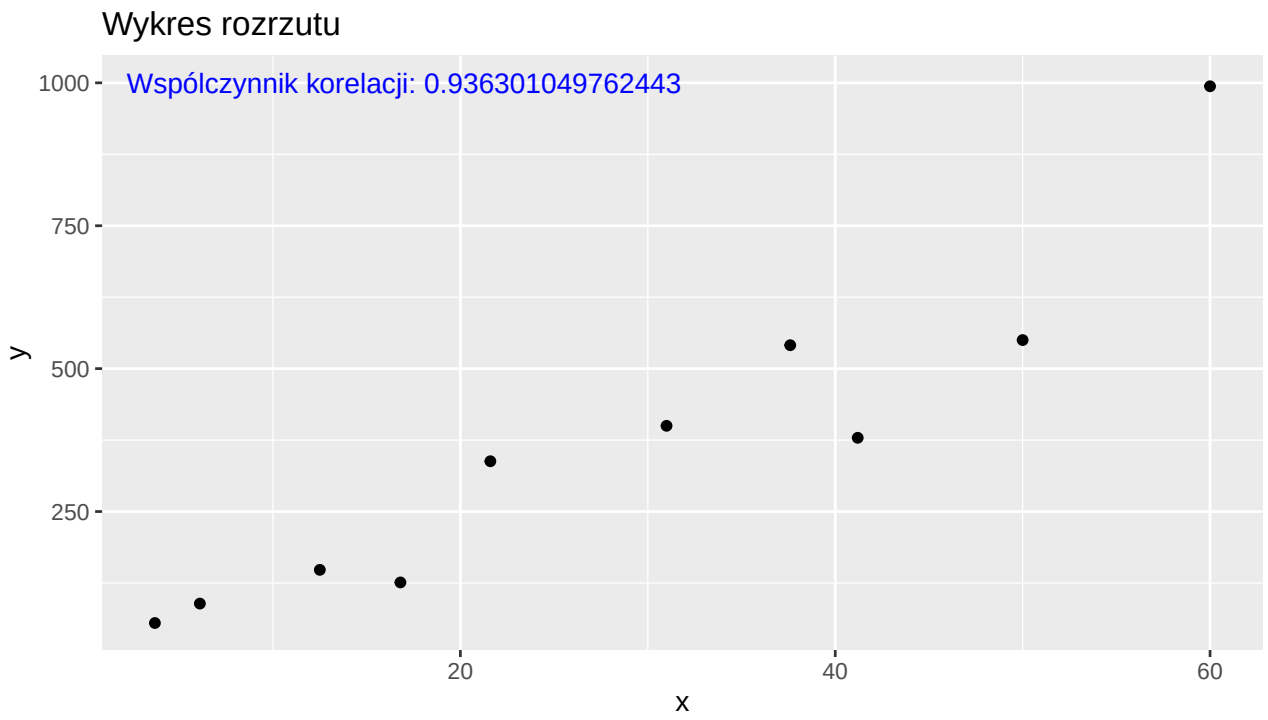
3 Wyniki

3.1 Badanie własności prostego modelu regresji liniowej

Rozpoczynamy od wczytania zadanych danych. (x - wydatki na reklamę, y - wielkość sprzedaży), a następnie rysujemy wykres rozrzutu w celu scharakteryzowania zależności między wydatkami na reklamę i wielkością sprzedaży.

	1	2	3	4	5	6	7	8	9	10
x	12.50	3.70	21.60	60.00	37.60	6.10	16.80	41.20	50.00	31.00
y	148.00	55.00	338.00	994.00	541.00	89.00	126.00	379.00	550.00	400.00

Tabela 1: Wczytane dane



Rysunek 1: Wykres rozrzutu dla wczytanych danych

Ponieważ współczynnik korelacji jest bardzo blisko 1, a także na podstawie rozmieszczenia punktów na wykresie rozrzutu, możemy wyciągnąć wniosek, że zależność między tymi zmiennymi ma charakter liniowy.

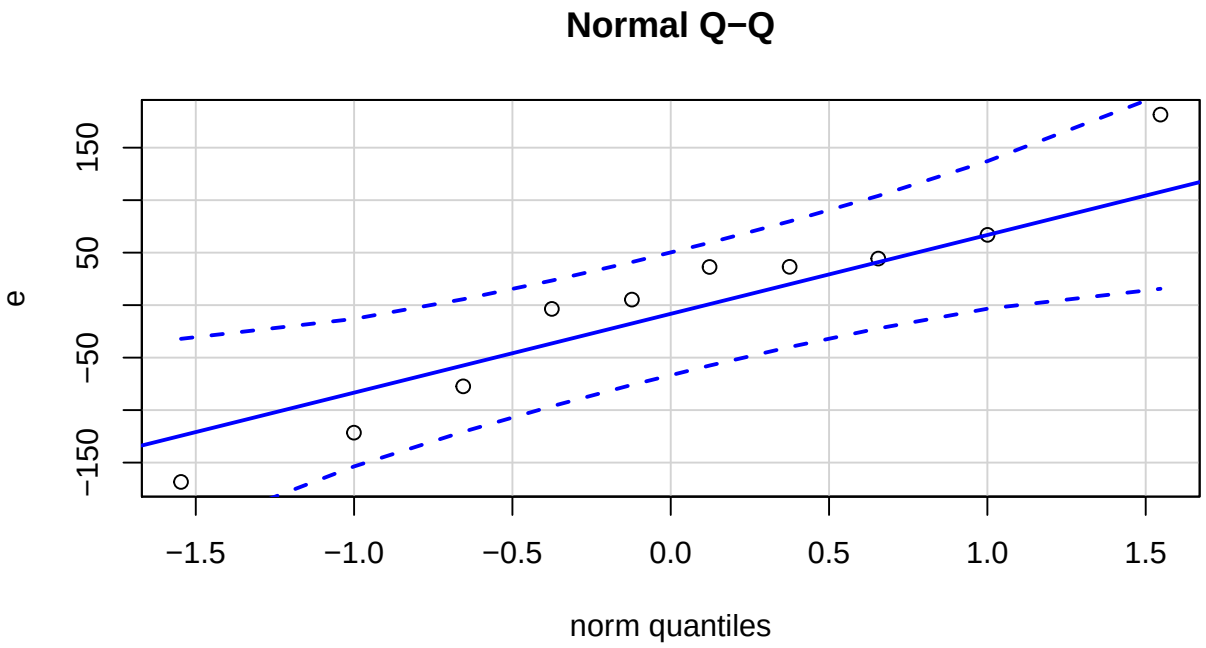
Napiszemy teraz funkcję `mnk`, która wyznaczy paramtery β_0 i β_1, wykorzystując metodę najmniejszych kwadratów.

```
mnk <- function(x,y){
  x_sr <- mean(x)
  y_sr <- mean(y)
  n <- length(x)
  s1 <- 0
  s2 <- 0
  for (i in c(1:n)) {
    s1 <- s1 + (x[i] - x_sr) * (y[i] - y_sr)
    s2 <- s2 + (x[i] - x_sr)^2
  }
  beta_1 <- s1/s2
  beta_0 <- y_sr - beta_1*x_sr
  return(c(beta_0,beta_1))
}
```

Korzystając z funkcji `mnk` otrzymujemy prostą regresji dla naszych danych:

Rysunek 2: Dopasowana prosta regresji

Sprawdzimy teraz, jakie własności mają reszty w naszym modelu. W tym celu sporządzimy wykres Normal Q–Q oraz wyznaczymy współczynnik determinacji.



Rysunek 3: Badanie własności reszt

Możemy założyć, że wektor reszt ϵ ma rozkład normalny. Współczynnik determinacji wynosi 0.8766597, zatem jest dość blisko 1. Stąd wnioskujemy, że uzyskany model jest dobrze dopasowany.

Korzystając z uzyskanego modelu, spróbujemy wyznaczyć prognozowaną wielkość sprzedaży dla nakładów 35, 45 i 55 [mln.$].

Nakład [mln.$]	Progn. wielk. sprz.
35.00	460.00
45.00	601.00
55.00	742.01

Tabela 2: Prognozowana wielkość sprzedaży

Teraz wyznaczymy przedziały ufności oraz przedziały predykcji dla otrzymanych wyników.

Nakład [mln.$]	Progn. wielk. sprz.	conf.lwr	conf.upr	pred.lwr	pred.upr
35.00	460.00	376.57	543.43	200.04	719.96
45.00	601.00	494.21	707.79	332.63	869.37
55.00	742.01	602.13	881.89	458.84	1025.18

Tabela 3: Przedziały ufnosci i predykcji - 0.95

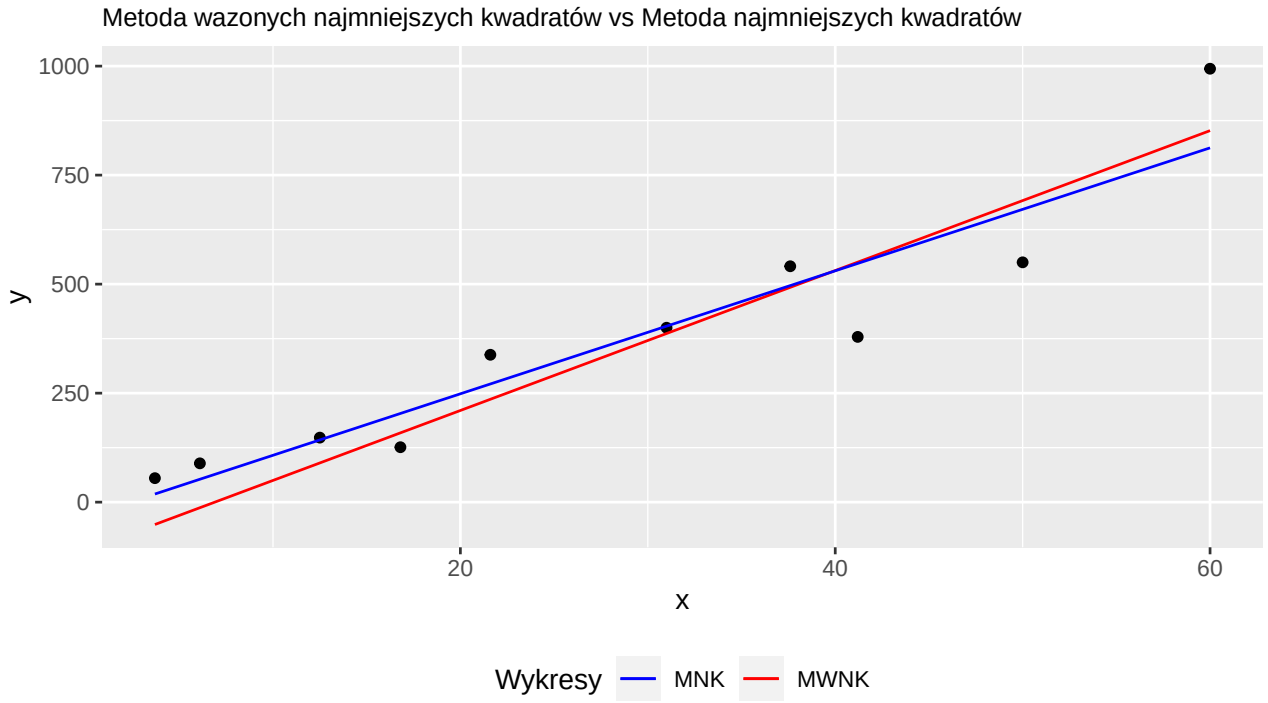
Odczytujemy, że np. przy nakładzie 35, z 95% prawdopodobieństwem wielkość sprzedaży będzie w przedziale $[200.04, 719.96]$.

Nakład [mln.$]	Progn. wielk. sprz.	conf.lwr	conf.upr	pred.lwr	pred.upr
35.00	460.00	338.61	581.39	81.74	838.26
45.00	601.00	445.62	756.39	210.51	991.50
55.00	742.01	538.47	945.54	329.98	1154.04

Tabela 4: Przedziały ufnosci i predykcji - 0.99

Analogicznie odczytujemy, że np. przy nakładzie 35, z 99% prawdopodobieństwem wielkość sprzedaży będzie w przedziale $[81.74, 838.26]$.

Znajdziemy teraz model regresji liniowej korzystając z metody najmniejszych ważonych kwadratów.



Rysunek 4: Porównanie MNK i MWNK

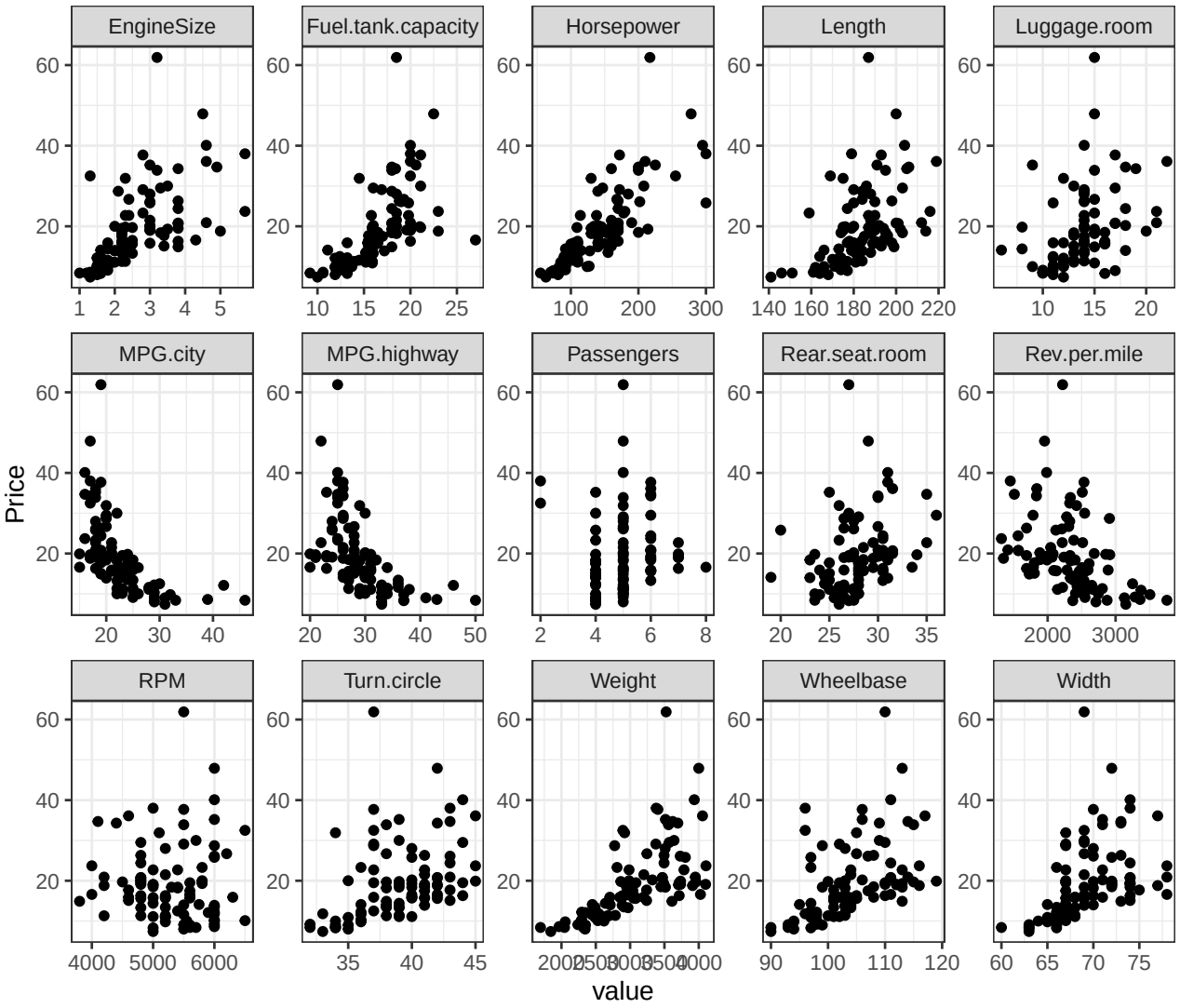
	R^2	σ
MNK	0.88	62.38
MWNK	0.82	113.16

Tabela 5: Porównanie MNK i MWNK

Obserwujemy, że krzywa uzyskana metodą ważonych najmniejszych kwadratów ma wyższy współczynnik β_1. Poza tym model uzyskany tą metodą charakteryzuje się niższym współczynnikiem determinacji oraz wyższym odchyleniem standardowym.

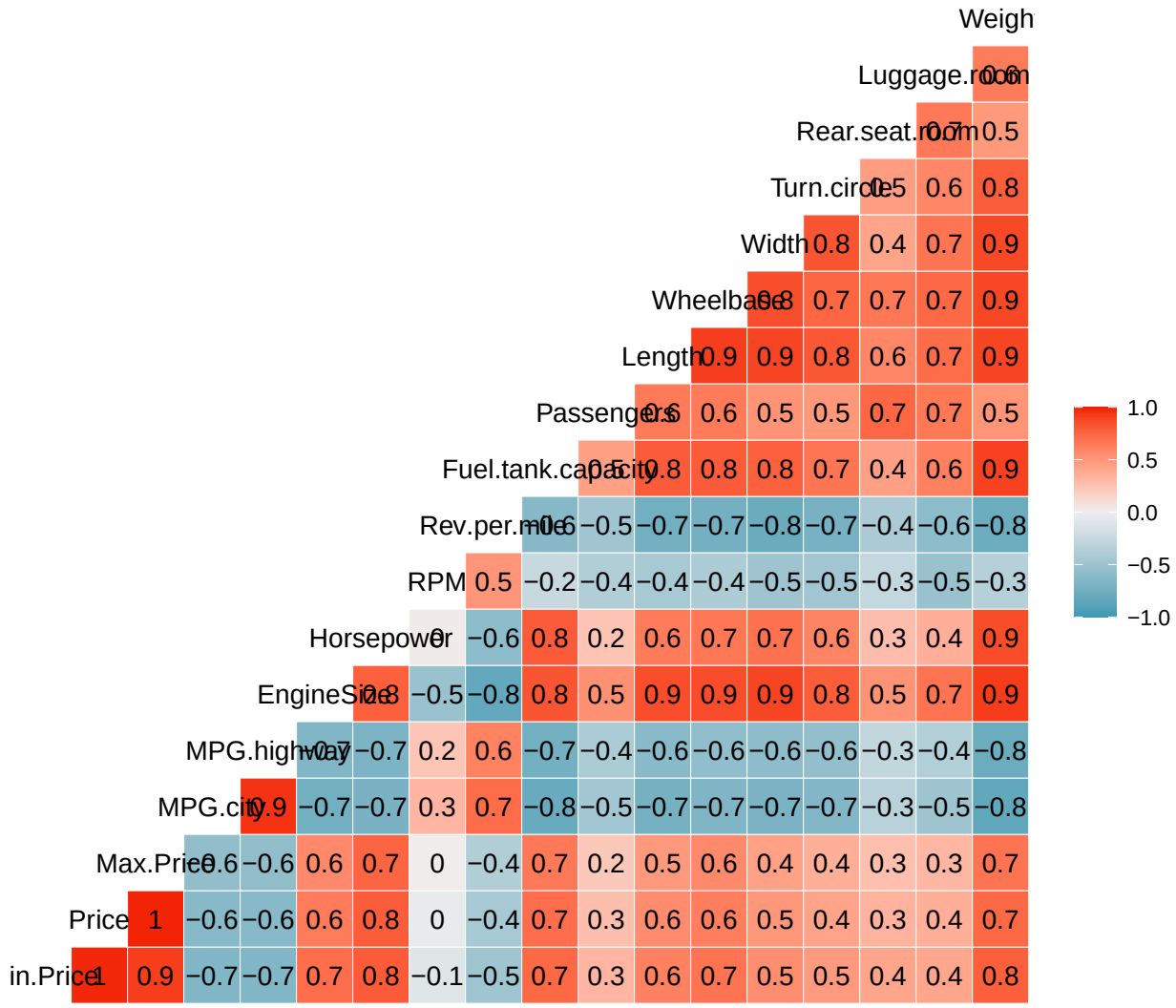
3.2 Porównanie modeli regresji liniowej dla danych Cars.93

Wczytujemy dane `Cars93` z pakietu `MASS`. Zbadamy zależności zmiennej `Price` od pozostałych cech (pomijamy cechy `Min.Price` i `Max.Price`, gdyż ta zależność jest oczywista). W tym celu przygotujemy wykresy rozrzutu dla cech ilościowych (typ `numeric` lub `integer`).



Rysunek 5: Dane

Widzimy, że dla zmiennej `Price` zależność zbliżona do liniowej występuje m.in. w przypadku `Horsepower`, `EngineSize`, `Fuel.tank.capacity`, `MPG.City`, `MPG.Highway`, `Weight` oraz `Wheelbase`. Sporządzimy teraz macierz korelacji, aby sprawdzić zasadność naszych obserwacji. Umieścimy też w tabeli te zmienne, dla których |współczynnik korelacji| ≥ 0.6 (względem zmiennej `Price`).

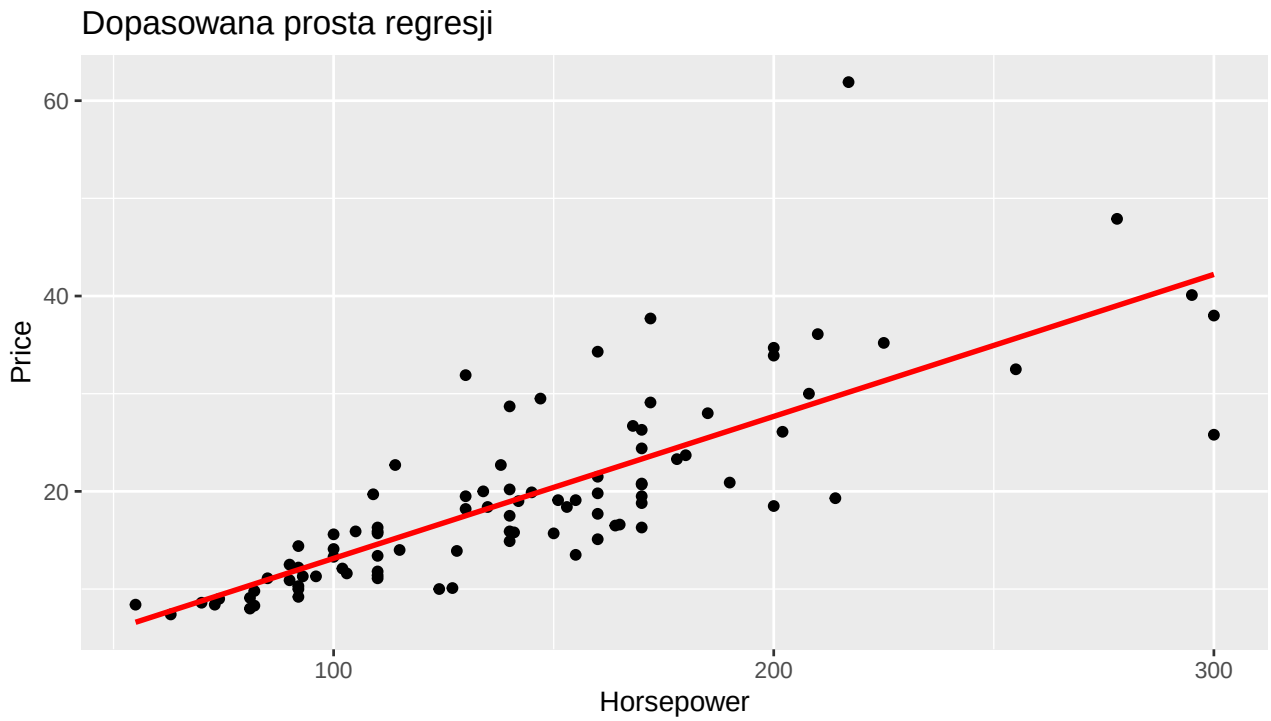


Rysunek 6: Macierz korelacji

	Price
Min.Price	0.97
Price	1.00
Max.Price	0.98
MPG.city	-0.62
MPG.highway	-0.63
EngineSize	0.64
Horsepower	0.79
Fuel.tank.capacity	0.71
Wheelbase	0.63
Weight	0.74

Tabela 6: Korelacja między Price i pozostałymi zmiennymi (korelacja ≥ 0.6)

Przygotujemy teraz modele regresji liniowej dla wybranych trzech cech - będą to `Horsepower`, `EngineSize` oraz `Wheelbase`. Rozpoczniemy od prostego modelu dla zmiennej `Horsepower`.



Rysunek 7: Model regresji dla 1 zm. objaśniającej

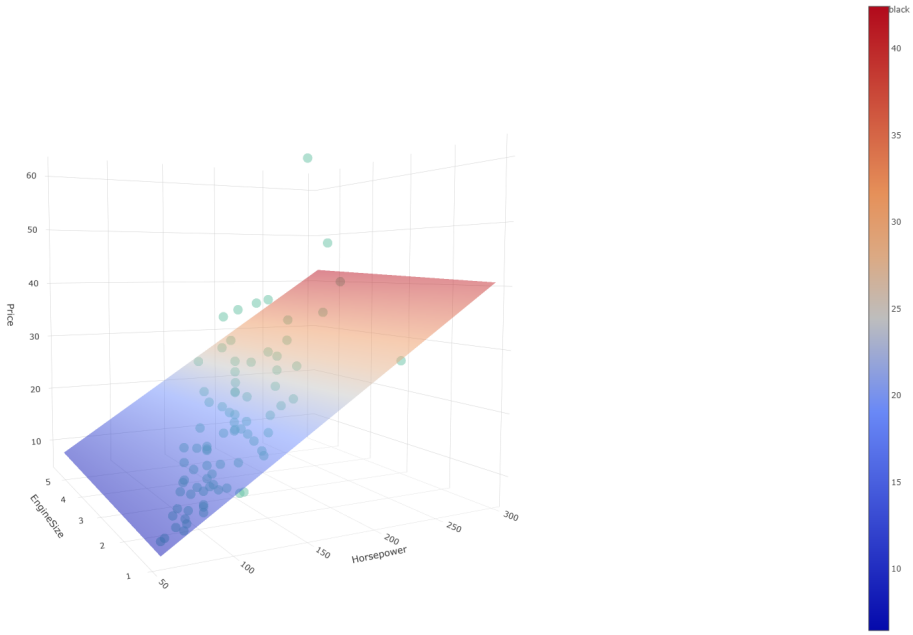
Model regresji dla 1 zmiennej objaśniającej (`Horsepower`) jest w przybliżeniu w postaci

$$\mathbf{Y}_i = -1.3988 + 0.1454 x_i + \epsilon_i \tag{1}$$

Model regresji dla 2 zmiennych objaśniających (`Horsepower`, `EngineSize`) jest w przybliżeniu w postaci

$$\mathbf{Y}_i = -1.6366 + 0.1394 x_{i,1} + 0.4085 x_{i,2} + \epsilon_i \tag{2}$$

Jest to równanie płaszczyzny. Możemy ją zwizualizować na wykresie.

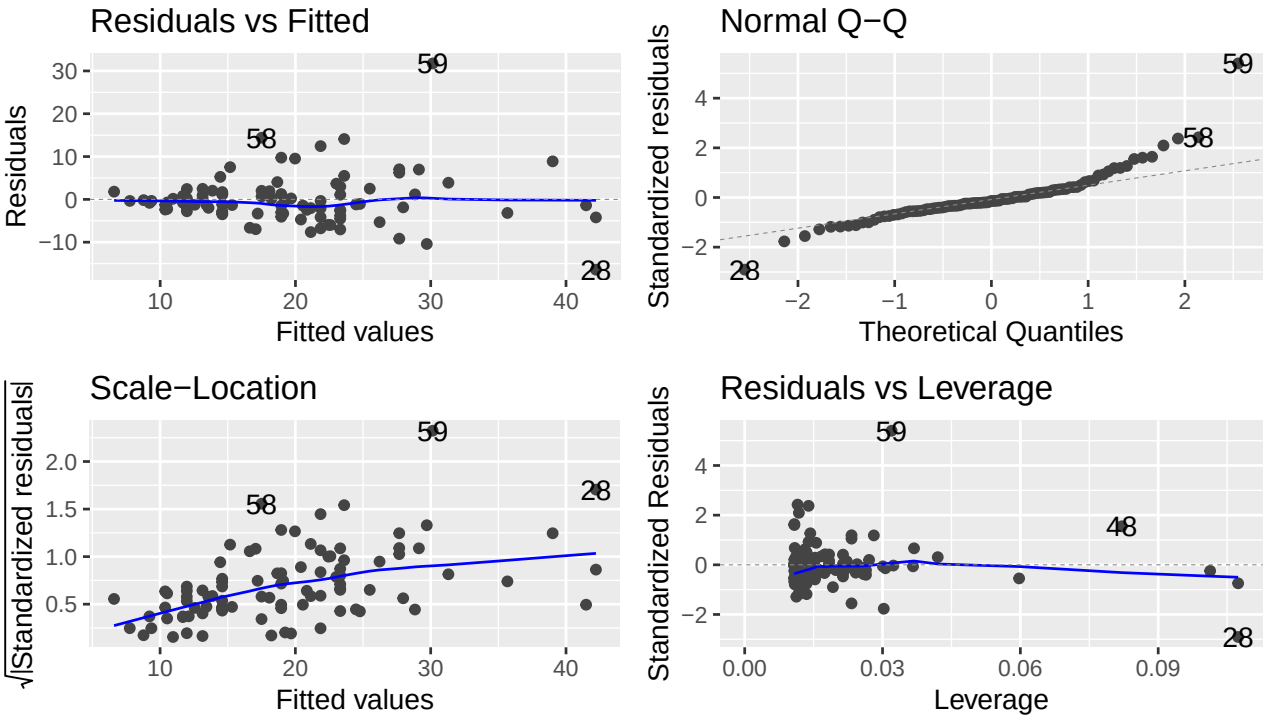


Rysunek 8: Model regresji dla 2 zm. objaśniających

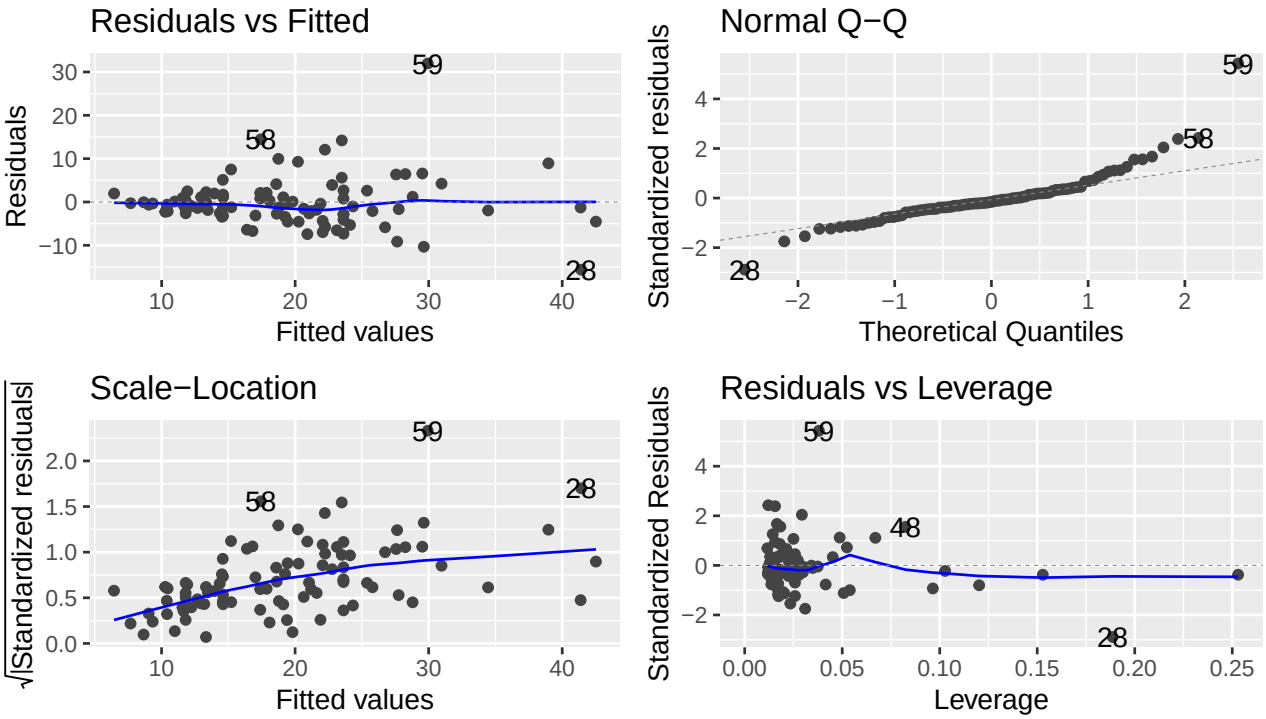
Model regresji dla 3 zmiennych objaśniających (`Horsepower`, `EngineSize`, `Wheelbase`) jest w przybliżeniu w postaci

$$\mathbf{Y}_i = -30.167 + 0.1438x_{i,1} - 1.2487x_{i,2} + 0.311x_{i,3} + \epsilon_i \tag{3}$$

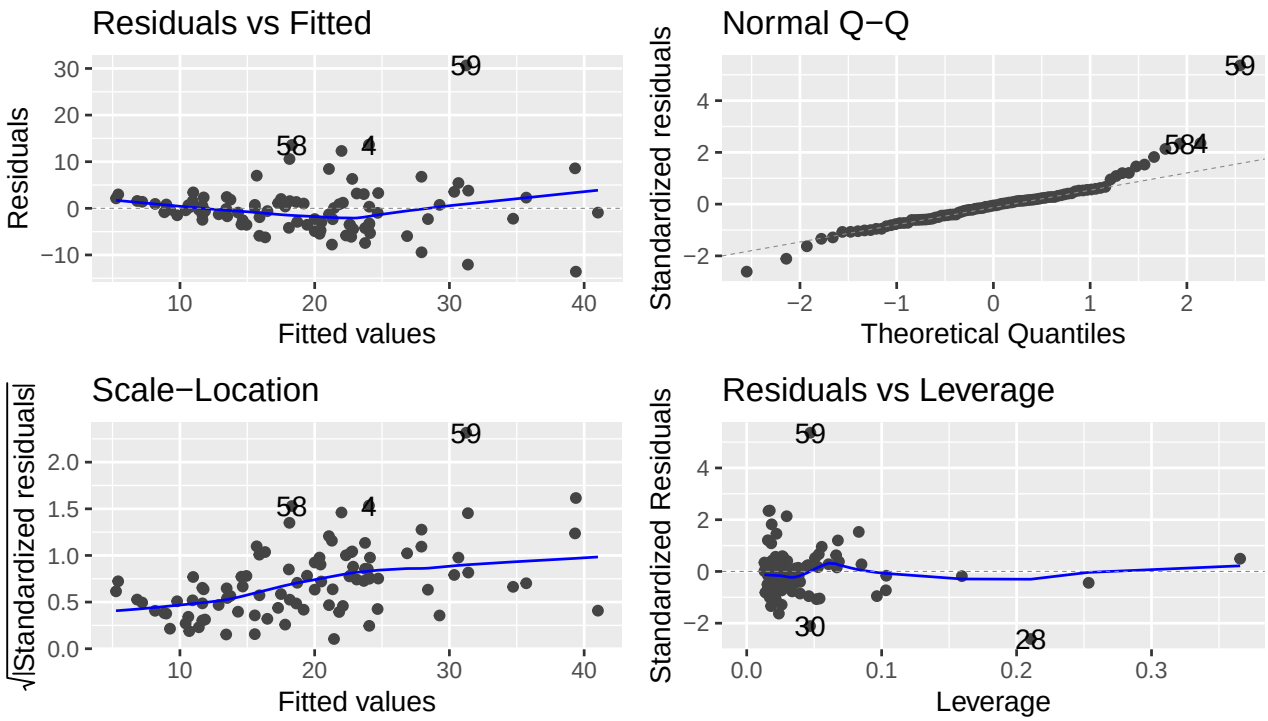
Spróbujemy teraz przeanalizować, który z tych trzech modeli jest najlepiej dopasowany. W tym celu sporządzimy wykresy, które pomogą nam zweryfikować poprawność dopasowania dla kolejnych modeli.



Rysunek 9: Weryfikacja poprawności dopasowania dla modelu 1



Rysunek 10: Weryfikacja poprawności dopasowania dla modelu 2



Rysunek 11: Weryfikacja poprawności dopasowania dla modelu 3

	R^2	σ
Model 1	0.62	5.98
Model 2	0.61	6.00
Model 3	0.63	5.86

Tabela 7: Współczynnik adjusted R^2 i odchylenie standardowe σ

Porównując wykresy Residuals vs Fitted oraz wykresy Normal Q–Q ciężko wyciągnąć jakieś wnioski. Na wykresie Scale–Location widać natomiast, że najbliżej poziomej linii jest krzywa w modelu 3. Także na wykresie Residuals vs Leverage delikatnie lepiej wypada model 3. Porównując współczynniki R^2 oraz σ również możemy wyciągnąć wniosek, że najlepiej dopasowanym modelem jest model 3, natomiast różnice nie są bardzo wyraźne.

4 Podsumowanie

Poniżej wypunktujemy najważniejsze wnioski, jakie można wyciągnąć z przeprowadzanych analiz:

- dobrze dopasowany model regresji liniowej pozwala nam na prognozowanie przyszłych wartości z zadanym prawdopodobieństwem,
- porównując wskaźniki adjusted R^2 oraz odchylenie standardowe σ możemy sprawdzać, jak dobrze dany model jest dopasowany do danych,
- porównanie wykresów Residuals vs Fitted, Normal Q–Q, Scale–Location oraz Residuals vs Leverage pozwala nam wyłonić najlepiej dopasowany model.