Komputerowa analiza szeregów czasowych raport 1

Natalia Klepacka, Joanna Kołaczek

21.12.2022

Spis treści

1 Wstęp **2**

2 Analiza jednowymiarowa zmiennej zależnej oraz zmiennej niezależnej **2**

3 Analiza zależności liniowej pomiędzy zmienną zależną a zmienną niezależną **4**
3.1 Wykres rozproszenia i określenie zależności 4
3.2 Punktowa estymacja współczynników 5
3.3 Przedziałowa estymacja współczynników 6
3.4 Ocena poziomu zależności 6
3.5 Predykcja oraz jej przedziały ufności 6
3.6 Interpretacja wyników 6

4 Analiza residuów **6**

5 Podsumowanie **8**

6 Źródła **9**

1 Wstęp

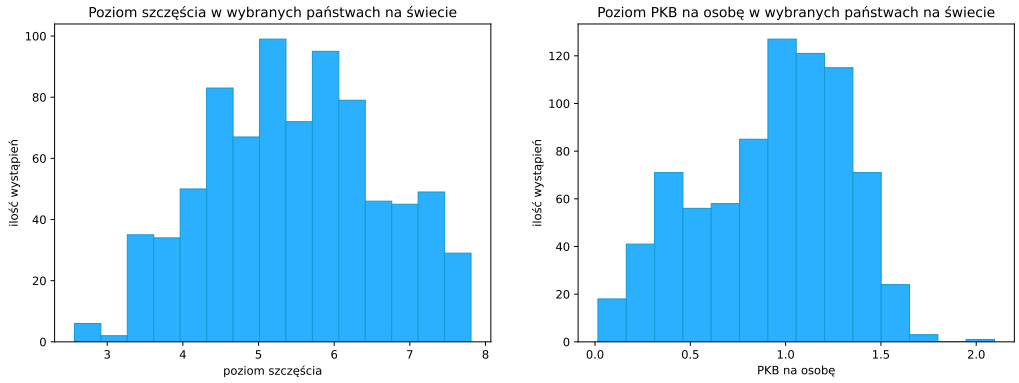
Niniejszy raport powstał na potrzeby realizacji laboratorium z Komputerowej Analizy Szeregów Czasowych, prowadzonych przez mgr Justynę Witulską, do wykładu prof. Agnieszki Wyłomańskiej. Będziemy analizować dane dotyczące poziomu szczęścia w wybranych krajach na świecie, oraz jego związku z wartością PKB na osobę (w dalszej części raportu, będziemy je określać skrótowo jako szczęście i PKB). Po usunięciu wartości brakujących, dysponujemy próbami o wielkości 791. Dane pochodzą *z tej strony*. Są to wyniki uzyskane przez Instytut Gallupa, w ankietach badających poziom szczęścia oraz jego możliwe indykatory, zebrane w latach 2015-2020. W raporcie przeprowadzimy analizę jednowymiarową dla dwóch zmiennych oraz zwizualizujemy je przy pomocy histogramu, dystrybuanty empirycznej oraz boxplotu. Następnie wyestymujemy współczynniki w klasycznym modelu regresji, aby na końcu sprawdzić czy uzyskane residua spełniają oczekiwane założenia.

Życzymy Czytelnikowi miłej lektury.

2 Analiza jednowymiarowa zmiennej zależnej oraz zmiennej niezależnej

2.1 Wizualizacja danych

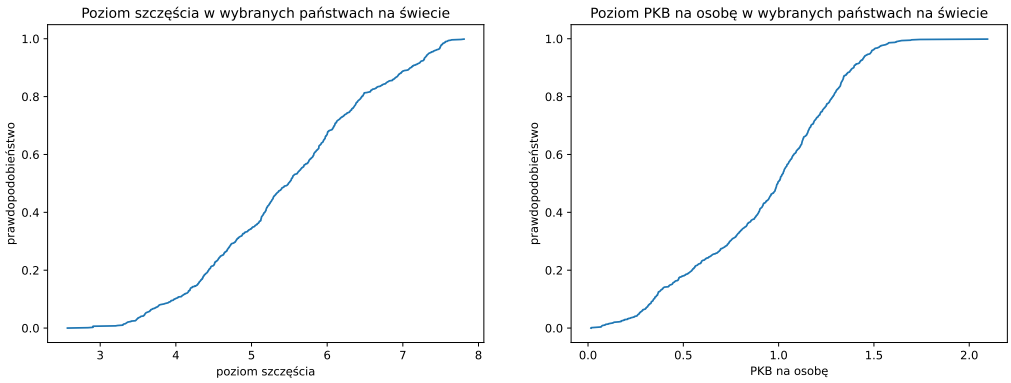
W przeprowadzanej przez nas analizie, zmienną zależną jest **szczęście**, natomiast zmienną niezależną jest **PKB**. Rozkład danych możemy zobaczyć na histogramach [1]. Zauważmy, że rozkład szczęścia wydaje się bardziej symetryczny niż rozkład PKB, który sprawia wrażenie lewoskośnego. Jednakże, niestety na pierwszy rzut oka, nie przypominają nam one żadnego ze znanych rozkładów.



Rysunek 1: Histogramy

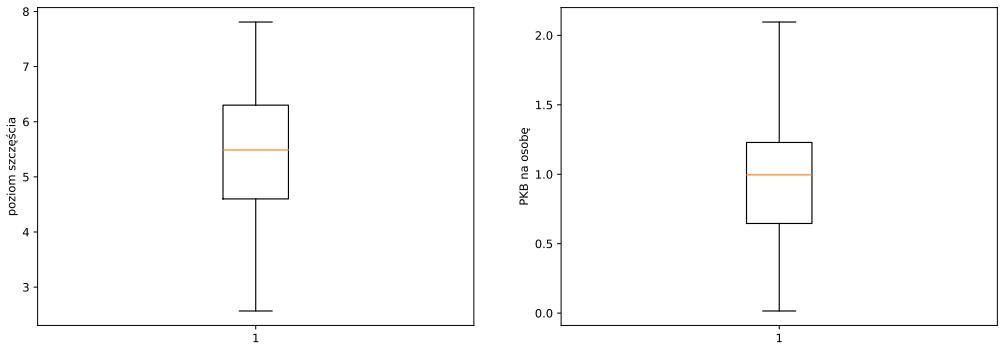
Dystrybuanty empiryczne [2] pokazują z jakim prawdopodobieństwem natrafimy na kolejno szczęście i PKB mniejsze bądź równe od danej wartości. Gdy spojrzymy na dystrybuantę szczęścia,

przypomina ona dystrybuantę rozkładu normalnego, o wariancji większej niż wariancja standardowego rozkładu normalnego. W przypadku dystrybuanty PKB, również przypomina ona rozkład normalny, jednak z przesuniętą średnią.



Rysunek 2: Dystrybuanta empiryczna

Wykres pudełkowy (ang. *boxplot*) [3] jest to graficzna reprezentacja mediany oraz kwartyli. Końce wąsów wskazują ostatnią wartość odległą od końca "pudełka" o co najwyżej półtorej wartości rozstępu międzykwartylowego. Co ciekawe, w naszych danych nie pojawiły się wartości odstające, zatem możemy sądzić, że dane zostały rzetelnie zebrane, a ryzyko przeprowadzenia błędnej analizy (nieodzwierciedlającej rzeczywistych trendów) jest mniejsze.



Rysunek 3: Boxploty

2.2 Podstawowe miary

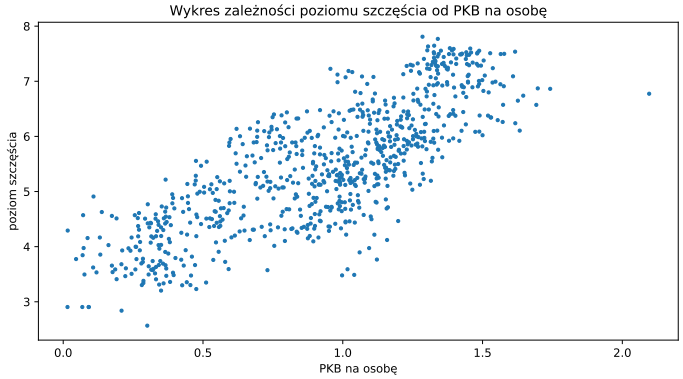
W tabeli [1] podsumowałyśmy najistotniejsze miary położenia, rozproszenia, spłaszczenia i skośności. Rzeczwiście skośność którą odczytałyśmy z histogramu dla PKB jest ujemna (zatem rozkład jest lewoskośny), natomiast skośność dla szczęścia jest niewielka. Przypuszczenia z wykresu dystrybuanty empirycznej również się sprawdziły - wariancja szczęścia jest większa od jeden.

Miary		Poziom szczęścia	PKB na osobe
położenia	średnia arytmetyczna	5.47	0.93
	średnia geometryczna	5.23	0.58
	średnia harmoniczna	5.35	0.81
	średnia ucinana 10%	5.47	0.94
	mediana Q2	5.48	0.99
	Q1	4.59	0.64
	Q3	6.30	1.22
rozproszenia	rozstęp	5.24	2.08
	rozstęp międzykwartylowy	1.70	0.58
	wariancja nieobciążona	1.26	0.14
	odchylenie standardowe	1.12	0.38
	współczynnik zmienności	20.52	41.33
spłaszczenia	kurtoza	2.25	2.31
skośności	skośność	-0.01	-0.36

Tabela 1: Zestawienie statystyk.

3 Analiza zależności liniowej pomiędzy zmienną zależną a zmienną niezależną

3.1 Wykres rozproszenia i określenie zależności



Rysunek 4: wykres rozproszenia

Na podstawie wykresu rozproszenia [4] możemy stwierdzić, że zależność pomiędzy zmiennymi prawdopodobnie jest liniowa.

3.2 Punktowa estymacja współczynników

Współczynniki regresji liniowej wyznaczamy ze wzorów

$$\begin{cases} \beta_1 = r\frac{S_y}{S_x} \\ \beta_0 = \bar{y} - r\beta_1\bar{x} \end{cases},$$

gdzie

$$r = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_x S_y},$$

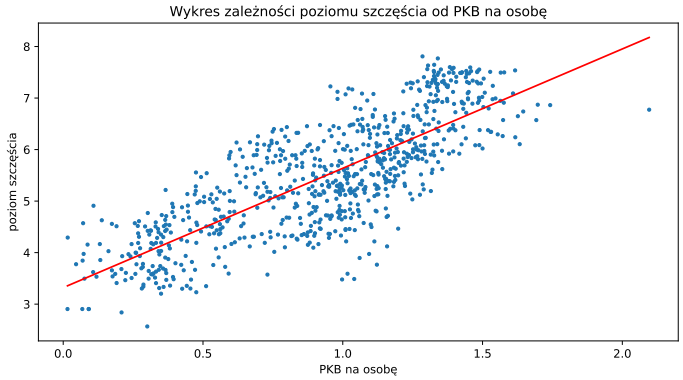
$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

$$S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Wzory te zostały wyznaczone przy pomocy metody najmniejszych kwadratów.



Rysunek 5: wykres rozproszenia z zaznaczoną prostą regresji

Z powyższych wzorów otrzymaliśmy

$$\begin{cases} \beta_1 \approx 2.32 \\ \beta_0 \approx 3.32 \end{cases}.$$

Współczynniki te opisują prostą widoczną na wykresie [5].

3.3 Przedziałowa estymacja współczynników

Z założenia o normalności rozkładu $\{\varepsilon\}_{i=1}^{n}$, przy braku znajomości jego wariancji możemy stwierdzić, że unormowane parametry β_0, β_1 mają rozkład t-studenta z n-2 stopniami swobody. Przedziały ufności przyjmują wtedy postać

$$\begin{cases} P\left(\hat{\beta}_0 - t_{n-2}(1-\frac{\alpha}{2})S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} < \beta_0 < \hat{\beta}_0 + t_{n-2}(1-\frac{\alpha}{2})S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) = 1-\alpha \\ P\left(\hat{\beta}_1 - t_{n-2}(1-\frac{\alpha}{2})\frac{S}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} < \beta_1 < \hat{\beta}_1 + t_{n-2}(1-\frac{\alpha}{2})\frac{S}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) = 1-\alpha \end{cases},$$

gdzie

$$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}}.$$

Przyjmując $\alpha = 0.05$ otrzymujemy zatem, że z 95% prawdopodobieństwem

$$\begin{cases} \beta_1 \in (2.19, 2.44) \\ \beta_0 \in (3.20, 3.44) \end{cases}.$$

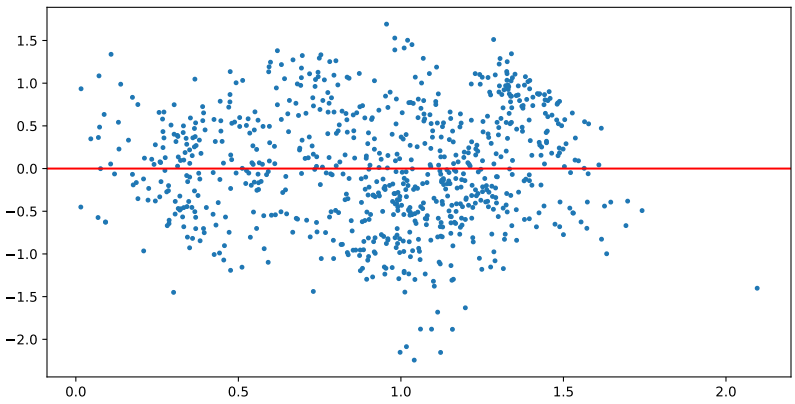
3.4 Ocena poziomu zależności

3.5 Predykcja oraz jej przedziały ufności

3.6 Interpretacja wyników

4 Analiza residuów

W tej części, przeprowadzimy analizę residuów. Dzięki temu możemy ocenić czy model regresji liniowej jest dobrym wyborem dla naszego zbioru danych. Pod nazwą residuum mamy na myśli różnicę między wyestymowaną wartością, a wartością rzeczywistą.



Rysunek 6: Residua

Na wykresie [6] przedstawione zostały residua z regresji liniowej szczęścia w zależności od PKB. Czerwona linia oznacza średnią, która jest bliska zeru. Na pierwszy rzut oka, widzimy że wariancja jest mniej więcej równa na całej długości osi zmiennej niezależnej. Aby upewnić się czy nasze przewidywanie jest słuszne, wykonamy test Levene'a jednorodności wariancji [1].

```
stat, p_value = scipy.stats.levene(random.sample(residuals,350),random.sample(
  residuals,350))
if p_value > 0.05:
  print("Wariancja jest raczej stala.")
else:
  print("Wariancja raczej nie jest stala.")

#output
Wariancja jest raczej stala.
```

Kod 1: Test Levene'a

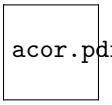
Po wykonaniu testu nie mamy podstaw aby odrzucić hipotezę zerową: "Wszystkie próbki są z populacji o równej wariancji". Teraz będziemy chciały sprawdzić czy residua są niezależne. Ponieważ na wykresie [6] trudno dostrzec zależność, wykonamy test Durbina-Watsona [2].

```
from statsmodels.stats.stattools import durbin_watson as dwtest
dwtest(resids=np.array(residuals))

#output
1.2617265161048754
```

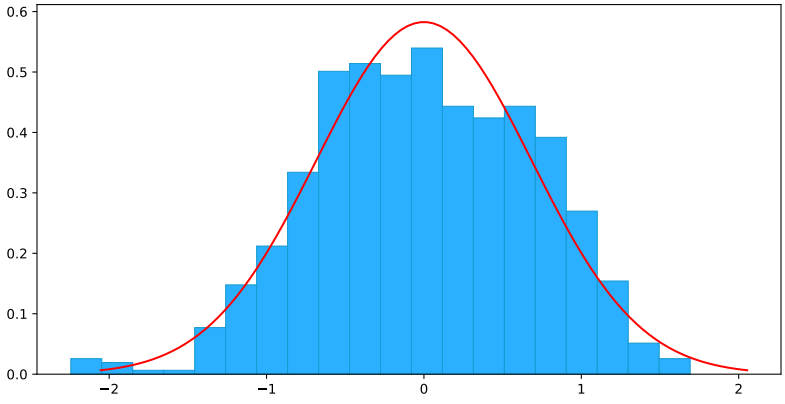
Kod 2: Test Levene'a

Jeżeli wynik jest bliski 2 oznacza to, że residua nie są skorelowane. W naszym przypadku, test sugeruje dodatnią korelację, co możemy również zobrazować na wykresie [7], który wykorzystuje funkcję `acf` z pakietu `statsmodels.tsa.stattools`, aby policzyć autokorelację.



Rysunek 7: Autokowariancja residuów

Na koniec, sprawdzimy jeszcze, czy rozkład residuów jest normalny. W tym celu porównamy ich histogram z gęstością rozkładu normalnego [8].



Rysunek 8: Rozkład residuów

Ponieważ możemy mieć wątpliwości wynikające z różnic na wykresie, wykonamy test na normalność [3], który potwierdza, że residua prawdopodobnie nie mają oczekiwanego rozkładu.

```
stat, p = scipy.stats.normaltest(residuals)
if p > 0.05:
```

```
    print('Dane raczej pochodza z rozkladu normalnego')
  else:
    print('Dane raczej nie pochodza z rozkladu normalnego')

  #output
  Dane raczej nie pochodza z rozkladu normalnego
```

Kod 3: Test Levene'a

5 Podsumowanie

Analizując przedstawione w raporcie statystyki możemy sformułować następujące wnioski i przypuszczenia dotyczące długości ogonów w populacji myszołowów rdzawosternych. Z histogramu widzimy, że badany rozkład przypomina rozkład normalny, jednak po obliczeniu skośności okazuje się, że różni się ona od skośności rozkładu normalnego, która wynosi zero. Podejrzewamy, iż może to być spowodowane licznością próby. Statystyczny myszołów rdzawosterny powinien mieć ogon o długości od 207,638 do 236.66 mm, (średnia arytmetyczna $\pm$ odchylenie standardowe). Biorąc średnią z populacji przewidujemy długość około 222,15 mm. Wartości skrajne nie wpływają znacząco na wartość średniej - wiemy to po obliczeniu średniej Winsorowskiej (222,182 mm) i ucinanej (222,104 mm).

6 Źródła

- Wykłady
- https://vincentarelbundock.github.io/Rdatasets/csv/Stat2Data/HawkTail.csv
- https://www.geo.fu-berlin.de/en/v/soga/Basics-of-statistics/Descriptive-Statistics/Measures-of-Position/index.html
- https://www.investopedia.com/terms/w/winsorized_mean.asp