concordance=TRUE

Politechnika Wrocławska
Wydział Matematyki
Komputerowa Analiza Szeregów Czasowych

Analiza zbioru danych rzeczywistych

Patryk Statkiewicz 262307

22.12.2022 r.

```
## Error: <text>:4:0: nieoczekiwany koniec wejścia
## 2:   pdf.options(encoding = 'CP1250')
## 3: library("ggplot2")
##    ^
```

Spis treści

1 Wprowadzenie

1.1 Wstęp

Każdy wie, że konie mechaniczne są bardzo ważnym czynnikiem w wyborze auta. Dla niektóych fanów szybkiej jazdy jest to nawet najważniejszy parametr w pojeździe. Najczęściej najwięcej koni mechanicznych mają auta sportowe, które są zazwyczaj droższe niż zwykłe auta osobowe. Rodzi się więc pytanie jak moc silnika ma się do innych parametrów auta, a co najważniejsze ceny.

```
## Warning in file(file, "rt"): nie można otworzyć pliku 'CarPrice_Assignment.csv': No
such file or directory
## Error in file(file, "rt"): nie można otworzyć połączenia
## Error in file$horsepower: obiekt typu 'closure' nie jest ustawialny
## Error in file$price: obiekt typu 'closure' nie jest ustawialny
## Error in file$horsepower: obiekt typu 'closure' nie jest ustawialny
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'X'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'X'
```

1.2 Opis danych

Do analizy został wybrany zbiór danych **"Car Price Prediction (Linear Regression)"** ze strony Kaggle `https://www.kaggle.com/code/ashydv/car-price-prediction-linear-regression/data/`. Dane zawierają 205 wierszy oraz 26 kolumn. Do naszej analizy będą potrzebne dwie kolumny odpowiadające za cenę auta oraz za moc silnika.

1.3 Opis zmiennych

Zbiór danych do analizy:

- **horsepower** (zmienna objaśniająca) - ilość koni mechanicznych w silniku danego auta wyrażona w *KM* (konie mechaniczne). Maksymalna moc silnika to 288 *KM*, a minimalna 48 *KM*.

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

- **price** (zmienna objaśniana) - cena za auto wyrażona w *RMB* (Renminbi, waluta Chin). Maksymalna cena to 45400 *RMB*, a minimalna 5118 *RMB*.

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

1.4 Cel analizy

Aby prawidłowo ocenić zależności pomiędzy ceną auta a ilością koni mechanicznych w silniku będziemy na początku sprawdzać statystyki opisowe dla każdego zbioru danych. Później zajmiemy się dopasowaniem modelu regresji liniowej do danych oraz zbadaniem przedziałów ufności. Ponadto przeprowadzimy analizę residów i postaramy się wysnuć predykcje na podstawie otrzymanych wyników. W naszej analizie przyjmiemy X jako kolumnę z ilością koni mechanicznych, a Y jako tabela wartości cen aut. Wektor

zmiennych X będzie zmienną objaśniającą, a wektor wartości Y zmienną objaśnianą. Korelacja pomiędzy danymi wyliczona analitycznie wynosi ≈ 0.8081. Jest na tyle wysoka, że na spokojnie możemy spróbować dopasować model regresji liniowej.

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

2 Statystyki opisowe dla danych

Poniżej przedstawiam definicje oraz wyniki statystyk opisowych, których używamy w analizie naszych danych.

- **Średnia arytmetyczna** $\bar{x}$ jest to iloraz sumy wszystkich n obserwacji przez ich ilość. Dla danych $x_1, \dots x_n$ wyraża się wzorem:
$$\bar{x} = \frac{x_1 + \dots + x_n}{n}. \tag{1}$$
Dla naszych zmiennych średnia arytmetyczna ilości koni mechanicznych wynik to $\bar{X} = 104.1171$, a dla cen aut wynosi $\bar{Y} = 13276.711$.

- **Mediana** jest wartością środkową w posortowanej próbie. W zależności, czy wielkość próby n jest parzysta, czy nieparzysta mediana wyraża się wzorem:
$$x_{med} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{gdy } n \text{ nieparzyste}, \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \text{gdy } n \text{ parzyste}. \end{cases} \tag{2}$$
Medianą ilości koni mechanicznych jest wynik $X_{med} = 95$, a mediana cen pojazdów to $Y_{med} = 10295$.

- **Wariancja** z próby $x_1, \dots x_n$ jest średnią arytmetyczną kwadratów odchyleń od wartości średniej oczekiwanej. Wyraża się wzorem:
Dla estymatora nieobciążonego:
$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{3}$$
Dla estymatora obciążonego:
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{4}$$
W naszym przypadku, dla danych rzeczywistych, obliczamy wariancję wzorem na estymator obciążony. Dla ilości koni mechanicznych jest równa $\sigma_X^2 \approx 1563.741$, a dla cen pojazdów wariancja wynosi $\sigma_Y^2 \approx 63821762$.

- **Odchylenie standardowe** σ z próby jest pierwiastkiem kwadratowym z wariancji. Odchylenie standardowe ilości koniów mechanicznych $\sigma_X \approx 39.54417$, a dla cen pojazdów to $\sigma_Y \approx 7988.852$.

- **Kwartyle** dzielą zbiór danych na 4 grupy:
  - drugi kwartyl (Q2) to mediana,
  - pierwszy kwartyl (Q1) to mediana grupy obserwacji mniejszych od Q2, $Q1_X = 70, \quad Q1_Y = 7788$,
  - trzeci kwartyl (Q3) to mediana grupy obserwacji większych od Q2, $Q3_X = 116, \quad Q3_Y = 16503$.

- **Rozstępem międzykwartalnym (IQR)** nazywamy liczbę równą różnicy kwartylu trzeciego i kwartylu pierwszego: $Q3 - Q1$. $IQR_X = 46, \quad IQR_Y = 8715$.

3 Dobranie prostej regresji

3.1 Metoda najmniejszych kwadratów

Aby prawidłowo dopasować prostą regresji liniowej będziemy szukać takich wartośći β_0, β_1, aby jak najlepiej spełniały warunek $\hat{Y}_i = \hat{\beta}_1 \cdot X_i + \hat{\beta}_0$. Użyjemy do tego metody najmniejszych kwadratów. Starając się znaleźć najlepsze estymatory β_0, β_1 skorzystamy z sumy kwadratów błędów:

$$S(\hat{\beta}_1, \hat{\beta}_0) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 \cdot X_i - \hat{\beta}_0)^2 \tag{5}$$

Następnie obliczymy pochodne funkcji $S(\hat{\beta}_1, \hat{\beta}_0)$ po β_0, β_1:

$$\frac{\delta S(\hat{\beta}_1, \hat{\beta}_0)}{\delta \hat{\beta}_1} = \sum_{i=1}^{n} -2X_i(Y_i - \hat{\beta}_1 \cdot X_i - \hat{\beta}_0) = 0, \qquad \frac{\delta S(\hat{\beta}_1, \hat{\beta}_0)}{\delta \hat{\beta}_0} = \sum_{i=1}^{n} -2(Y_i - \hat{\beta}_1 \cdot X_i - \hat{\beta}_0) = 0. \tag{6}$$

Z równania po prawej wyznaczamy wartość estymatora $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$. Podstawiając do równania po lewej stronie możemy już obliczyć wartość estymatora $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{7}$$

```
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'X'
## Error in mean(Y): nie znaleziono obiektu 'Y'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'X'
```

Stosując powyższą metodę współczynnik $\hat{\beta}_1$ wynosi 163.2631, a współczynnik $\hat{\beta}_1$ jest równy −3721.761. Podstawiając otrzymane wartości do naszego modelu $\hat{Y}_i = \hat{\beta}_1 \cdot X_i + \hat{\beta}_0$ otrzymujemy prostą regresji:

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

3.2 Jakość dopasowania

Zdefiniumy trzy trzy sumy:

- **Całkowita suma kwadratów**: $\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- **Suma kwadratów błędów**: $\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- **Suma kwadratów regresji**: $\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$

Prawdą jest, że SST = SSR + SSE. Współczynnikiem determinacji określającym stopień w jakim występuje zależność liniowa między zmienną objaśniającą, a zmienną objaśnianą nazywamy wielkość równą $R^2 = \frac{\text{SSR}}{\text{SST}}$. Im R^2 jest bliższe wartości 1 to tym lepiej dopasowaliśmy model regresji. Ze wcześniejszych informacji można również wywnioskwać, że im większe SSR tym większe R^2. Nasze $R^2 \approx 0.6530884$, więc z tego wynika, że nasz model regresji nie jest idealny ale jest do zaakceptowania.

4 Przedziały ufności

Jako że badamy dane rzeczywiste nasza σ jest nieznana. W takim przypadku przedział ufności wyznaczamy za pomocą kwantyla rozkładu t-studenta $t_{n-2,1-\frac{\alpha}{2}}$, gdzie n to długość próby i α to poziom istotności. Obliczymy również wartość S^2, która jest nieobciążonym estymatorem σ^2 i jest wyrażony wzorem $S^2 = \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}$. Procedurę wpadania danej wartości do wyznaczonego przedziału wykonamy $M = 1000$ razy dla trzech różnych poziomów istotności $\alpha_1 = 0.01, \alpha_2 = 0.05, \alpha_3 = 0.1$.

4.1 Przedział ufności β_0

$$\frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}} \sim N(0,1) \tag{8}$$

$$P(-t_{n-1,1-\frac{\alpha}{2}} S\sqrt{\tfrac{1}{n}+\tfrac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} + \hat{\beta}_0 < \beta_0 < t_{n-1,1-\frac{\alpha}{2}} S\sqrt{\tfrac{1}{n}+\tfrac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} + \hat{\beta}_0) = 1-\alpha \tag{9}$$

4.2 Przedział ufności β_1

$$\frac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}} \sim N(0,1) \tag{10}$$

$$P(-t_{n-1,1-\frac{\alpha}{2}} \tfrac{S}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}} + \hat{\beta}_1 < \beta_1 < t_{n-1,1-\frac{\alpha}{2}} \tfrac{S}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}} + \hat{\beta}_1) = 1-\alpha \tag{11}$$

4.3 Końcowa analiza wyników

Załóżmy, że średnią dolną granicę przedziału dla β_0 zdefiniujemy jako $A^-_{\beta_0}$ oraz górną jako $A^+_{\beta_0}$ i odpowiednio dla $\beta_1, A^-_{\beta_1}, A^+_{\beta_1}$. Po $M = 1000$ symulacjach dla każdego α wyniki są następujące:

$\cdot$	$\alpha_1 = 0.01$	$\alpha_2 = 0.05$	$\alpha_3 = 0.1$
$A^-_{\beta_0}$	-5429.822	-4988.035	-4808.366
$A^+_{\beta_0}$	-1997.945	-2390.224	-2627.025
$A^-_{\beta_1}$	148.6998	152.2243	154.917
$A^+_{\beta_1}$	177.708	174.1438	172.2801
$P(A^-_{\beta_0} < \beta_0 < A^+_{\beta_0})$	0.992	0.95	0.892
$P(A^-_{\beta_1} < \beta_1 < A^+_{\beta_1})$	0.99	0.943	0.89

Tabela 1: Tabela przedziałów oraz prawdopodobieństwa w zależności od poziomu istotności α.

Widzimy, że im mniejszy przedział istotności, tym większe są przedziały oraz naturalnie prawdopodobieństwo, że nasze β_0 trafi do naszego przedziału.

5 Analiza residuów

```
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'Y'
```

Różnice pomiędzy wartością Y_i a teoretyczną wartością $\hat{Y}_i$ nazywami błędem e_i. Wyrażamy je wzorem:

$$e_i = y_i - \hat{y}_i.$$

Suma wartości resztkowych (residuów) $\sum_{i=1}^{n} e_i$ powinna się sumować do 0. Suma błędów w naszym modelu wynosi $-1.95 \cdot 10^{-13}$; jest bardzo bliska wartości 0. Analitycznie wyliczona wariancja $\mathrm{Var}(e_i) \approx 22163811$. Tak duża wariancja związana jest z nieidealnymi rzeczywistymi danymi. Y_i można wyrazić jako $Y_i = \beta_1 \cdot X_i + \beta_0 + e_i$. W teoretycznym modelu regresji liniowej residua powinny być z rozkładu normalnego $N(0, \sigma^2)$ oraz być niezależne. Sprawdzenie tych warunków pozwoli ocenić nam jak rzeczywiście zachowują się wartości resztkowe i poprawność dobranego modelu regresyjnego:

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

Jak widać na wykresie błędy wydają się oscylować wokół 0, ale pod koniec ich wartość się zwiększa. Może być to spowodowane nieidealnymi danymi rzeczywistymi oraz wielkościami danych rzędu 10^5.

5.1 Wykres pudełkowy

Łatwiej informacje będzie można zdobyć z wykresu pudełkowego:

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

Z wykresu boxplot można odczytać, że niektóre wartości są odstające. Policzymy wartości, które mogą nam je wyrzucić. $\mathrm{IQR}_e = 3996.662$, $\mathrm{Q2}_e = -571.7144$, $\mathrm{Q3}_e + 1.5 \cdot \mathrm{IQR} = 7701.817$, $\mathrm{Q1}_e - 1.5 \cdot \mathrm{IQR} = -8284.83$. Dwie ostatnie wartości to kolejno górny i dolny wąs. Wartości, które nie mieszczą się w zadanym przedziale wyrzucimy.

```
## Error in if (e[i] == out[j]) {: argument jest długości zero
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'X'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'Y'
## Error in -ind: niepoprawny argument przekazany do operatora jednoargumentowego
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'X1'
## Error in mean(Y1): nie znaleziono obiektu 'Y1'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'X1'
```

5.2 Usunięcie wartości odstających

Wyrzuciliśmy 17 wartości, jak na 206 danych nie jest to mały wynik, ale też nie jest za duży. Teraz wykresy po wyrzuceniu wyglądają następująco:

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

Po odrzuceniu wartości odstających otrzymaliśmy już inne wyniki $\beta_0 = -2493.058$, $\beta_1 = 143.2368$ $R^2 \approx 0.7248339$. Zwiększenie się wartości R^2 świadczy o lepszym dopasowaniu modelu regresji liniowej do naszych danych.

```
## Error in rnorm(n, mean(e), sqrt(var(e))): nie znaleziono obiektu 'n'
## Warning in mean.default(e): argument nie jest wartością liczbową ani logiczną: zwracanie
wartości NA
## Error in var(e): 'x' ma wartość NULL
```

Po odrzuceniu wartości odstających postaramy się zobaczyć, czy nasze residua mają rozkład normalny z parametrami $\mu = 0, \sigma_e^2 = \text{Var}(e_i)$. Wykorzystamy do tego gęstość prawdopodobieństwa rozkładu normalnego:

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

Widzimy, że gęstości się pokrywają, więc można założyć, że nasze residua mają rozkład normalny z zadanymi parametrami.

5.3 Empiryczna funkcja autokowariancji

```
## Warning in mean.default(e): argument nie jest wartością liczbową ani logiczną: zwracanie
wartości NA
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'n'
```

Żeby sprawdzić, czy dane nie są skorelowane można na nie nałożyc funkcję autokowariancji zadaną wzorem:

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (e_{i+|h|} - \hat{e}) \cdot (e_i - \hat{e}).$$

Wykres porównujący wartości teoretyczne funkcji autokowariancji z wartościami empirycznymi wygląda następująco:

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

Wnioskując z wyglądu wykresu, wartości empirycznej autokowariancji kręcą się wokół wartości teoretycznych, ale poza pojedynczymi przypadkami całkiem odbiegają od wartości teoretyczych. Pomimo, że dla $h = 0$ wartość empiryczna jest bardzo blisko, tak dla $|h| = 1$, błąd już jest bardzo duży. Na podstawie wykresu ciężko z pełną pewnością powiedzieć, że błędy nie są skorelowane.

6 Predykcje

```
## Error in seq(min(X), max(X), length.out = n): nie znaleziono obiektu 'X'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'n'
## Error in rnorm(n, 0, sqrt(var(e))): nie znaleziono obiektu 'n'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'n'
## Warning in mean.default(YY): argument nie jest wartością liczbową ani logiczną: zwracanie
wartości NA
## Error in mean(xx): nie znaleziono obiektu 'xx'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'n'
```

```
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'xx_sr'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'n'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'n'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'n'
## Error in eval(expr, envir, enclos): nie znaleziono obiektu 'pom'
```

Predykcje to jest przedział, stworzony za pomocą przedziałów ufności, w którym się spodziewamy następnych obserwacji. Oczywiście ma znaczenie, czy σ jest znana, czy nie jest. Tak jak w poprzednim przypadku. musimy rozpatrywać przykład z σ-nieznana. Ważny jest też poziom istotności, w naszym przypadku α będzie równa 0.05.

```
## Error in ggplot(): nie udało się znaleźć funkcji 'ggplot'
```

Z wykresu można zobaczyć, że nasze wysymulowane sztuczne obserwacje mieszczą się w zadanym przedziale. Zatem możemy stwierdzić, że nasze predykcje były poprawne.

7 Podsumowanie

Taka obligatoryjna analiza danych ma bardzo wiele plusów. Nie dość, że można wykryć trendy w danych to również po większych przemyśleniach, obliczeniach można jeszcze te trendy sprecyzować. Dobrym przykładem jest usunięcie wartości odstających. Po dłuższej analizie po której doszliśmy do usuwania ekstremalnych wartości, nasz nowy model regresji liniowej był lepszy niż bez usunięcia tych danych, o czym mówił współczynnik determinacji R^2. Na podstawie całej analizy przedstawionej w sprawozdaniu można wywnioskować, że ilość koni mechaniznych w pojeździe ma znaczny wpływ na cenę auta. Rzadko w danych rzeczywistych spotyka się tak wysoki współczynnik determinacji.