Jakub Kaczor Yana Negulescu

Wykorzystanie poznanych metod dotyczących analizy zależności liniowej do wybranych danych rzeczywistych

18 grudnia 2023

1. Wstęp i opis danych

Analizowany zbiór danych zawiera informacje o samochodach produkowanych w USA, Europie i Japonii w latach 1970–1982. Wersja z której korzystają autorzy pochodzi z repozytorium autorów pakietu do wizualizacji `seaborn` i za pomocą tego pakietu autorzy uzyskują do nich dostęp.

Tabela 2: Zmienne w zbiorze danych wraz z opisem

Zmienna	Opis
mpg	średnia przebytych mil na galon paliwa
cylinders	liczba cylindrów
displacement	pojemność silnika
horsepower	moc (w koniach mechanicznych)
weight	masa pojazdu
acceleration	przyspieszenie
model_year	rok produkcji
origin	kraj produkcji
name	nazwa modelu

Wśród zbioru znaleziono kilka obserwacji brakujących, jest ich 6, a brakuje w nich wartości dla kolumny `horsepower`. Druga zasada dynamiki Newtona orzeka zależność pomiędzy masą, siłą i przyspieszeniem. Wobec tego autorzy spodziewają się ścisłej relacji pomiędzy masą pojazdu, przyspieszeniem, a jego mocą i w tym sprawozdaniu pochylą się nad uzupełnieniem brakujących danych poprzez predykcję ich wartości z użyciem modelu regresji liniowej.

Kolejnym interesującym autorów problemem jest wydajność pojazdu. W danych jest zmienna bezpośrednio z tym związana, jest to zmienna `mpg`. Potrzeba

Tabela 1: Obserwacje z brakującymi wartościami

	origin	model_year	name
32	usa	71	ford pinto
126	usa	74	ford maverick
330	europe	80	renault lecar deluxe
336	usa	80	ford mustang cobra
354	europe	81	renault 18i
374	usa	82	amc concord dl

znaleźć minimalną liczbę zmiennych dobrze opisujących `mpg` w zależności liniowej. Pomysłem autorów jest, że to jak samochód jest w stanie daleko zajechać zależy od tego, jak ciężki jest, a—średnio—forma napędu nie ma większego znaczenia, co spróbują potwierdzić w analize reziduów modelu liniowego.

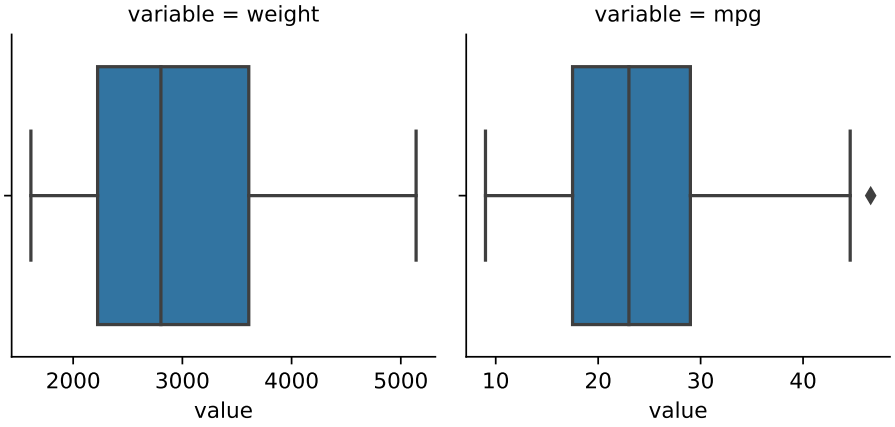
2. Analiza jednowymiarowa

Jak wspomniano w Sekcja 1, będziemy badać zależność

$$\text{moc} \sim \text{przyspieszenie} \cdot \text{masa}.$$

Do tego potrzeba nowej zmiennej będącej iloczynem przyspieszenia i masy. Nazwiemy ją `force` (pol. siła).

Do analizy wspomnianych w Sekcja 1 relacji będziemy potrzebowali zmiennych `acceleration`, `weight` oraz `mpg`. Jak okaże się w Sekcja 3.1, `acceleration` nie przyda się nam tak bardzo, więc dokładniej przyjrzymy się tylko `weight` i `mpg`.



2.1. Podstawowe statystyki dla zmiennej zależnej ‘mpg’

Średnia próbkowa dla zmiennej `mpg` wynosi: 23.5146. Nie znajduję się ona daleko od mediany, co jest widocznie na wykresie pudełkowym. To świadczy o małej skośności rozkładu.

Odchylenie standardowe próbkowe wynosi: 7.8160, stąd wariancja wynosi 61.0896. To świadczy o tym że większość obserwacji jest z przedziału: [15.6986, 31.3306].

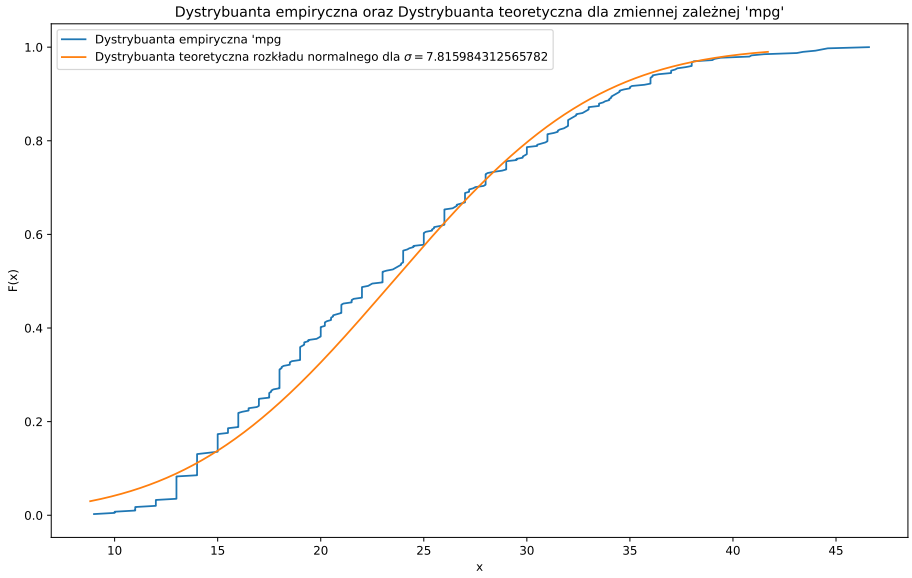
Mediana próbkowa wynosi: 23.00. Oraz kwartyli rzędu 25 i 75 wynoszą odpowiednio: 17.50, 29.00.

Element minimalny oraz maksymalny: 9.00, 46.60.

Kurtoza próbkowa wynosi: -0.5108. Oznacza to, że rozkład ma płaski szczyt i mniej grubych ogonów w porównaniu z rozkładem normalnym.

Skośność próbkowa wynosi: 0.4571. To oznacza, że rozkład jest prawostronnie skośny.

Aby lepiej zobaczyć na ile rozkład jest podobny do rozkładu normalnego z odpowiednimi parametrami ($\mu = 23.5146$, $\sigma = 7.8160$), narysujmy wykres porównujący dystrybuanty.



2.2. Podstawowe statystyki dla zmiennej niezależnej ‘weight’

Średnia próbkowa dla zmiennej `weight` wynosi: 2970.4246. Nie znajduję się ona daleko od mediany, co jest widocznie na wykresie pudełkowym. To świadczy o małej skośności rozkładu.

Odchylenie standardowe próbkowe wynosi: 846.8418, stąd wariancja wynosi 717140.9905. To świadczy o tym że większość obserwacji są z przedziału: [2123.5828, 3817.2664].

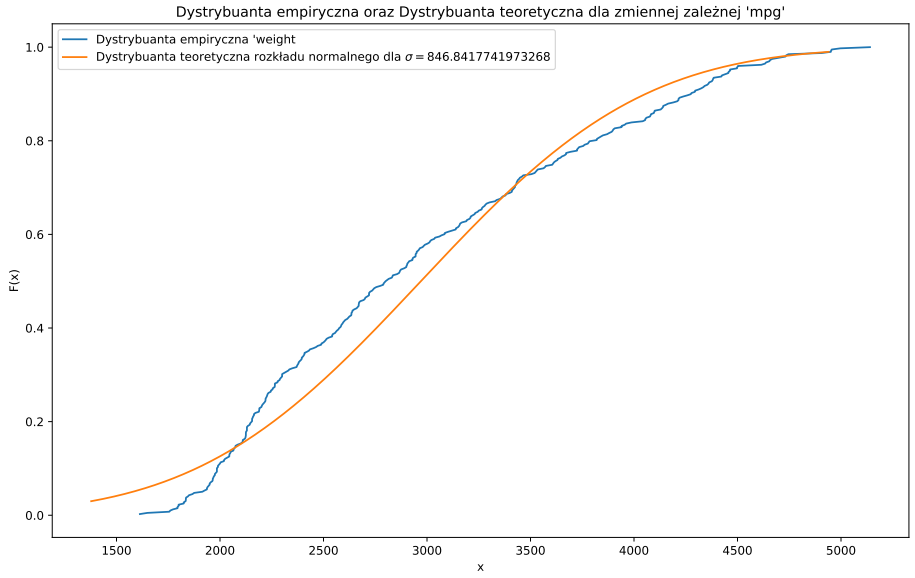
Mediana próbkowa wynosi: 2803.50. Oraz kwartyli rzędu 25 i 75 wynoszą: 2223.75, 3608.00.

Element minimalny oraz maksymalny: 1613.00, 5140.00.

Kurtoza próbkowa wynosi: -0.7855. Oznacza to, że rozkład ma płaski szczyt i mniej grubych ogonów w porównaniu z rozkładem normalnym.

Skośność próbkowa wynosi: 0.5311. To oznacza, że rozkład jest prawostronnie skośny.

Aby lepiej zobaczyć na ile rozkłąd jest podobny do rozkładu normalnego z odpowiednimi parametrami ($\mu = 2970.4246$, $\sigma = 846.8418$), narysujmy wykres dystrybuant dla porównania.



2.3. Wydzielenie danych

W dalszej części raportu będziemy dopasowywać model i testować jego poprawność. Aby ocenić, czy model nie jest przeuczony, wydzielimy podzbiór na którym model będzie dopasowywany i podzbiór na którym będziemy to dopasowanie testować, odpowiednio w proporcjach 80:20 w sposób losowy.

3. Analiza zależności

3.1. Moc, masa i przyspieszenie

Postulujemy zależność, zgodnie z klasycznym modelem regresji liniowej,

$$\text{horsepower} = \beta_1 \text{force} + \beta_0 + \varepsilon,$$

dla pewnych współczynników β_0, β_1 i zmiennej losowej $\varepsilon \sim \text{Normal}(0, \sigma^2)$ przy nieznanej wariancji σ^2.

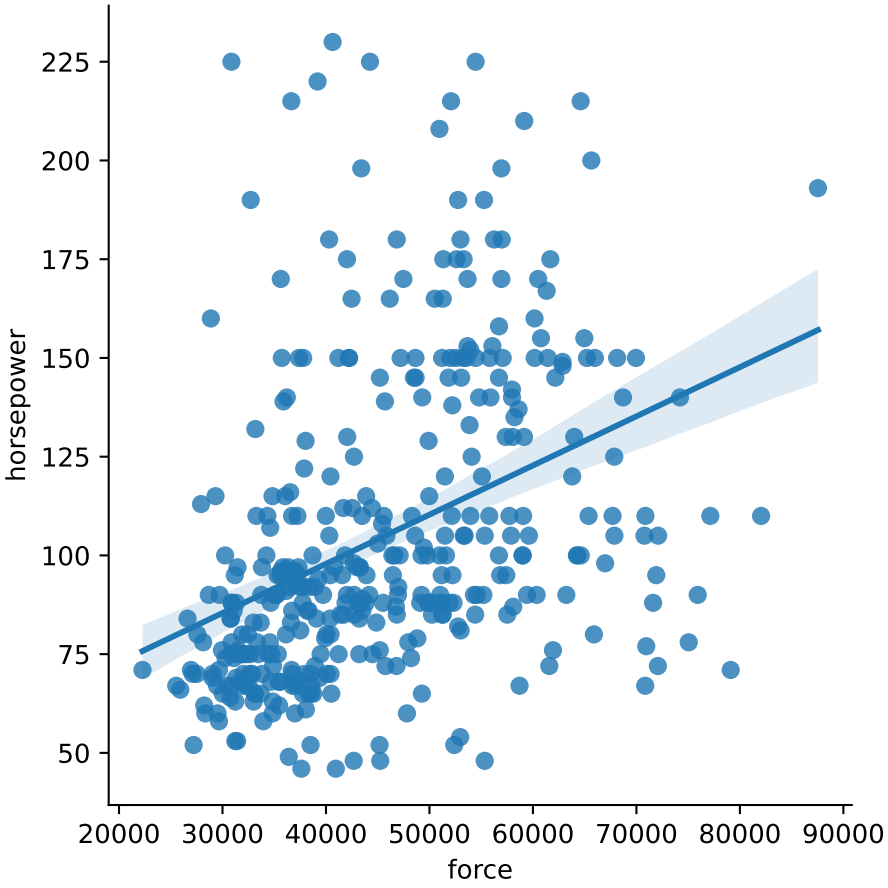
Jak można zobaczyć na Rysunek 1, postulowany model nie pasuje. Jako że nieodpowiedniość jest widoczna gołym okiem, porzucamy dalszą analizę relacji na rzecz kolejnej.

3.2. Wydajność

W podobny sposób co w Sekcja 3.1, postulujemy

$$\text{mpg} = \beta_1 \text{weight} + \beta_0 + \varepsilon,$$

dla pewnych współczynników β_0, β_1 i zmiennej losowej $\varepsilon \sim \text{Normal}(0, \sigma^2)$ przy nieznanej wariancji σ^2.



Rysunek 1: Regresja liniowa przy zmiennej objaśnianej `horsepower` i zmiennej objaśniającej `force`.

3.3. Estymacja punktowa oraz przedziałowa

Wyestymowaliśmy parametry β_0 i β_1 które odpowiednio wynoszą 46.1317 i -0.0076.

Oraz obliczyliśmy przedział ufności: dla β_0 to [44.3675, 47.8959], a dla β_1—[-0.0082, -0.0071]

3.4. Ocena poziomu zależności

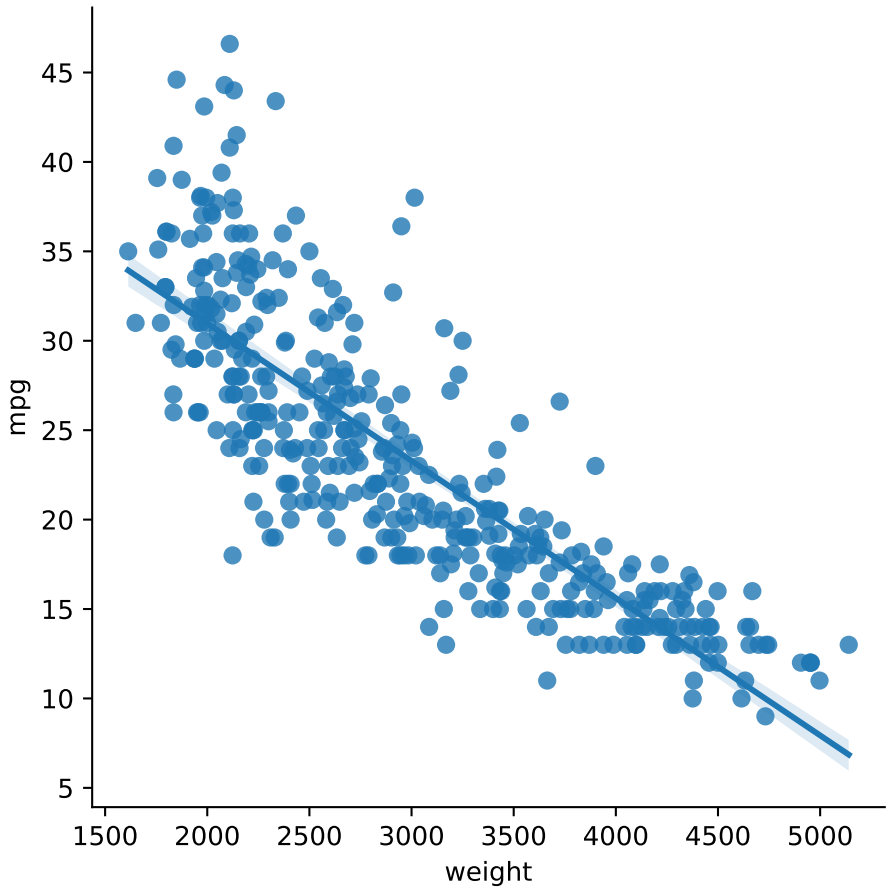
Żeby ocenić poziom zależności policzyliśmy:

SST = 4971.6395, co wskazuje na dużą zmienność danych.

SSE = 1465.5649, co wskazuje, że istnieje pewna zmienność w danych, która nie jest wyjaśniona przez model.

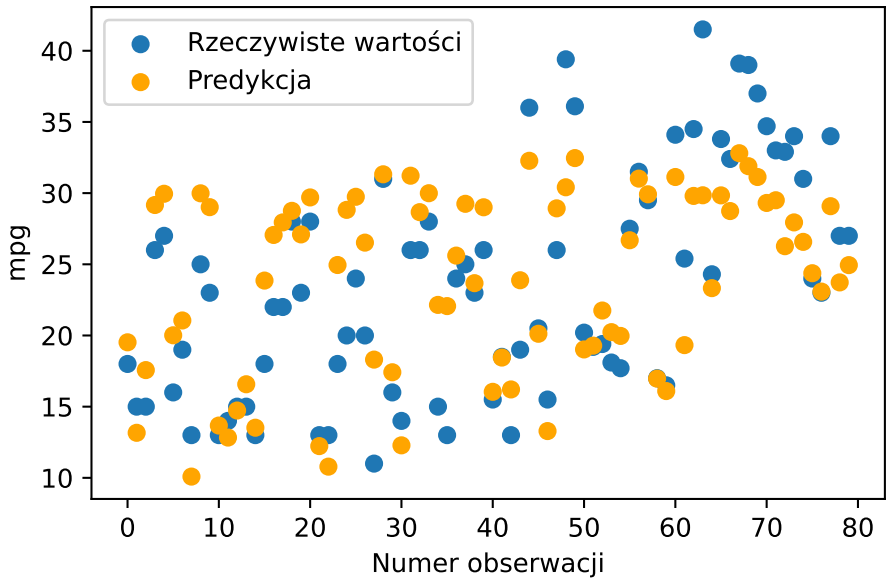
SSR = 3236.4607, oznacza to że model w miare dobrze wyjaśnia zmienność danych.

Współczynnik korelacji Pearsona $r^2 = 0.6510$. To oznacza, że około 65.09% zmienności zmiennej zależnej jest wyjaśnione przez model. Ta wartość sugeruje że model ma sensowną moc predykcyjną.



4. Predykcja

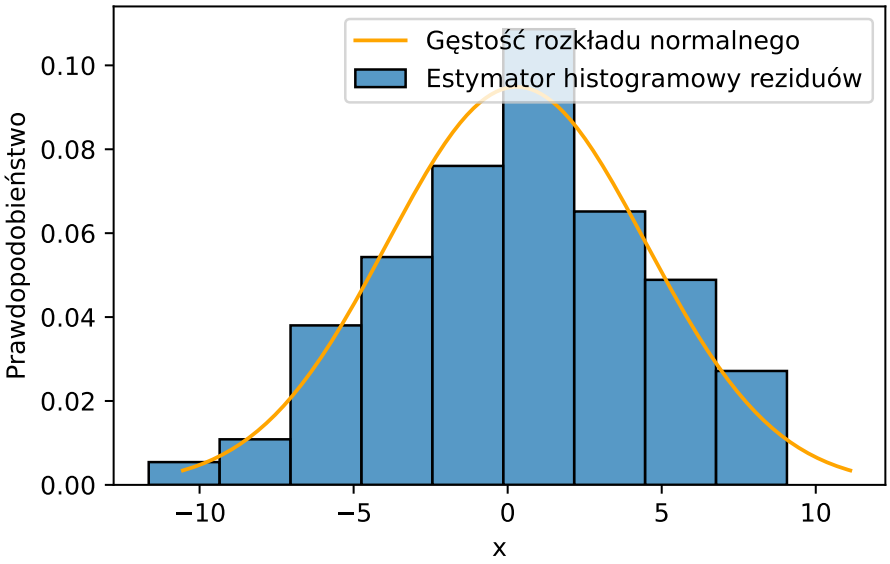
Korzystając z modelu opisanego w Sekcja 3.2, wykonamy predykcję na danych testowych i ocenę błedów. Rezultaty można zobaczyć na Rysunek 3. Średni błąd bezwzględny wyniósł 0.29, natomiast błąd względny 4.46%.



Rysunek 2: Wartości rzeczywiste oraz prognozowane dla zmiennej `mpg`. Nie zauważa się wzorców, co może sugerować odchylenia spowodowane szumem.

5. Analiza reziduów

W tym rozdziale sprawdzimy normalność reziduów modelu przedstowionego w Sekcja 3.2. Spodziewamy się rozkładu normalnego.



Rysunek 3: Estymator histogramowy reziduów oraz gęstość rozkładu normalnego o średniej i wariancji równych średniej i wariancji próbkowej reziduów.

Średnia reziduów wynosi 0.29, natomiast odchylenie standardowe to 4.20. Oprócz tego, jako że średnia i wariancja nie są znane, autorzy przeprowadzili test normalności Shapiro-Wilka. Ustalono poziom istotności jako $\alpha = 0.05$. Wartość statystyki testowej tego testu to 0.99 z p-wartością równą 0.8582, co nie stanowi podstaw do odrzucenia hipotezy, że rezidua faktycznie pochodzą z rozkładu normalnego.

6. Podsumowanie

W tym sprawozdaniu przyjrzano się zbiorowi danych `mpg` i rozważono głównie dwa aspekty oparte o modele regresji liniowej. Pierwszy, inspirowany fizycznym prawem okazał się niedobrze odwzorowywać rzeczywiste zależności w danych, natomiast drugi odwrotnie—był trafny. Opisano zmienne mające udział w modelach z użyciem podstawowych statystyk opisowych. Dopasowanie modelu sprawdzono z użyciem poznanych wskaźników, a jego poprawność dodatkowo potwierdzono przy analize reziduów. Całość została wykonana na danych treningowych, a proces oceny został przeprowadzony na danych testowych, co dodatkowo wzmacnia model o pewność, że nie jest przeuczony.