Komputerowa analiza szeregów czasowych - raport 1

"Analiza danych z wykorzystaniem regresji liniowej"

Przedmiot i prowadzący:
Komputerowa analiza szeregów czasowych,
piątki 11.15 - 13.00 (grupa nr 5),
Mgr inż. Wojciech Żuławiński

Aleksandra Hodera (268733) Aleksandra Polak (268786)

Spis treści

1 Opis danych, cel

Poniżej zostały przedstawione oraz omówione dane, które analizowałyśmy w raporcie. Można je znaleźć pod poniższym linkiem: Dane dotyczące wskaźnika szczęścia w poszczególnych państwach. Na podstawie ich analizy starałyśmy się zidentyfikować czynniki, które w najbardziej znaczący sposób wpływają na poziom szczęścia obywateli danego państwa.

1.1 Przedstawienie i opis zbioru danych

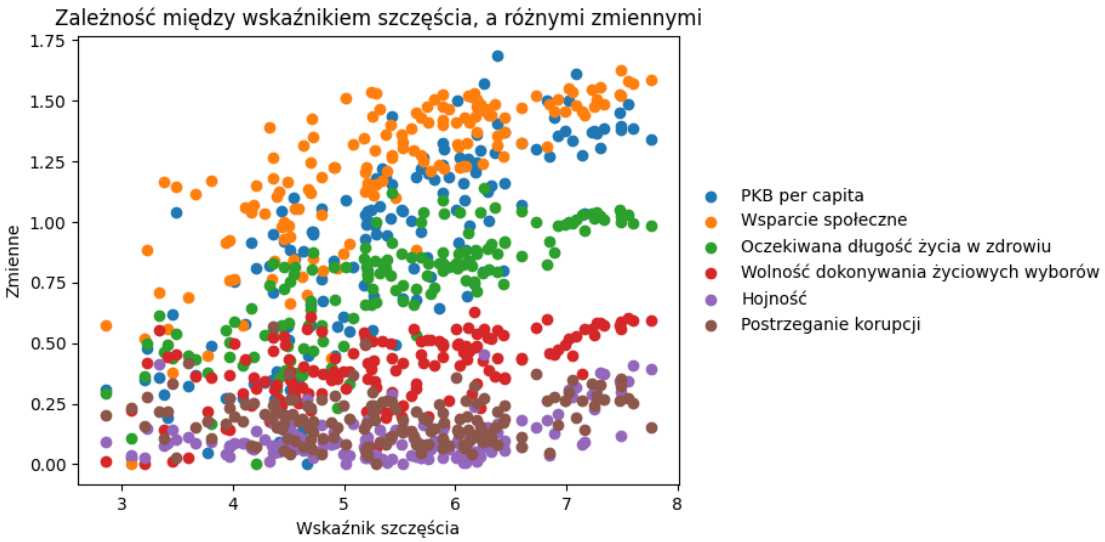
W naszym raporcie analizowałyśmy dane obrazujące związek między wskaźnikiem szczęścia (Score) w 2019 roku, a zestawem 6 niezależnych zmiennych wpływających na pozycję danego państwa w rankingu. W zestawieniu tym zostało ujętych 156 państw, które zostały uporządkowane od najbardziej do najmniej "szczęśliwych"(Overall rank).

1.2 Opis oraz analiza poszczególnych czynników

W zestawieniu zostało ujętych 6 niezależnych zmiennych mających wpływ na poziom wskaźnika szczęścia w danym państwie. Są to:

- GDP per capita, PKB per capita - Produkt Krajowy Brutto na 1 mieszkańca,
- Social support, Wsparcie społeczne - wsparcie społeczne udzielane mieszkańcom przez państwo,
- Healthy life expectancy, Oczekiwana długość życia w zdrowiu,
- Freedom to make life choices, Wolność dokonywania życiowych wyborów - wskaźnik postrzegania wolności i możliwości dokonywania własnych wyborów,
- Generosity, Hojność - wskaźnik hojności i życzliwości mieszkańców,
- Perceptions of corruption, Postrzeganie korupcji.

Na poniższym wykresie przedstawiłyśmy wpływ poszczególnych czynników na wartość wskaźnika szczęścia.

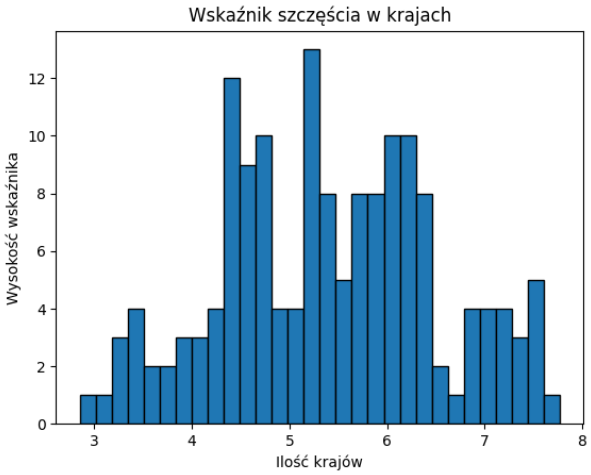


Rys. 1

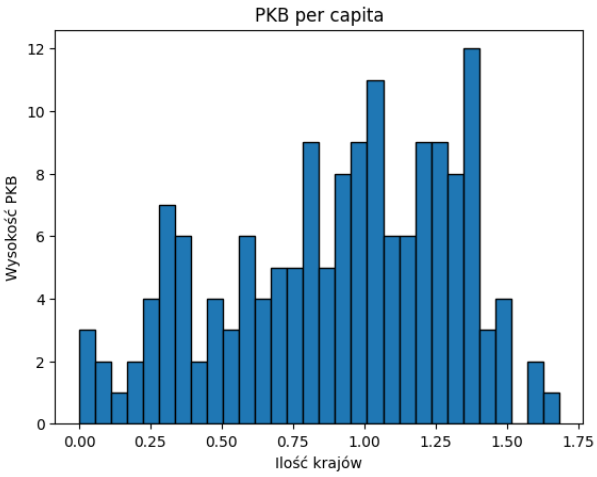
1.2.1 Uzasadnienie wyboru PKB per capita, jako głównego czynnika warunkującego wskaźnik szczęścia

Jak możemy zauważyć na wykresie 1, czynnikami, które w najbardziej znaczącym stopniu wpływają na pozycję państwa w rankingu są PKB per capita oraz wsparcie społeczne. W naszym raporcie postanowiłyśmy zająć się analizą pierwszego z tych czynników, ponieważ zauważamy, że zachowuje się on w sposób najbardziej zbliżony do liniowego. Cała poniższa analiza danych opiera się na badaniu wpływu PKB per capita na wskaźnik poziomu szczęścia.

1.3 Wykresy danych



Rys. 2



Rys. 3

Na rysunkach numer 2 i 3 zostały zwizualizowane odpowiednio dane dotyczące wskaźnika szczęścia i PKB per capita w zależności od liczby krajów, w których dana wartość jest osiągana. Na tej podstawie spróbowałyśmy dopasować teoretyczne rozkłady. Wyniki widoczne są poniżej.

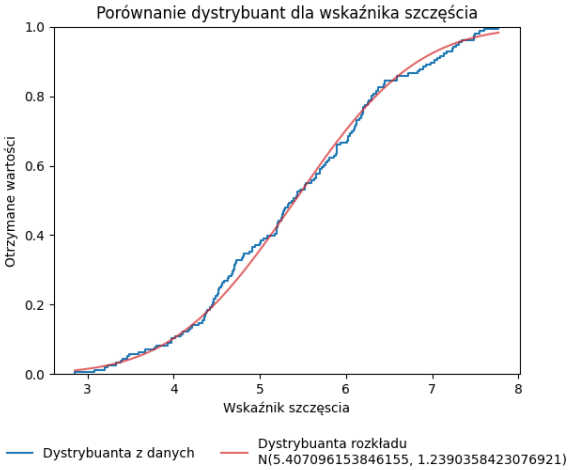
1.4 Dopasowanie rozkładów do danych

Na podstawie analizy wyglądu histogramu dopasowałyśmy następujące rozkłady:

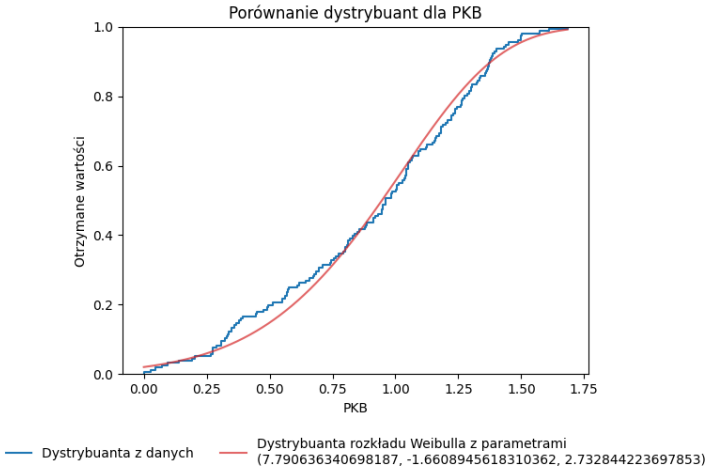
- dla wskaźnika szczęścia:
  rozkład normalny ze średnią 5.407096153846155 i wariancją 1.2390358423076921
- dla PKB per capita:
  rozkład Weibulla z parametrami 7.790636340698187, -1.6608945618310362, 2.732844223697853

Następnie porównałyśmy dystrybuanty empiryczne i teoretyczne oraz histogramy i teoretyczne gęstości. Wyniki są widoczne na następnej stronie.

1.4.1 Porównanie dystrybuanty teoretycznej z empiryczną

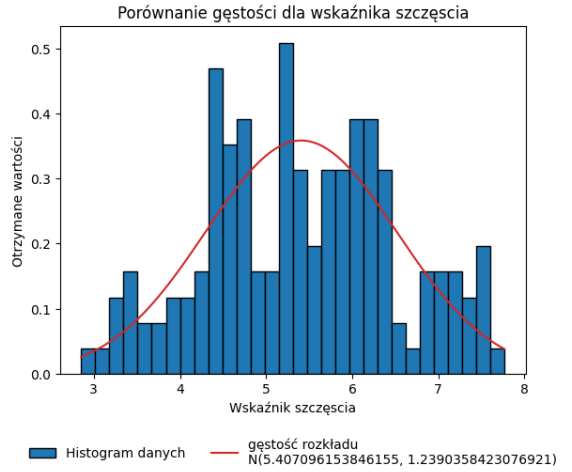


Rys. 4

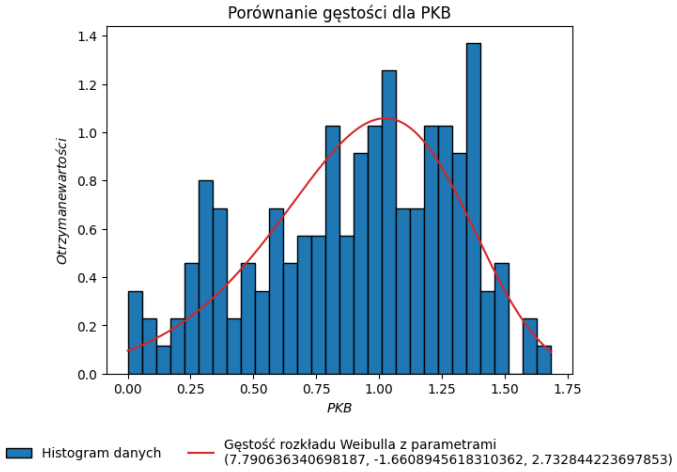


Rys. 5

1.4.2 Porównanie histogramu z gęstością teoretyczną



Rys. 6



Rys. 7

1.4.3 Test Kołmogorowa-Smirnowa dla rozkładu normalnego

Dodatkowo, dla danych opisujących wskaźnik szczęścia przeprowadziłyśmy test Kołmogorowa-Smirnowa. Jego hipoteza zerowa pozwala nam stwierdzić, że otrzymując p-wartość $> \alpha = 0.05$ mamy do czynienia z rozkładem normalnym. W naszym przypadku otrzymałyśmy p-wartość $\approx 0.653 > 0.5$, czyli test ten potwierdził, że poprawnie dobrałyśmy rozkład.

Na podstawie wykresów numer 4, 5, 6, 7 oraz wyniku testu KS możemy wnioskować, że wskazane przez nas rozkłady są najprawdopodobniej prawidłowo dobrane oraz, że nasza metoda estymacji parametrów tych rozkładów jest prawidłowa.

2 Statystyki opisowe

2.1 Najważniejsze pojęcia, definicje, wzory

2.1.1 Średnia arytmetyczna

Jest to iloraz sumy wszystkich zmiennych i rozmiaru próby. Zwana również wartością średnią lub wartością oczekiwaną.

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

2.1.2 Średnia ucinana

Obliczając ją, w pierwszej kolejności porządkujemy zmienne od najmniejszej do największej, a następnie odrzucamy k skrajnych obserwacji (tyle samo na obu krańcach). Z pozostałych zmiennych obliczamy średnią arytmetyczną.

$$\frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_i$$

2.1.3 Średnia winsorowska

Obliczając ją, w pierwszej kolejności porządkujemy zmienne od najmniejszej do największej, a następnie k skrajnych obserwacji (tyle samo na obu krańcach) zastępujemy odpowiednio wartością minimalną lub maksymalną z pozostałych danych. Z tych zmiennych obliczamy średnią arytmetyczną.

$$\frac{1}{n}\left[(k+1)x_{k+1} + \sum_{i=k+2}^{n-k-1} x_i + (k+1)x_{n-k}\right]$$

2.1.4 Wariancja

Mówi nam o tym jak duże jest zróżnicowanie zmiennych w próbie, jak bardzo obserwacje odbiegają od średniej. Im mniejsza wariancja, tym obserwacje są bardziej skupione wokół średniej.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad lub \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

2.1.5 Odchylenie standardowe

Jest to pierwiastek kwadratowy z wariancji. Jego interpretacja jest analogiczna do interpretacji wariancji.

$$S = \sqrt{S^2}$$

2.1.6 Mediana

Mediana (kwartyl Q_2) – jest zwana również wartością środkową zbioru, dzieli zbór na dwie równe części. Jest to wartość od której połowa zmiennych jest mniejsza, a druga połowa większa.

$$x_{med} = \begin{cases} x_{\frac{n+1}{2}} \ , \ gdy\ n\ jest\ nieparzyste \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) \ , \ gdy\ n\ jest\ parzyste \end{cases}$$

2.1.7 Kwartyle

Kwartyle (rzędu pierwszego: Q_1 oraz rzędu trzeciego: Q_3) – dzielą zbiór na cztery równe części. Kwartyl Q_1 to mediana obserwacji mniejszych od Q_2, kwartyl Q_3 to mediana obserwacji większych od Q_2.

2.1.8 IQR - rozstęp międzykwartylowy

Jest różnica między kwartylem rzędu trzeciego (Q_3), a kwartylem rzędu pierwszego (Q_1). Mówi on nam o zmienności 50 środkowych danych, nie uwzględnia wartości mniejszych od Q_1 i większych od Q_3.

$$IQR = Q_3 - Q_1$$

2.1.9 Wartość minimalna i maksymalna

Odpowiednio najmniejsza (x_{min}) i największa (x_{max}) zaobserwowana wartość.

2.1.10 Rozstęp z próby

Jest to różnica wartości maksymalnej i wartości minimalnej próby

$$x_{max} - x_{min}$$

2.1.11 Odchylenie przeciętne od wartości średniej

Jest to iloraz sumy wartości bezwzględnej z różnicy kolejnych obserwacji oraz średniej arytmetycznej z próby i rozmiaru próby.

$$d_1 = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$$

2.1.12 Współczynnik zmienności

Mówi nam o zróżnicowaniu obserwacji. Im wyższa wartość współczynnika zmienności tym dane są bardziej zróżnicowane, im niższa tym dane są bardziej jednorodne. ν większe od 60 oznacza dużą zmienność.

$$\nu = \frac{S}{\overline{x}} * 100\%$$

2.1.13 Współczynnik skośności (asymetrii)

Mówi nam o symetrii/asymetrii próby. Wyróżniamy symetrię prawostronną, lewostronną oraz brak asymetrii. Współczynnik skośności o wartości bliskiej 0 świadczy o braku asymetrii rozkładu, wartość powyżej 0 - o prawostronnej asymetrii, a wyniki poniżej 0 – o lewostronnej asymetrii rozkładu.

$$\alpha_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{\frac{3}{2}}} \quad lub \quad \alpha_2 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{S}\right)^3$$

2.1.14 Kurtoza

Jest to miara kształtu próby, ilości skrajnych wartości. Mówi nam o tym, w jakim stopniu obserwacje są skoncentrowane wokół średniej. Punktem odniesienia jest rozkład normalny, dla którego kurtoza wynosi 3. Gdy wartość kurtozy jest większa od 3 mówimy o rozkładach ciężkoogonowych, natomiast mniejsza od 3 to rozkłady lekkoogonowe.

$$K_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} \quad lub \quad K_2 = \frac{n-1}{(n-2)(n-3)}\left[(n+1)K_1 - 3(n-1)\right] + 3$$

2.1.15 Współczynnik korelacji

Współczynnik korelacji Pearsona pomiędzy każdą parą wartości można wyliczyć ze wzoru:

$$r_{xy} = \frac{\sum_{i-1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i-1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i-1}^{n}(y_i - \overline{y})^2}}$$

Wartości współczynnika korelacji mieszczą się w zakresie pomiędzy -1, a 1. Wartości zbliżone do 1 wskazują silną korelację dodatnią, natomiast wartości zbliżone do -1 silną korelację ujemną. Wartości bliskie zera wskazują korelację zerową.

2.2 Analiza jednowymiarowa dla poszczególnych zmiennych - podstawowe statystyki

W poniższej tabeli przedstawiłyśmy wartości statystyk opisowych dla wskaźnika szczęścia oraz PKB per capita. Wszystkie wartości zostały zaokrąglone do 3 miejsc po przecinku.

Statystyka opisowa	Wskaźnik szczęścia	PKB per capita
Średnia arytmetyczna	5.407	0.905
Wariancja	1.231	0.158
Odchylenie standardowe	1.110	0.397
Mediana	5.380	0.960
Kwartyl Q_1	4.541	0.595
Kwartyl Q_3	6.187	1.234
IQR	1.6460	0.640
Wartości minimalna	2.853	0.000
Wartość maksymalna	7.769	1.684
Rozstęp z próby	4.916	1.684
Odchylenie przeciętne od wartości średniej	0.917	0.333
Współczynnik zmienności	20.520	43.872
Współczynnik skośności	0.011	-0.382
Kurtoza	2.373	2.216

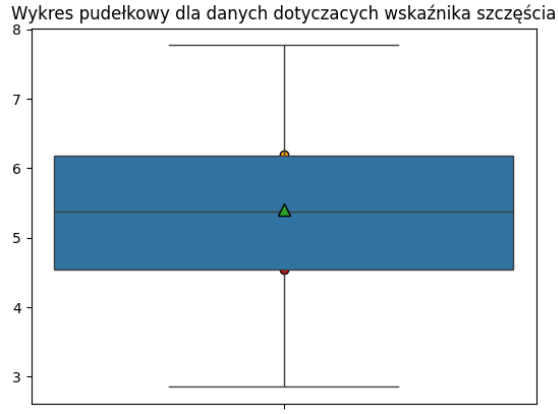
Tabela 1

Dodatkowo obliczyłyśmy współczynnik korelacji tych 2 zmiennych. Wyniósł on: $r_{xy} \approx 0.794$

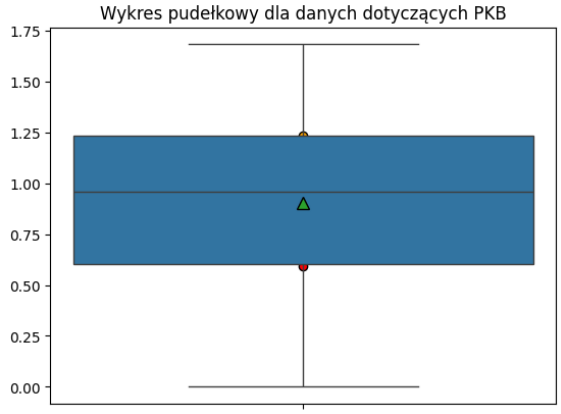
2.3 Analiza jednowymiarowa dla poszczególnych zmiennych - wykresy

2.3.1 Wykresy pudełkowe

Dodatkowo narysowałyśmy wykresy pudełkowe dla wskaźnika szczęścia oraz PKB per capita.



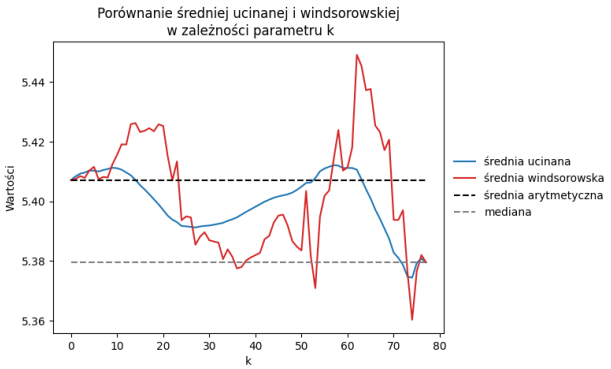
Rys. 8



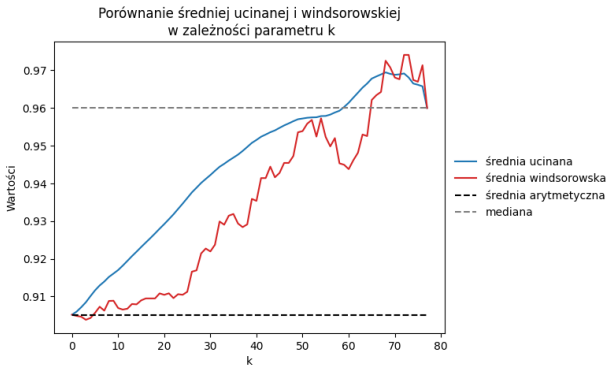
Rys. 9

2.3.2 Porównanie wartości poszczególnych średnich

W tej części porównałyśmy wartości średniej ucinanej i winsorowskiej w zależności od parametru k, dla wskaźnika szczęścia i PKB per capita. Dla lepszego zwizualizowania na wykresach naniosłyśmy dodatkowo średnie arytmetyczne i mediany.



Rys. 10



Rys. 11

2.4 Podsumowanie - statystyki opisowe

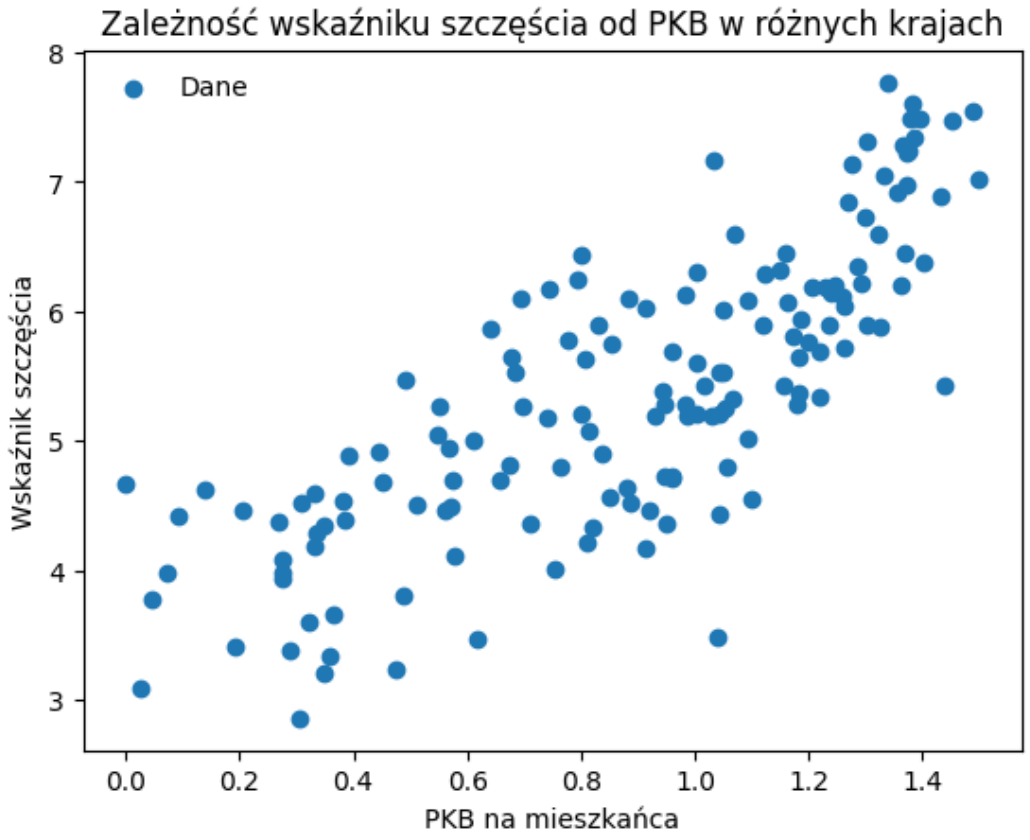
W tabeli 1 zostały przedstawione podstawowe statystyki opisowe. Na podstawie części z nich zostały również sporządzone wykresy numer 8-11. Analizując współczynnik skośności możemy wnisokować, że dane dotyczące wskaźnika szczęścia wykazują brak asymetrii, natomiast te opisujące PKB per capita są lewostronnie skośne (asymetria ujemna). Kurtoza w obu przypadkach jest poniżej 3, dlatego możemy wnioskować, że mamy do czynienia z rozkładami lekkoogonowymi. Z wykresów 7 i 8 możemy odczytać, że kwartyle rzędu pierwszego Q_1 (czerwone kropki) i trzeciego Q_3 (żółte kropki) pokrywają się z liniami wyznaczającymi krańce prostokąta, podobnie jest

w przypadku kwartylu rzędu drugiego Q_2, czyli mediany (zielone trójkąty). Na obu wykresach nie widać także obserwacji odstających, czyli, nawet jeśli występują, to nie odbiegają w znaczący sposób. Na wykresach numer 10 i 11 możemy zauważyć, że średnia winsorowska oraz ucinana oscylują w okolicach mediany oraz średniej arytmetycznej. Patrząc na wykres 11, dotyczący PKB per capita, widzimy, że czym większa wartość k, tym obie średnie bardziej zbliżają się do mediany i oddalają od średniej arytmetycznej. W przypadku wykresu 10, dotyczącego wskaźnika szczęścia, nie istnieje taka zależność.

3 Regresja liniowa

3.1 Regresja liniowa - definicja

Regresja liniowa jest najprostszym przykładem regresji. Zakłada ona, że zależność pomiędzy zmienną objaśniającą (PKB per capita), a objaśnianą (wskaźnik szczęścia) jest liniowa. Analiza regresji liniowej opiera się na znalezieniu takich współczynników b_0 i b_1, aby prosta $y = b_0 + b_1 x$, jak najlepiej dopasowywała się do danych. Oznacza to, że dążymy do tego, żeby błędy dopasowania e_i były jak najmniejsze. W celu sprawdzenia czy istnieje zależność liniowa pomiędzy PKB per capita, a wskaźnikiem szczęścia narysowaliśmy wykres zależności tych dwóch zmiennych.



Rys. 12

Wyniki przedstawione na wykresie numer 12 pozwalają nam przypuszczać, że do naszych danych da się dobrać prostą regresji. Jest to zgodne z założeniami, które poczyniłyśmy na samym początku.

3.2 Metoda najmniejszych kwadratów

Chcąc wyznaczyć wartości b_0 i b_1 korzystamy z metody najmniejszych kwadratów. Jest to odpowiednik metody największej wiarygodności dla e_i z rozkładu normalnego. Otrzymane parametry mają następującą postać:

$$\begin{cases} b_0 = \overline{y} - b_1\overline{x} \\ b_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n}(y_i-\overline{y})x_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \end{cases}$$

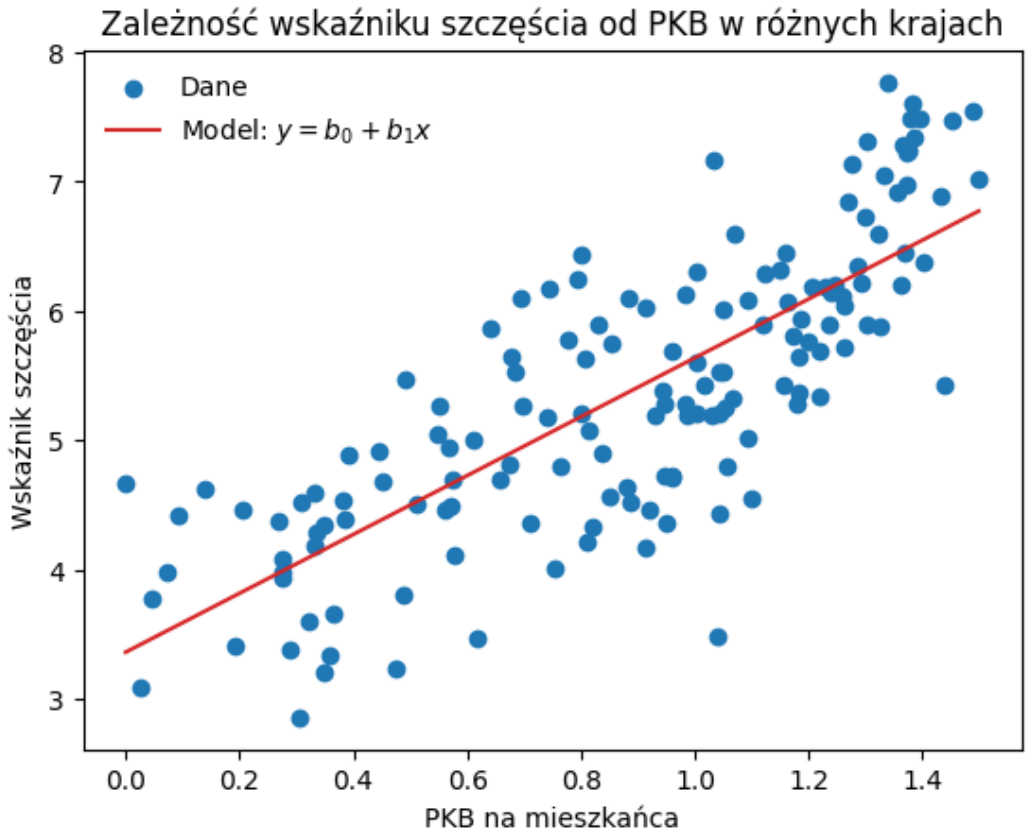
W naszym przypadku:

$$\begin{cases} b_0 \approx 3.359 \\ b_1 \approx 2.278 \end{cases}$$

Natomiast błędy e_i mają postać:

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

Aby sprawdzić czy prosta o współczynnikach b_0 i b_1 dopasowuje się do naszych danych naniosłyśmy prostą regresji na wykres numer 12.



Rys. 13

Patrząc na wykres numer 13 możemy wnioskować, że parametry b_0 i b_1 najprawdopodobniej zostały poprawnie wyznaczone. W celu upewnienia się co do prawdziwości tej tezy, w dalszej części raportu policzymy współczynnik determinancji R^2 oraz przyjrzymy się błędom.

3.3 R^2 - wyznaczenie oraz analiza wartości

W celu upewnienia się czy prosta regresji została poprawnie dobrana obliczyłyśmy współczynnik determinancji R^2. Statystyka ta mówi nam o tym, jaki procent punktów leży na wyznaczonej przez nas prostej regresji. Czym jest bliższy 1, tym prosta jest lepiej dopasowana.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{y})^2}{\sum_{i=1}^{n}(y - \overline{y})^2} = 0.879$$

Otrzymaliśmy wartość zbliżoną do 1, więc możemy wnioskować, że prosta regresji jest poprawnie dobrana.

3.4 Teoretyczny model regresji liniowej

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

gdzie:

- $x_1, ..., x_n$ - wielkości deterministyczne,
- β_0, β_1 - parametry modelu, wielkości deterministyczne,
- ϵ_i - zmienne losowe, ich realizacje to e_i.

Natomiast estymatory $\hat{\beta_0}$ i $\hat{\beta_1}$ mają postać:

$$\begin{cases} \hat{\beta_0} = \overline{Y} - \hat{\beta_1}\overline{x} \\ \hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{cases}$$

Tutaj również estymatory $\hat{\beta_0}$ i $\hat{\beta_1}$ uzyskane metodą najmniejszych kwadratów są takie same jak estymatory uzyskane metodą największej wiarygodności przy założeniu, że $\epsilon_i \sim \mathrm{N}(0, \sigma^2) \quad \forall i$. Postać oraz założenia dotyczące ϵ_i również zostaną przedstawione w dalszej części.

4 Przedziały ufności

4.1 Najważniejsze pojęcia, definicje, wzory

4.1.1 Poziom ufności/istotności α

Jest to przyjęte odgórnie, dopuszczalne ryzyko odrzucenia prawdziwej hipotezy zerowej. Poziom ten oznaczamy jako α.

4.1.2 Przedziały ufności

Przedziały ufności zostały wyznaczone przy założeniu, że wariancja nie jest znana. Dlatego skorzystamy z poniższej statystyki testowej:

$$T = \frac{\hat{\beta_1} - \beta_1}{\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} \sim \mathrm{T}_{n-2},$$

gdzie:

- T_{n-2} to rozkład T-Studenta z $n-2$ stopniami swobody,
- $s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y})}{n-2},\ n > 2$

Chcemy znaleźć takie wartości A i B, żeby spełnione było równanie:

$$P(\beta_1 \in [A, B]) = 1 - \alpha$$

$$A = \hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$
$$B = \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}} \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}},$$

gdzie: $t_{n-2,1-\frac{\alpha}{2}}$ to kwantyl rzędu $1 - \frac{\alpha}{2}$ rozkładu T-studenta z $n-2$ stopniami swobody.

Czyli przedział ufności na poziomie ufności $1-\alpha$ dla β_1 ma postać:

$$\left[\hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}} \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right]$$

Analogiczne obliczenia przeprowadziłyśmy dla β_0 i otrzymałyśmy przedział ufności:

$$\left[\hat{\beta}_0 - t_{n-2,1-\frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \hat{\beta}_0 + t_{n-2,1-\frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right]$$

4.2 Otrzymane wyniki

Podstawiając nasze dane do powyższych wzorów, dla $\alpha = 0.05$ (poziom ufności to $1 - \alpha = 0.95$) otrzymałyśmy następujące przedziały ufności:

- dla β_0:

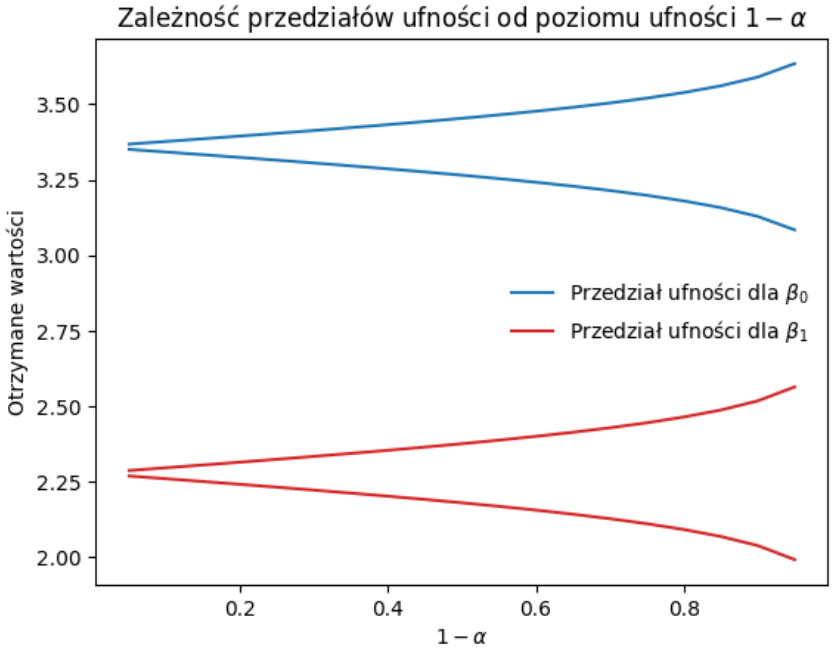
$$A = 3.08384281612917,$$
$$B = 3.63427077478727,$$
$$\beta_0 \in [3.08384281612917, 3.634270774787275]$$

- dla β_1:

$$A = 1.9920360810674222,$$
$$B = 2.5636683309145547,$$
$$\beta_1 \in [1.9920360810674222, 2.5636683309145547]$$

4.3 Przedziały ufności w zależności od poziomu istotności

Analizując nasze dane sprawdziłyśmy także jak zmieniają się przedziały ufności w zależności od poziomu istotności. Wyniki przedstawiłyśmy na wykresie numer 14.



Rys. 14

Możemy zauważyć, że wraz ze wzrostem poziomu ufności $1-\alpha$ (spadkiem wartości α) rosną przedziały ufności.

5 Analiza residuów

5.1 Residua - definicja, założenia

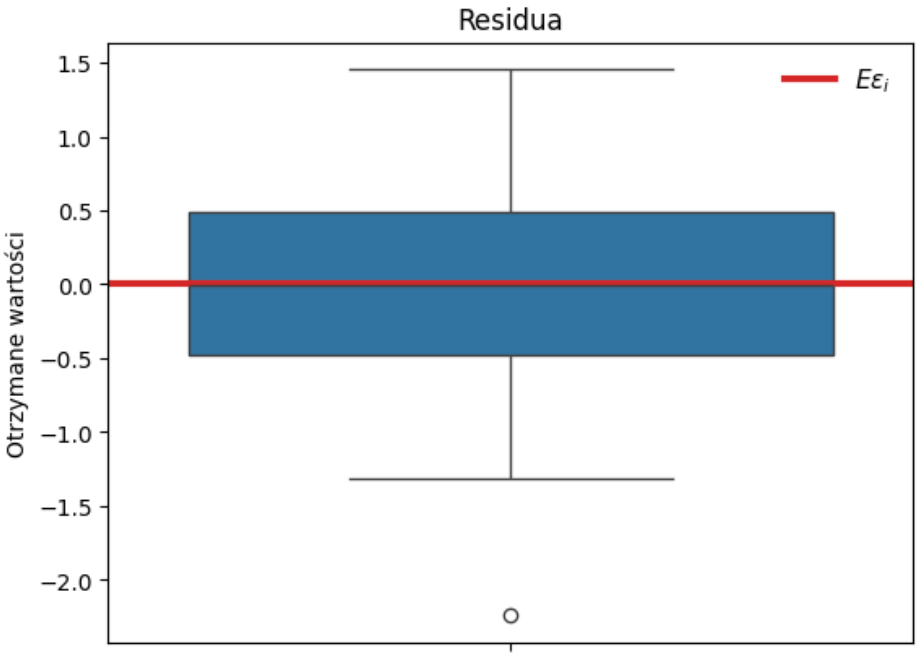
Residua to inaczej reszty. Pozwalają nam one określić czy prosta regresji jest dobrze dopasowana do danych.

Założenia dotyczące residuów:

1. $\epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad \forall i = 1, ..., n,$
2. $\epsilon_1, ...\epsilon_n$ - niezależne (nieskorelowane),
3. $E\epsilon_i = 0, \quad \forall i = 1, ..., n,$
4. $Var\epsilon_i = \sigma^2, \quad \forall i = 1, ..., n.$

5.2 Wizualizacja wyników - wykresy residuów

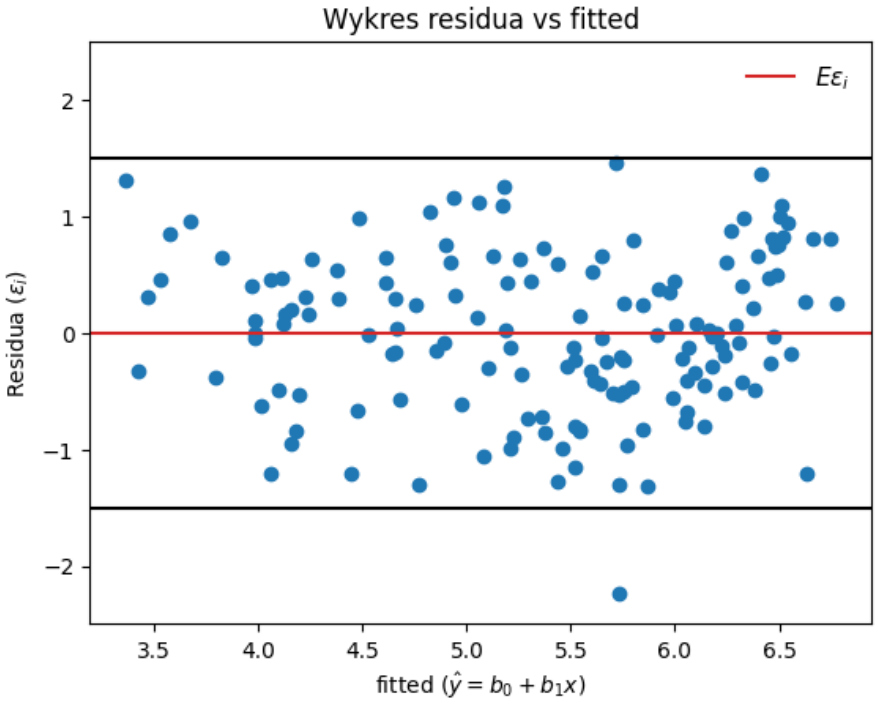
W celu lepszego zobrazowania wykonałyśmy wykres pudełkowy.



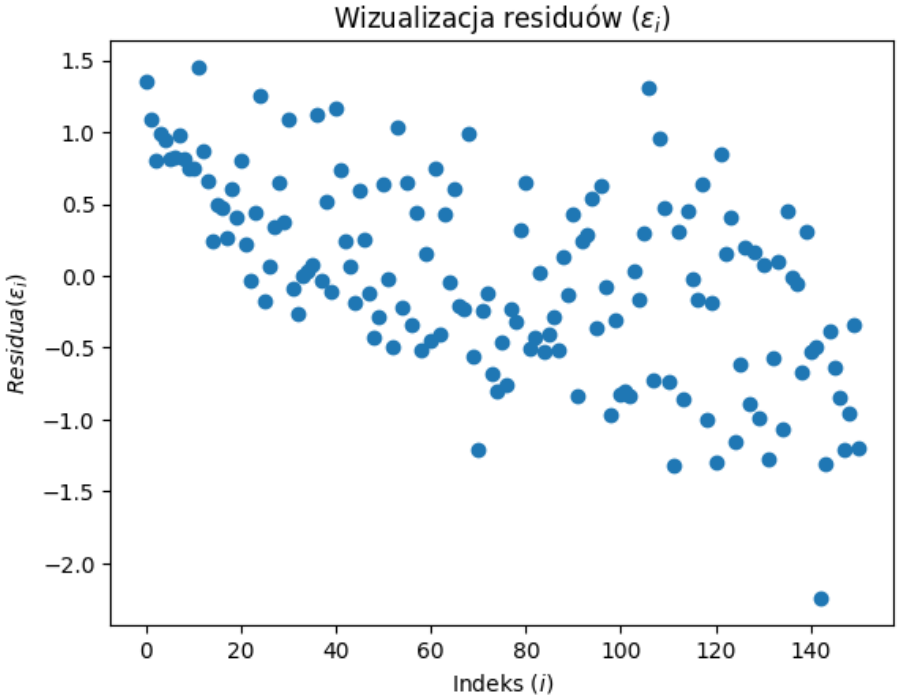
Rys. 15

Na wykresie numer 15 możemy zauważyć, że występuje jedna obserwacja odstająca. Ma ona wartość około -2.242 i dotyczy Madagaskaru.

Dodatkowo sporządziłyśmy 2 wykresy obrazujące zachowanie residuów. Są one widoczne na następnej stronie.


Wykres residua vs fitted
$E\varepsilon_i$
Residua (ε_i)
fitted ($\hat{y} = b_0 + b_1 x$)

Rys. 16


Wizualizacja residuów (ε_i)
Residua(ε_i)
Indeks (i)

Rys. 17

5.3 Sprawdzenie założeń

1. $\epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad \forall i = 1, ..., n$

   W celu sprawdzenia czy $\epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad \forall i = 1, ..., n$ przeprowadziliśmy testy: Kołmogorowa-Smirnowa, Shapiro-Wilka, Jarque-Bera, a także porównaliśmy dystrybuanty empiryczną i teoretyczną oraz histogram residuów i gęstość teoretyczną. Dodatkowo założyliśmy, że ϵ_i są iid $\forall i$.

   - test Kołmogorowa-Smirnowa

     Pomocniczo obliczyłyśmy odchylenie standardowe ze wzoru 2.1.5:

     $$s \approx 0.680$$

     W wyniku przeprowadzonego testu otrzymałyśmy p-wartość ≈ 0.934.

   - test Shapiro-Wilka

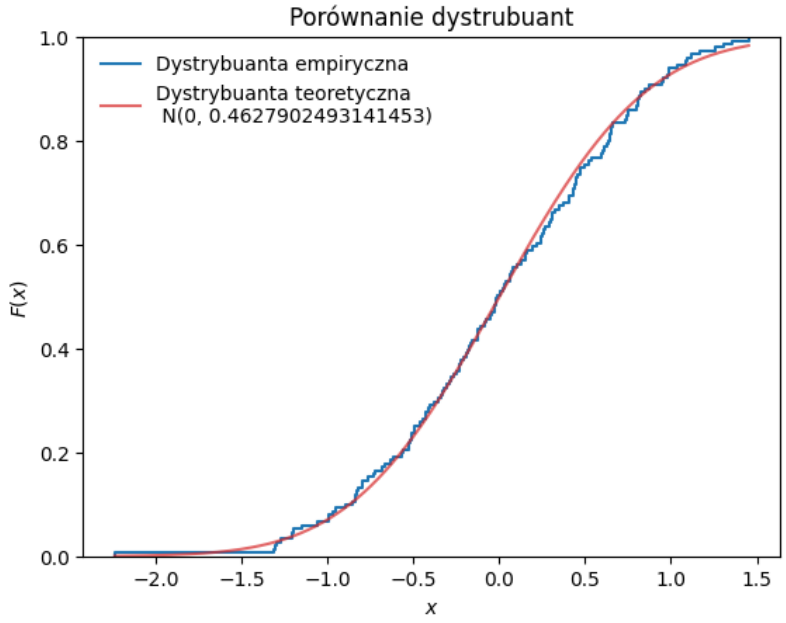
     W wyniku przeprowadzonego testu otrzymałyśmy p-wartość ≈ 0.350.

   - test Jarque-Bera

     W wyniku przeprowadzonego testu otrzymałyśmy p-wartość ≈ 0.466.

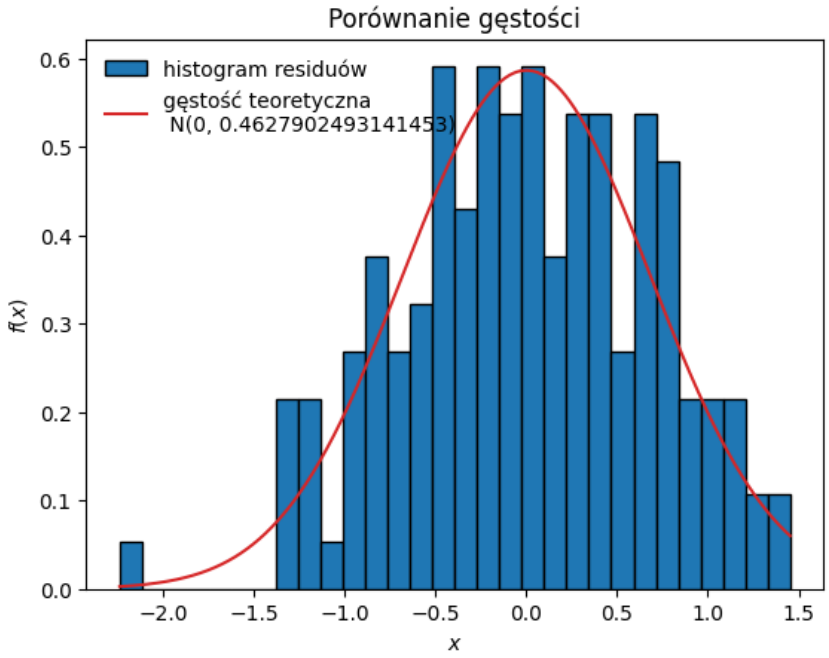
     Hipoteza zerowa tych testów pozwala nam stwierdzić, że otrzymując p-wartość $> \alpha = 0.05$ mamy do czynienia z rozkładem normalnym. Możemy zauważyć, że we wszystkich 3 przypadkach p-wartość jest znacznie większa od $0.05 = \alpha$, czyli możemy wnioskować, że mamy do czynienia z rozkładem normalnym.

   - Porównanie dystrybuant



Rys. 18

- Porównanie histogramu i gęstości teoretycznej



Rys. 19

Analizując wykresy numer 18 i 19 możemy wywnioskować, że $\epsilon_i \sim \mathrm{N}(0, 0.4627902493141453)$. Potwierdza nam to założenie, że $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$.

Zarówno wszystkie 3 testy, jak i porównanie dystrybuant oraz histogramu z gęstością teoretyczną pozwala nam stwierdzić, że założenie 1) jest spełnione.

2. $\epsilon_1, ...\epsilon_n$ - niezależne (nieskorelowane)

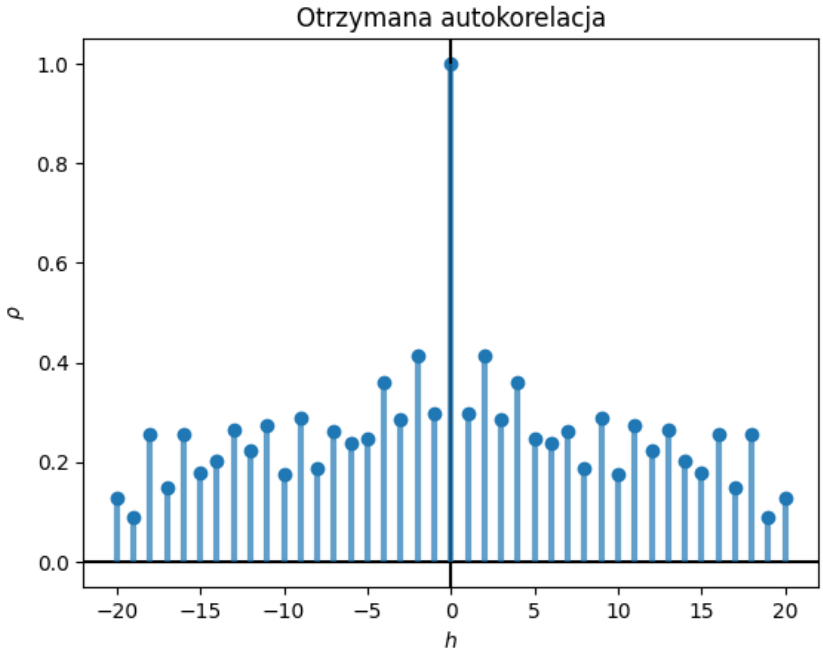
W celu sprawdzenia czy residua są nieskorelowane sprawdzimy zachowanie funkcji autokorelacji. W tym celu skorzystamy z funkcji autokowariancji:

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} \left(x_{t+|h|} - \overline{x}\right)\left(x_t - \overline{x}\right), \quad -n < h < n$$

Następnie na tej podstawie policzymy funkcję autokorelacji:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

Obliczone wartości naniosłyśmy na wykres widoczny na następnej stronie.



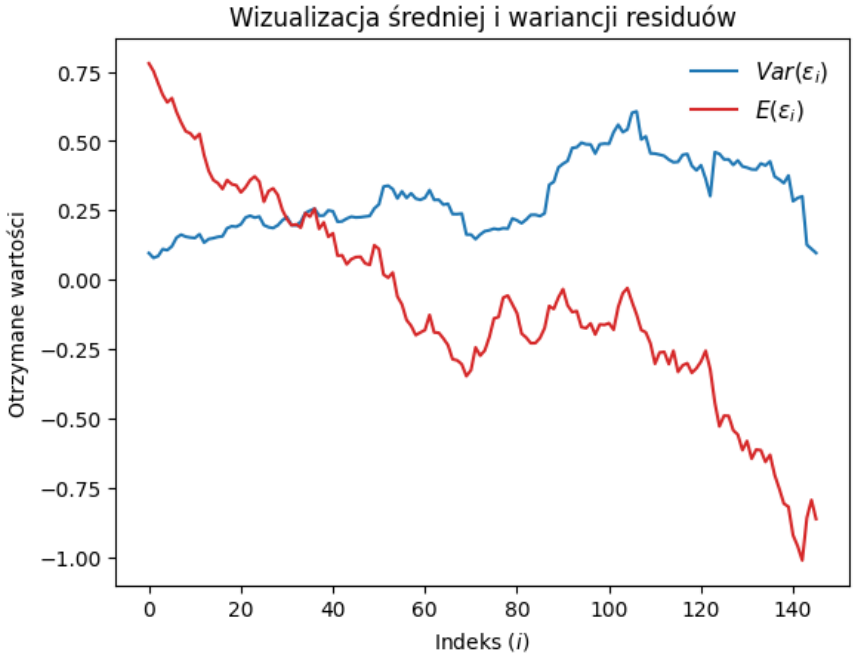
Rys. 20

Analizując wykres numer 20 widzimy, że autokorelacja dla wszystkich h jest różna od 0, czyli zmienne są skorelowane. Wynika stąd, że założenie 2) nie jest spełnione.

3. $E\epsilon_i = 0, \quad \forall i = 1, ..., n$

4. $Var\epsilon_i = \sigma^2, \quad \forall i = 1, ..., n$

   Analizując wykres numer 16 możemy zauważyć, że występuje 1 obserwacja odstająca, co jest zgodne z naszą wcześniejszą obserwacją. Widzimy również, że wariancja utrzymuje się na, w miarę, stałym poziomie (założenie 4) oraz ze średnia oscyluje wokół 0 (założenie 3). Natomiast patrząc na wykres numer 17 widzimy, że wariancja rośnie wraz ze wzrostem indeksu (zaprzeczenie założenia 4), natomiast średnia maleje (zaprzeczenie założenia 3), aby potwierdzić nasze obserwacje sporządziłyśmy wykres wariancji i średniej ruchomej - analizowałyśmy zachowanie residuów w kolejnych przedziałach o długości 20. Wyniki zostały przedstawione na wykresie numer 21, widocznym na następnej stronie.



Rys. 21

Wykres ten potwierdza nasze założenia na temat tego, że średnia maleje wraz ze wzrostem indeksu (czyli nie jest stale równa 0), wariancja natomiast jest zmienna w zależności od indeksu (czyli nie spełnia założenia, że jest stała). Na tej podstawie możemy wnioskować, że residua naszego modelu nie spełniają założenia numer 1 i 2 przedstawionych w punkcie 5.1.

Podsumowując, nie wszystkie założenia dotyczące residuów są spełnione. Możemy zauważyć, że residua mają rozkład $\epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad \forall i = 1, ..., n$, więc założenie 1) jest spełnione. Analizując zachowanie funkcji autokorelacji widzimy, że zmienne są skorelowane, czyli założenie 2) nie jest spełnione. Wnioski na temat spełnienia założenia 3) i 4) zależą od analizowanego wykresu.

5.4 Obserwacje odstające

Tak, jak zostało wspomniane wcześniej, istnieje jedna obserwacja odstająca. Dotyczy ona Madagaskaru.

6 Predykcja

6.1 Predykcja - definicja

Predykcja to według definicji przewidywanie przyszłych realizacji zmiennych losowych bądź ich cech. W naszym przypadku, aby przeprowadzić predykcję zakładamy, że residua spełniają założenia przedstawione w punkcie 5.1. Wszystkie obserwacje dzielimy na dwie grupy w zależności od tego czy ich PKB per capita jest większe czy mniejsze niż 1.5. Działanie to spowoduje oddzielenie 5 państw z najwyższym (wyższym niż 1.5) PKB per capita. Państwa te to: Kuwejt, Luksemburg, Katar, Singapur, Zjednoczone Emiraty Arabskie. W efekcie otrzymujemy zestaw danych, na którego podstawie został już wcześniej dobrany model regresji, który następnie może być użyty do przewidywania odciętych wartości.

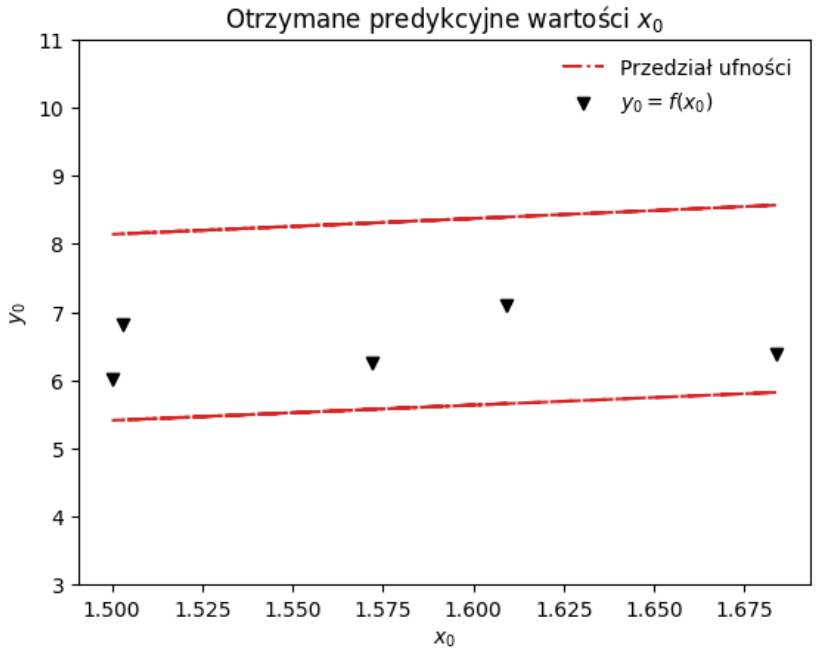
6.2 Przedziały ufności

Dla każdego z odrzuconych punktów obliczyłyśmy jego przedział ufności (definicja znajduje się w punkcie 4.1), czyli zakres wartości, w którym, z określonym prawdopodobieństwem $1-\alpha$, spodziewana jest prawdziwa wartość zmiennej.

Poniżej został przedstawiony wzór na przedział ufności dla prognozowanych wartości na poziomie istotności $1-\alpha$, dla $\alpha = 0.05$. Wyprowadzenie i oznaczenia są analogiczne do tych w punkcie 4.1.2. Zakładamy nieznaną wartość odchylenia standardowego.

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{n-2,1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}, \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{n-2,1-\frac{\alpha}{2}} s\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\right]$$

6.3 Analiza wyników dla odciętych wartości

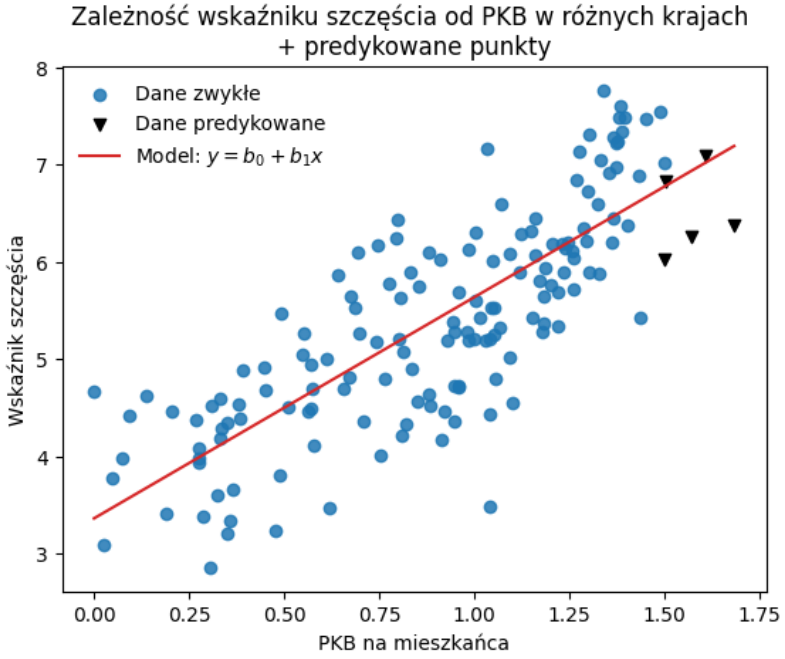


Rys. 22

Na wykresie numer 22 możemy zauważyć, że wszystkie odcięte wartości wpadają do przedziału ufności na poziomie istotności $1-\alpha$, dla $\alpha = 0.05$.

6.4 Analiza wyników dla wszystkich wartości

Następnie naniosłyśmy na wykres regresji nowe, predykowane dane.



Rys. 23

Analizując wykres numer 23 możemy zauważyć, że linia regresji, sporządzona na podstawie okrojonych danych jest dobrze dobrana także do danych odciętych, czyli możemy wnioskować, że linia regresji jest dobrze dopasowana.

7 Podsumowanie, wnioski

Zgodnie z naszymi początkowymi przewidywaniami widzimy, że istnieje liniowa zależność pomiędzy PKB per capita, a indeksem szczęścia. Dodatkowo, możemy zauważyć, że zazwyczaj, czym wyższe PKB per capita, tym wyższy wskaźnik szczęścia. Pokazuje to, że nasza teza dotycząca dużego wpływu PKB była prawdziwa. Warunki 2) - 4) dotyczące residuów w teoretycznym modelu regresji liniowej nie zostały spełnione. Jedynym warunkiem, który został spełniony jest założenie, że dane pochodzą z rozkładu $N(0, \sigma^2)$ (zakładając, że residua są iid). Analiza powyższych danych może wpłynąć na polepszenie poziomu szczęścia w państwach znajdujących się na dole zestawienia (takich jak na przykład Południowy Sudan czy Afganistan)

Źródła:

- Jacek Koronacki, Jan Mielniczuk "Statystyka dla studentów kierunków technicznych i przyrodniczych", WNT, Warszawa, 2018
- Wykłady dr hab. inż. Krzysztofa Burneckiego oraz laboratoria dr inż. Aleksandry Grzesiek z przedmiotu „Statystyka stosowana”.
- Wykłady dr hab. inż. Agnieszki Wyłomańskiej oraz laboratoria mgr inż. Wojciecha Żuławińskiego z przedmiotu „Komputerowa analiza szerwgów czasowych”.