Politechnika Wrocławska

Raport 1

Komputerowa analiza szeregów czasowych

Prowadzący kurs: mgr inż. Justyna Witulska

Aleksander Rzyhak

nr indeksu: 268766

22 grudnia 2023

1 Wstęp i opis danych

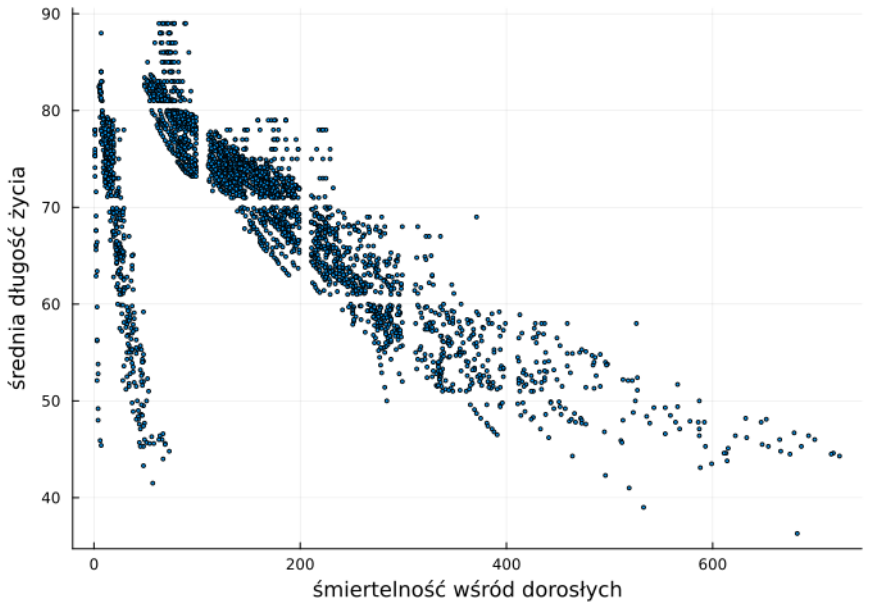
Celem reportu była analiza wybranych danych na podstawie dotychczasowej wiedzy zdobytej na wykładzie oraz laboratoriach z Komputerowej analizy szeregów czasowych. Najważniejszą częścią sprawozdania było badanie zależności liniowej między danymi oraz używając regresji liniowej estymacja współczynników tej zależności. Wykonano również oddzielną analizę jednowymiarową dla zmiennej zależnej oraz zmiennej niezależnej.

Dane użyte w raporcie dotyczą średniej długości życia oraz różnych parametrów mających na nią wpływ. Zbiór danych został znaleziony na stronie kaggle.com[1], a przygotowane zostały na bazie stron internetowych WHO (World Health Organisation) oraz United Nations. Informacje podane są dla 193 różnych państw w latach 2000-2015. Wszystkich wartości jest 2938.

W sprawozdaniu omówiona zostanie zależność między średnią długością życia (w latach, z dokładnością do jednego miejsca po przecinku), a śmiertelnością wśród dorosłych (wyznaczana jako ilość śmierci osób w wieku między 15, a 60 lat na 1000 osób w danym roku i w danym kraju, z dokładnością do liczby całkowitej). Po odrzuceniu wartości, w których co najmniej jedna z powyższych wartości jest brakująca dostajemy zbiór z 2928 elementami.

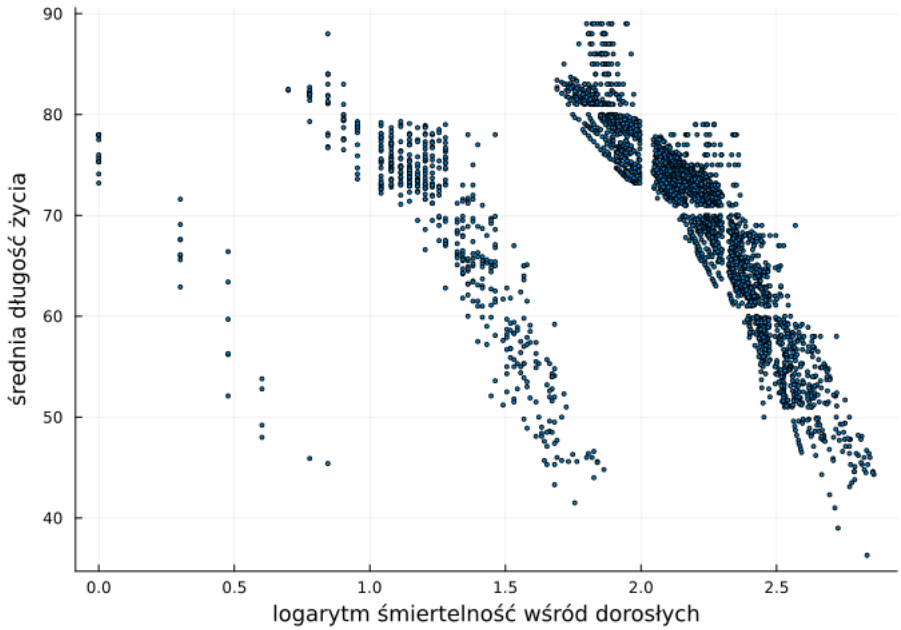
2 Poprawa błędnie zapisanych danych

Podczas wizualizacji danych zauważono dziwne zachowanie średniej długości życia w zależności od śmiertelności wśród dorosłych. Na rysunku 1 widać wyraźny podział danych na grupy. Wykonując wykres zależności średniej długości życia od logarytmu śmiertelności dorosłych (rysunek 2) można jeszcze wyraźniej zauważyć podział na 3 grupy. Również zauważono, że dla zmiennej niezależnej nie występują wartości podzielne przez 10.



Rysunek 1: Zależność średniej długości życia od śmiertelności wśród dorosłych (niepoprawiona)

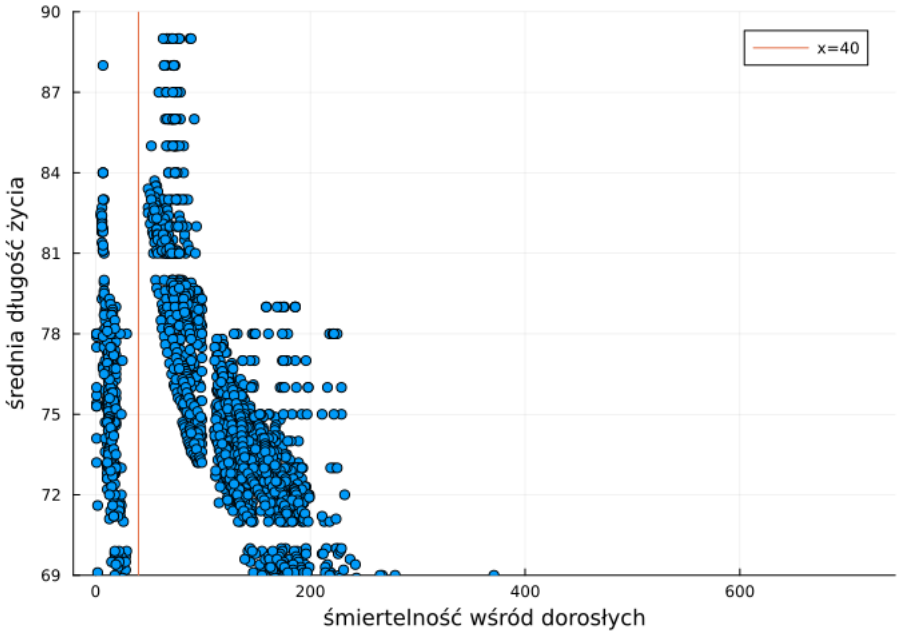
[1] https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who



Rysunek 2: Zależność średniej długości życia od logarytmu dziesiętnego śmiertelności wśród dorosłych (niepoprawiona)

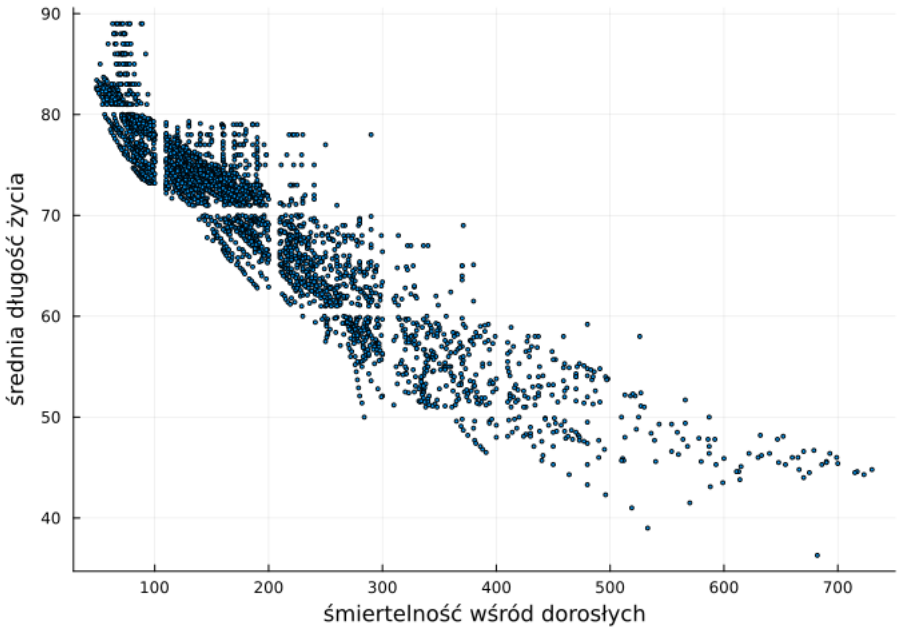
Analizując to zachowanie, stwierdzono że dane dotyczące zmiennej niezależnej zostały niepoprawnie zapisane, ponieważ dla wartości podzielnych przez 10 zostały one zapisane pomijając ostatnie zera (przykładowo 730 zapisano jako 73, a 300 jako 3). Wykorzystując tą wiedze przekształcono dane tak, aby były poprawne.

Dla wartości średniej długości życia mniejszych niż 70, wartości dotyczące zmiennej niezależnej mniejsze niż 10 zostały przemnożone przez 100, a wartości pomiędzy 10, a 100 zostały przemnożone przez 10. Największym problemem były dane dla wartości zależnej większej lub równej 70, ponieważ prawdziwe wartości jak i niepoprawne wartości mogły być mniejsze od 100. Narysowano wykres (rysunek 3), aby znaleźć prostą "odzielającą" wartości poprawne od niepoprawnych. Widać, że prosta $x = 40$ dosyć dobrze grupuje te dane. Korzystając z tego oraz faktu że wartości śmiertelności wśród dorosłych na tym przedziale przyjmują wartości mniejsze niż 300, dla wartości zależnej większych lub równych 70, wartości dotyczące zmiennej niezależnej mniejsze niż 3 zostały przemnożone przez 100, a wartości mniejsze niż 40 przez 10.



Rysunek 3: Zależność średniej długości życia (tylko dla wartości większych niż 70) od śmiertelności wśród dorosłych (niepoprawiona)

Po wykonanych przekształceniach otrzymane dane wyglądające jak na rysunku 4. Wykonane przekszatłcenia poprawiły wiarygodność danych, ale dalej mogą występować błędy. Najprawdopodobniej jednak ograniczono je na tyle mocno, że nie powinny przeszkadzać w dalszej analizie.



Rysunek 4: Zależność średniej długości życia od śmiertelności wśród dorosłych

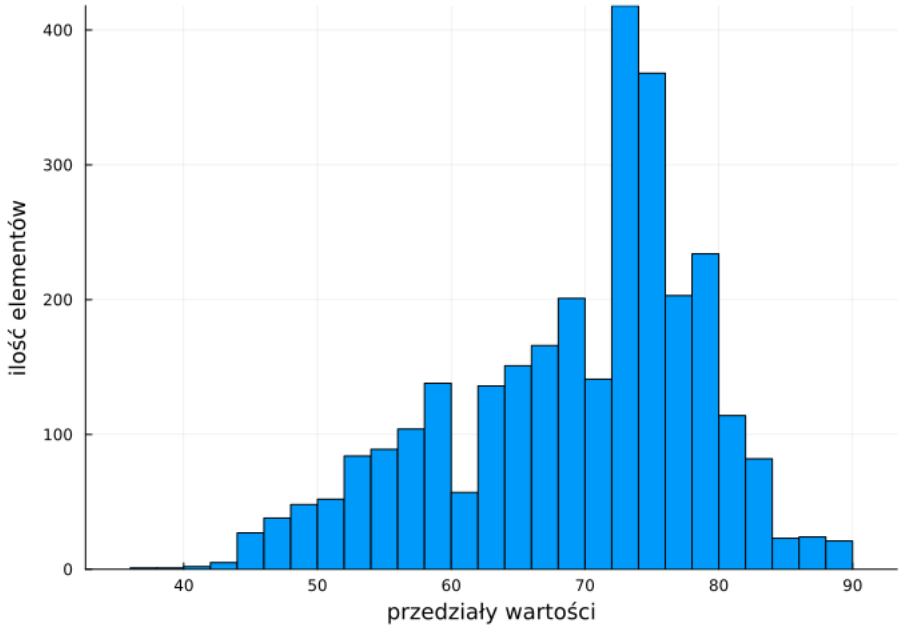
3 Analiza jednowymiarowa

3.1 Zmienna zależna

średnia arytmetyczna	wariancja	odchylenie standardowe	skośność	kurtoza[2]
69,22	90,70	9,52	-0,64	-0,24
pierwszy kawrtyl	mediana	trzeci kwartyl	minimum	maximum
63,1	72,1	75,1	36,3	89,8

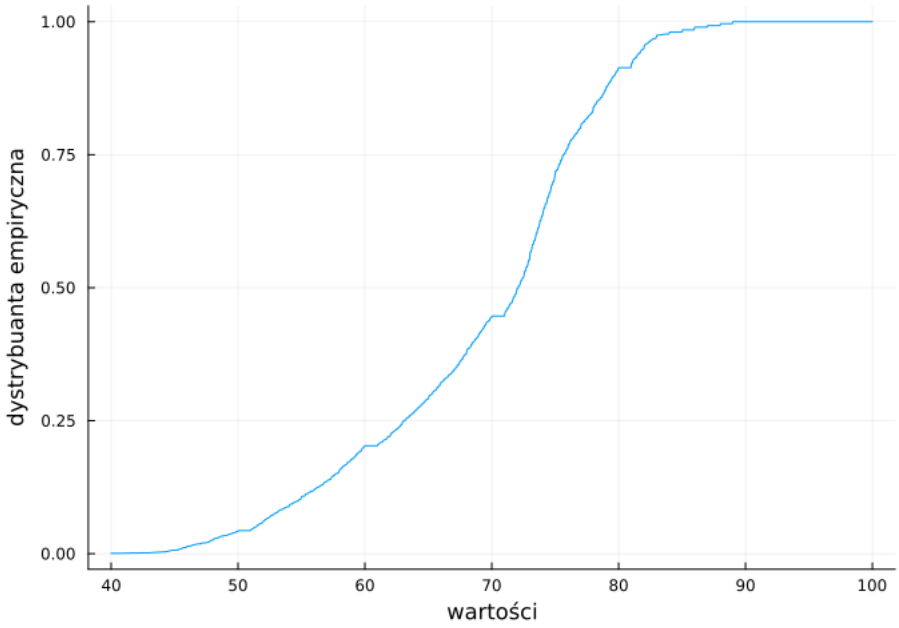
Tabela 1: Podstawowe statystyki śmiertelności wśród dorosłych

W tabeli (Tab. 1) wyliczone zostały podstawowe miary położenia, rozproszenia, skośności oraz spłaszczenia dla danych dotyczących średniej długości życia. Znalezione wartości wskazują na nieznaczną lewoskośność, można to zauważyć na histogramie poniżej (Rys. 5). Wartość kurtozy nieodstaje mocno od wartości dla rozkładu normalnego. Wszystkie wartości są mocno dodatnie (większe niż 36,3), co można łatwo zauważyć na wykresie dystrybuanty empirycznej (Rys. 6), dodatnie wartości można również wywnioskować z interpretacji danych dotyczącej średniej długości życia. Rozproszenie wartości jest dosyć niskie oraz nie występuje dużo wartości odstających co widać dokładnie na wykresie pudełkowym (Rys. 7). Nie udało się znaleźć znanego rozkładu opisującego dane, jednak nie mogą być one normalne przez brak symetrii i z interpretacji danych, które nie mogą przyjmować wartości ujemnych.

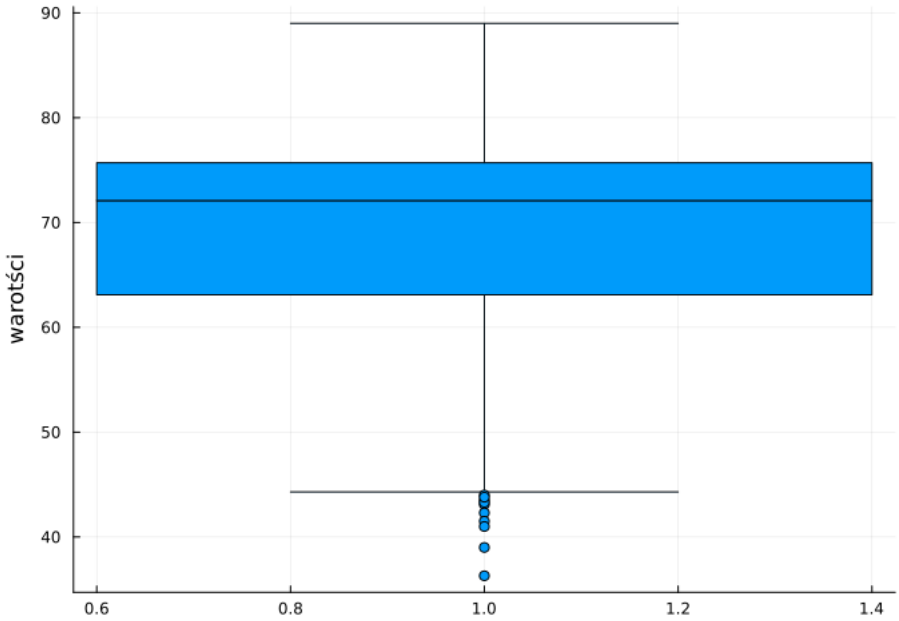


Rysunek 5: Histogram ilościowy średniej długości życia

[2]Kurtoza zdefiniowana w wariancie dla którego rozkład normalny ma kurtozę równą 0



Rysunek 6: Dystrybuanta empiryczna średniej długości życia



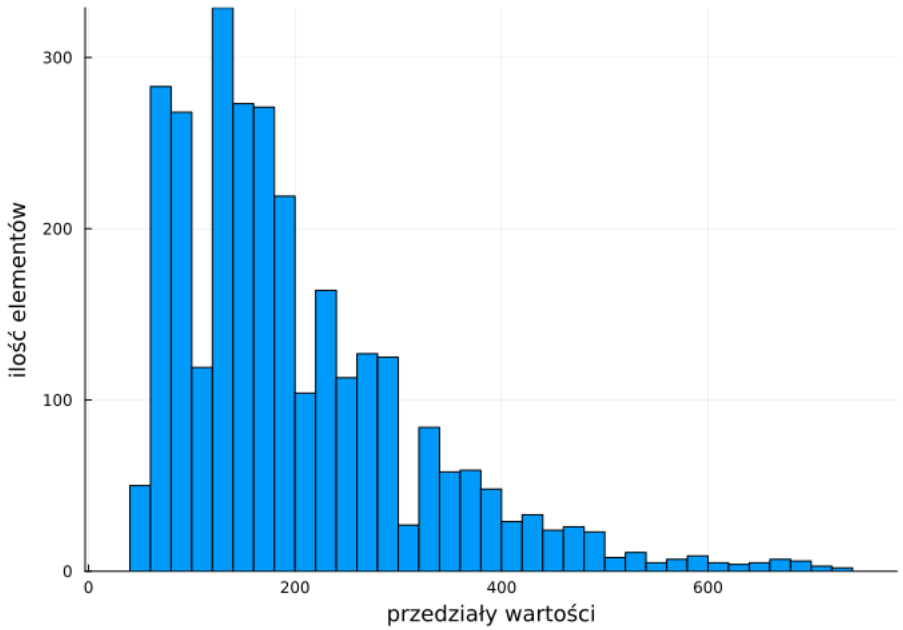
Rysunek 7: Wykres pudełkowy średniej długości życia

3.2 Zmienna niezależna

średnia arytmetyczna	wariancja	odchylenie standardowe	skośność	kurtoza[3]
200,84	14103,64	118,76	1,37	2,13
pierwszy kawrtyl	mediana	trzeci kwartyl	minimum	maximum
120	169	260	49	730

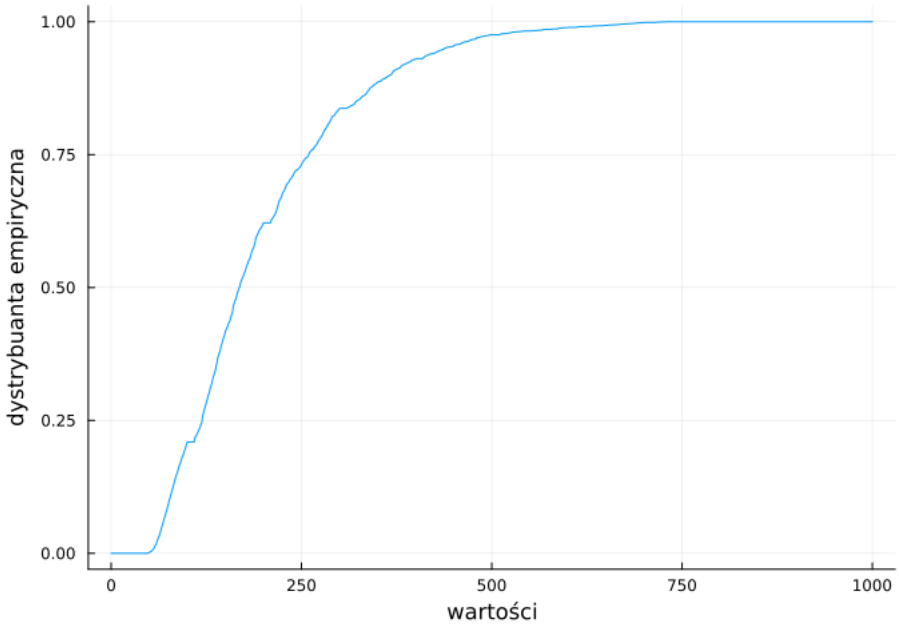
Tabela 2: Podstawowe statystyki średniej długości życia

W tabeli 2 wyliczone zostały podstawowe miary położenia, rozproszenia, skośności oraz spłaszczenia dla danych dotyczących śmiertelności wśród dorosłych. Znalezione wartości wskazują na dosyć sporą prawostronną skośność. Widać to na narysowanym poniżej histogramie (Rys. 8). Przyjmowane są tylko wartości dodatnie co najłatwiej zauważyć na wykresie dystrybuanty empirycznej (Rys. 9), ale również wynika to z interpertacji śmiertelności wśród dorosłych. Z wartości kurtozy można odczytać, że rozkład opisujący dane jest bardziej leptokurtyczny niż rozkład normalny. Wykres pudełkowy (Rys. 10) nie wskazuje na szczególnie dużą ilość wartości odstających, ponieważ taka ilość wartości poza wąsami wykresu jest podobna do ilości wartości poza wąsami dla rozkładu wykładniczego. Warto również zauważyć, wizualne braki w danych (przykładowo pojedyncze mocno niższe słupki histogramu). Wynikają one najprawdopodobniej z błędów lub braków w danych. Nie udało się dopasować żądnego znanego rozkładu odpowiadającemu danym, ale z pewnością nie jest to rozkład normalny (brak symetrii), ani wykładniczy.

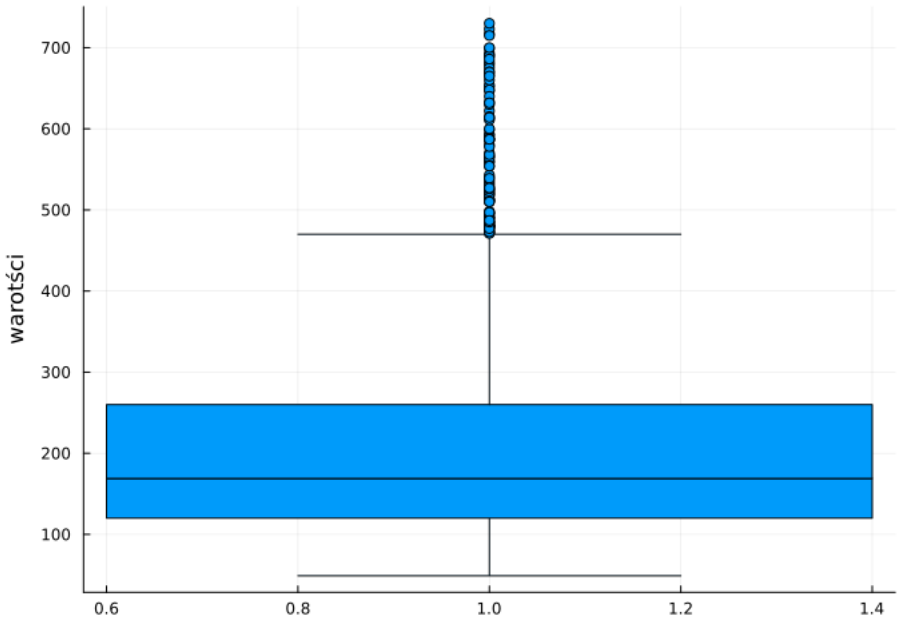


Rysunek 8: Histogram ilościowy śmiertelności wśród dorosłych

[3]Patrz dopisek 2



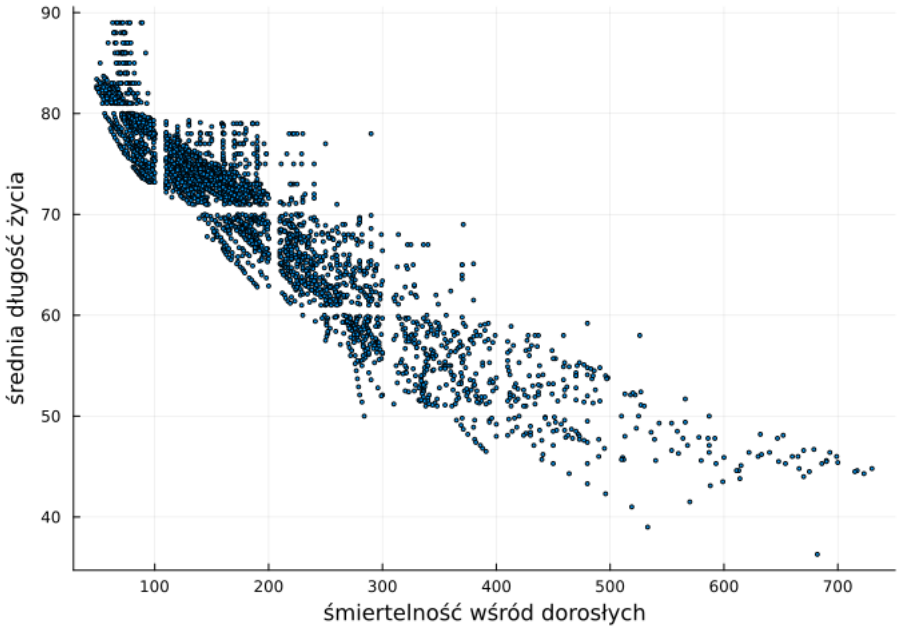
Rysunek 9: Dystrybuanta empiryczna śmiertelności wśród dorosłych



Rysunek 10: Wykres pudełkowy śmiertelności wśród dorosłych

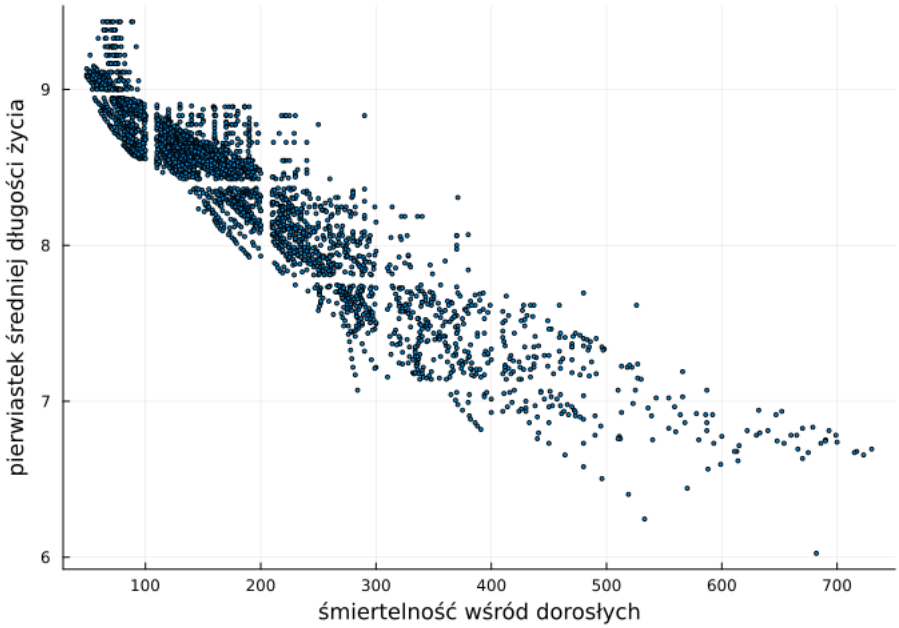
4 Analiza zależności liniowej

4.1 Wizualizacja i zależność



Rysunek 11: Wykres rozproszenia badanych danych

Na powyższym wykresie (Rys. 11) przedstawiono zależność między średnią długością życia, a śmiertelnością wśród dorosłych. Wizualnie jest dosyć bliska liniowej. Wyestymowano współczynnik korelacji Pearsona, który wynosi w przybliżeniu $r = -0{,}928$. Wskazuje on na duża zależność liniową.



Rysunek 12: Wykres rozproszenia badanych danych

Zależność można jednak poprawić. Biorąc pierwiastek z zmiennej zależnej (czyli przyjmując zależność kwadratową) otrzymujemy dane rozproszone jak na wykresie powyżej (Rys. 12) oraz współczynnik korelacji Paersona równy w przybliżeniu $r' = -0{,}934$. Nie zwiększa to drastycznie zależności jednak wiadomo, że dane nie mogą mieć zależności liniowej i wynika to z wiedzy, że ani zmienna zależna, ani niezależna nie powinny przyjomwać wartości ujemnych. Zależność kwadratowa wydaje się lepsza, ponieważ parabola będzie malała do zera wolniej niż prosta. Dalszą część analizy będzie zrobiona przy założeniu zależności kwadratowej, czyli będziemy ją wykonywać dla pierwiastka z wartości dotyczących średniej długości życia.

4.2 Regresja liniowa

W tej części raportu zostanie znaleziona prosta, która przybliża zależność między badanymi wartościami. Wykonane zostanie to za pomocą regresji liniowej, czyli metody pozwalającej na estymację współczynników prostej przybliżającej dane. Przyjmujemy że zmienna zależna w naszym modelu jest w postaci:

$$Y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i, \text{ dla } i = 1, 2, 3, \ldots, n$$

gdzie:
n - długość próby
Y_i - zmienne opisujące zmienną zależną,
x_i - deterministyczne wartości,
$\mathcal{E}_i$ - zmienne nieskorelowane o średniej 0 i wariancji σ^2 (inaczej jest to biały szum).
β_0, β_1 - współczynniki prostej regresji

Dla takiego modelu otrzymujemy prostą regresji opisaną równaniem:

$$y = \beta_0 + \beta_1 x$$

4.2.1 Estymacja punktowa

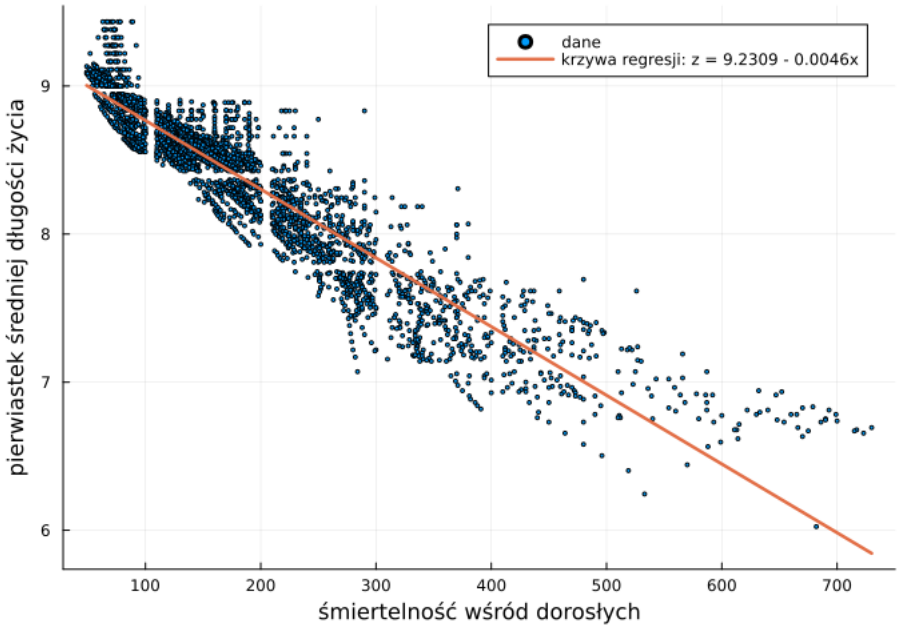
Korzystając z metody najmniejszych kwadratów, na wykładzie zostały wyprowadzone nieobciążone estymatory punktowe szukanych parametrów:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{x}$$

gdzie:
$\hat{\beta}_0, \hat{\beta}_1$ - estymatory kolejno β_0, β_1
$\bar{x}$ - średnia arytmetyczna $x_1, x_2, \ldots, x_n$.
$\bar{Y}$ - średnia arytmetyczna zmiennych $Y_1, Y_2, \ldots, Y_n$

Dla badanych obserwacji zmiennej zależnej będącej pierwiastkiem z średniej długości życia i zmiennej niezależnej będącej śmiertelnością wśród dorosłych wyestymowano wartości szukanych parametrów. Ich przybliżone wartości wynoszą: $\beta_1 = -4{,}6391 \cdot 10^{-3}$ i $\beta_0 = 9{,}2309$. Na wykresie (Rys. 13) przedstawiono znalezioną prostą regresji w porównaniu do przybliżanych nią przekształconych danych. Na wykresie można zauważyć, że prosta dosyć dobrze przybliża trend danych.

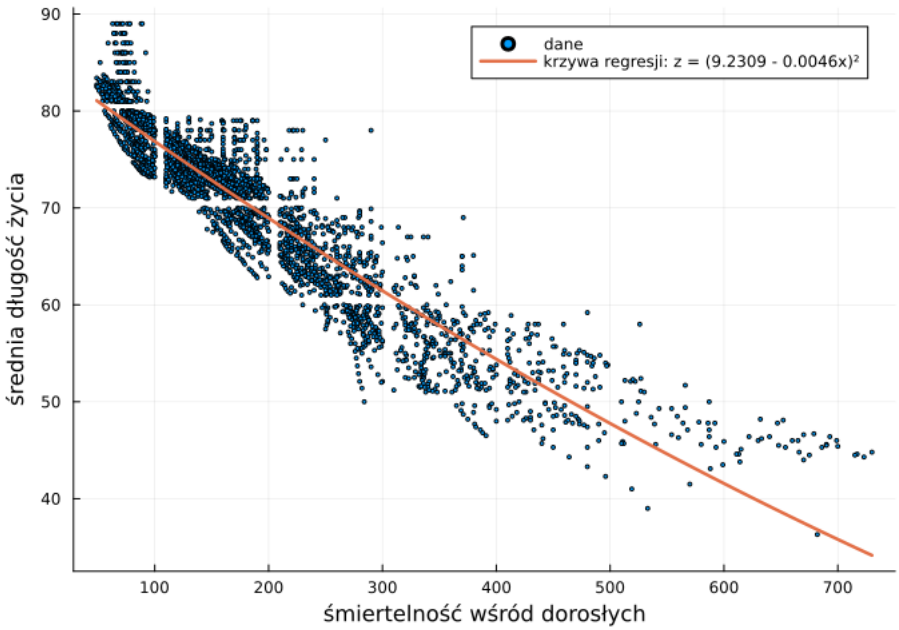


Rysunek 13: Przybliżenie przekształconych danych prostą regresji

Pamiętając że dane dotyczące zmiennej zależnej przekształcono pierwiastkiem. można dostać nieprzekszatłconą zależność między danymi. Na wykresie (Rys. 14) przedstawiono tą zależność.

Widać na nim, że krzywa dosyć dobrze odwzorowuje trend w danych. Równanie opisujące zależność wygląda następująco:

$$z = y^2 = (\beta_0 + \beta_1 x)^2$$



Rysunek 14: Przybliżenie danych krzywą regresji

4.2.2 Estymacja przedziałowa

W tej części analizy przyjmujemy dodatkowo, że residua w modelu regresji liniowej (zmienne losowe $\mathcal{E}_i$) są wzajemnie niezależne oraz z rozkładu normalnego $\mathcal{N}(0, \sigma^2)$, dla nieznanej wariancji $\sigma^2 \in \mathbb{R}$.

Przy takich założeniach znamy rozkład poniższych statystyk:

$$\frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim \mathcal{T}(n-2)$$

$$\frac{\hat{\beta}_0 - \beta_0}{S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{T}(n-2)$$

gdzie:
$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ - nieobciążony estymator σ^2,
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Bazując na tych statystykach możemy skonstruować przedziały ufności na poziomie istotności $\alpha \in (0,1)$:

$$\beta_1 \in \left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

$$\beta_0 \in \left[\hat{\beta}_0 - t_{1-\frac{\alpha}{2}} S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2}} S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

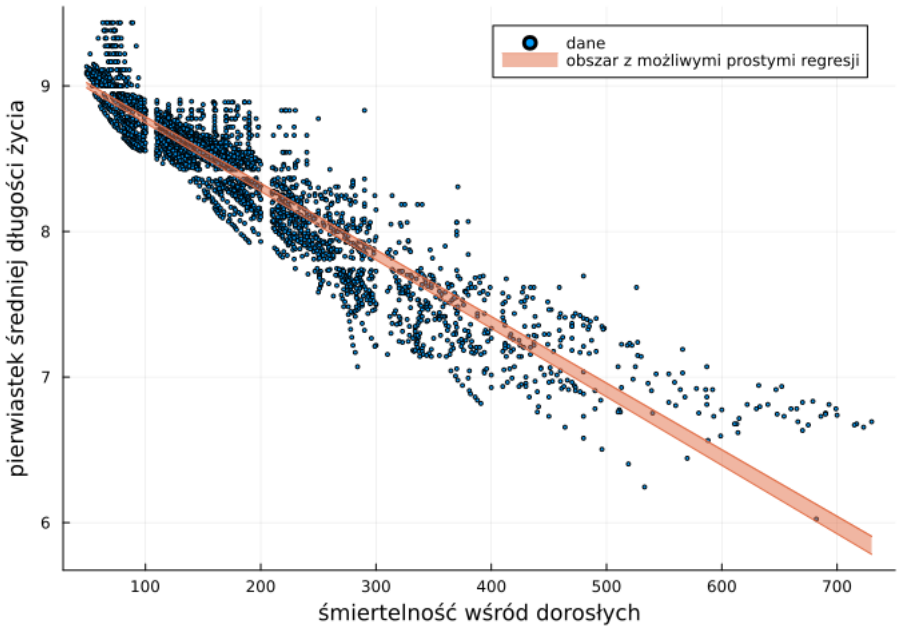
gdzie:
$t_{1-\frac{\alpha}{2}}$ - kwantyl rzędu $1-\frac{\alpha}{2}$ rozkładu $\mathcal{T}(n-2)$.

Dla badanych danych, korzystając z wyliczonych w sekcji 4.2.1 estymatorów $\hat{\beta}_1$ i $\hat{\beta}_0$, dostajemy następujące przedziały ufności na poziomie istotności $\alpha = 0{,}05$:

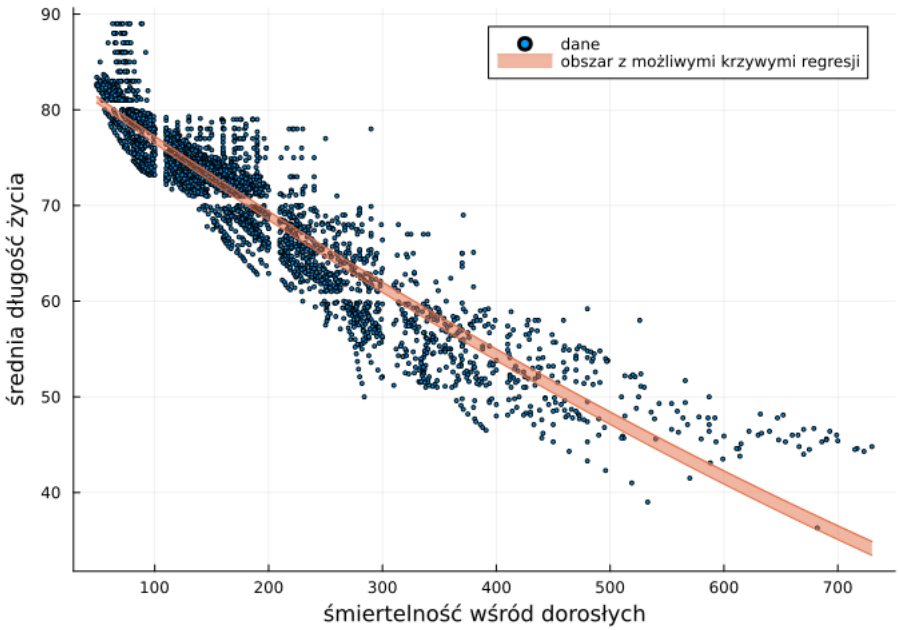
$$\beta_1 \in \left[-4{,}7034 \cdot 10^{-3};\ -4{,}5749 \cdot 10^{-3}\right]$$

$$\beta_0 \in [9{,}2159;\ 9{,}2459]$$

Na wykresie poniżej (Rys. 15), można zobaczyć obszar zawierający możliwe proste regresji, których współczynnki należą do przedstawionych wyżej przedziałów ufności, w zestawieniu z badanymi przekształconymi danymi. Natomiast na rysunku 16 widać możliwe krzywe regresji przybliżające już nieprzekształcone dane. Analizując oba wykresy oraz same wartości przedziałów ufności, można stwierdzić, że różnice między możliwymi krzywymi regresji nie są szczególnie duże. Oznacza to, że prawdopodobnie dane są dosyć dobrze dopasowywane zależnością kwadratową.



Rysunek 15: Wykres pokazujące możliwe proste regresji, dla współczynników należących do odpowiednich przedziałów ufności



Rysunek 16: Wykres pokazujące możliwe krzywe regresji, dla współczynników należących do odpowiednich przedziałów ufności

4.3 Ocena poziomu zależności

współczynnik korelacji Pearsona	SST	SSE	SSR
-0,9341	1018,17	129,73	888,44

Tabela 3: Współczynniki mieżące zależność liniową

W powyższej tabeli (Tab. 3) przedstawiono wartości, za pomocą których można próbować oceniać jakość dopasowania regresji liniowej dla badanych przekształconych danych. Wyliczone zostały korzystając z poniższych wzorów:

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

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SSE + SSR$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

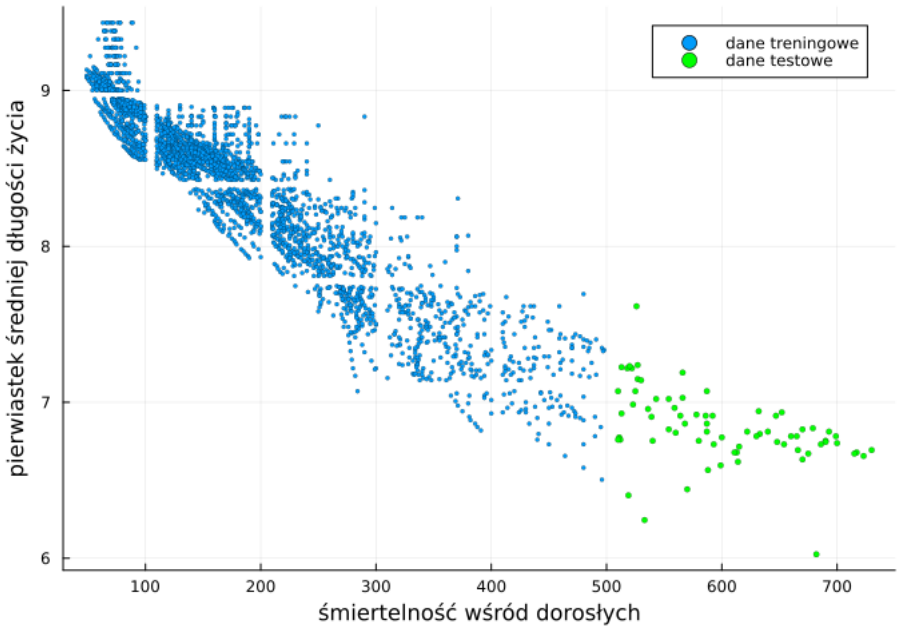
gdzie:
r - próbkowy współczynnik korelacji Pearsona
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, dla wartości estymatorów $\hat{\beta}_1$ i $\hat{\beta}_0$ wyliczonych dla badanych danych

Współczynnik korelacji Pearsona jest wysoki co do wartości bezwględnej (co zauważono w sekcji 4.2.1), mówi to o wysokiej liniowej zależności. Jego ujemna wartości mówi o tym, że gdy obserwacje zmiennej niezależne rosną to obserwacje zmiennej zależnej maleją. Zgadza się to z tym co widać na wykresach rozproszenia. Wartość SST określa całkowitą zmienność danych. Ciężko jednak ją samą w sobie interpretować bo nie jest w żaden sposób unormowana. Aby określić jak dobrze model jest dopasowany do danych porównujemy wartości SSE i SSR do SST. W badanym przypadku SSE (które jest sumą kwadratów błędów przybliżenia) jest dosyć małe w porównaniu do SST, co sugeruje dobre dopasowanie. Wartością, która pozwala ocenić jaka część zmienności została objaśniona przez model jest SSR. Wartość ta jest dosyć duża, co ponownie sugeruje dobre dopasowanie. Podobne wnioski o dosyć dobrym dopasowaniu, można wyciągnąć patrząc na wykresy narysowane w poprzednich sekcjach.

5 Predykcja dla danych testowych

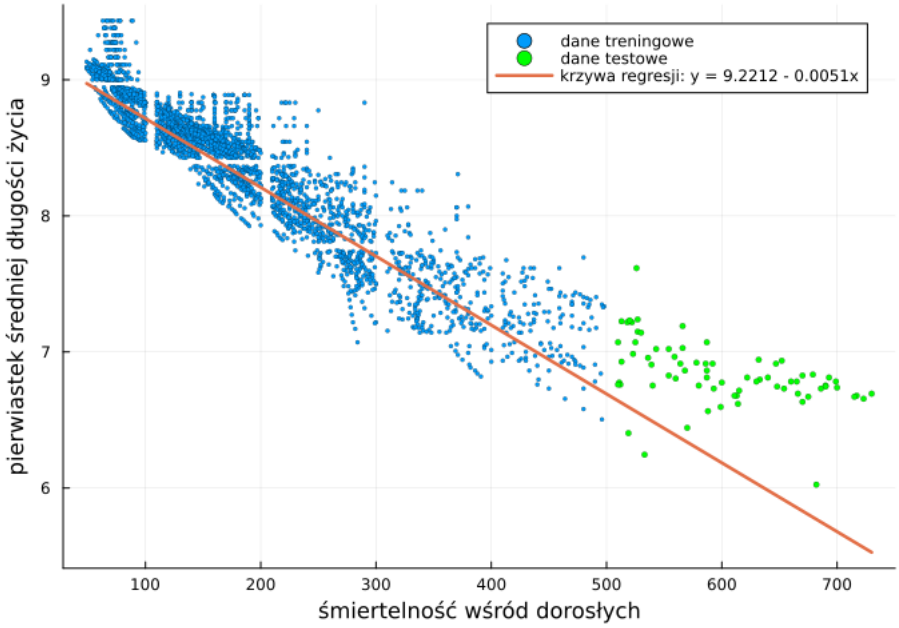
W tej części sprawozdania, dalej rozważany jest model regresji liniowej z residuami, będącymi niezależnymi zmiennymi losowymi z rozkładu $\mathcal{N}(0, \sigma^2)$ z nieznaną wariancją $\sigma^2 \in \mathbb{R}$. Zamiast estymować współczynniki regrersji liniowej na bazie całych danych, zostaną one wyestymowane tylko na sporej ich części. Następnie skonstruowane zostaną przedziały ufności dla predykcji danych, które nie zostały uwzględnione w estymacji parametrów.

Działając na przekształconych danych, do dopasowania współczynnkiów regresji liniowej wzięto te obserwacje (dane treningowe) dla których śmiertelność wśród dorosłych była mnijesza niż 500. Na wykresie poniżej (Rys. 17) przedstawiono jak wyglądają tak rozdzielone dane. Danych które nie zostaną uwzględnionych (dane testowe) w estymacji współczynników jest 72 (około 2,5% wszystkich obserwacji).



Rysunek 17: Podział danych

Na podstawie tak podzielonych daych wyestymowano wartości współczynników i wynoszą one około: $\beta_1 = 5{,}0601 \cdot 10^{-3}$ i $\beta_0 = 9{,}2212$. Można dostrzec, że współczynniki różnią się zauważalnie (choć nie bardzo mocno) od tych wyznaczonego w sekcji (4.2.1). Na wykresie poniżej (Rys. 18) pokazano prostą przybliżającą badane obserwacje.



Rysunek 18: Prosta regresji dla częśći danych

Z wykładu znamy rozkłady statystyk powiązanych z predykowanymi zmiennymi:

$$\frac{\hat{Y}_0 - Y_0}{S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim \mathcal{T}(n-2)$$

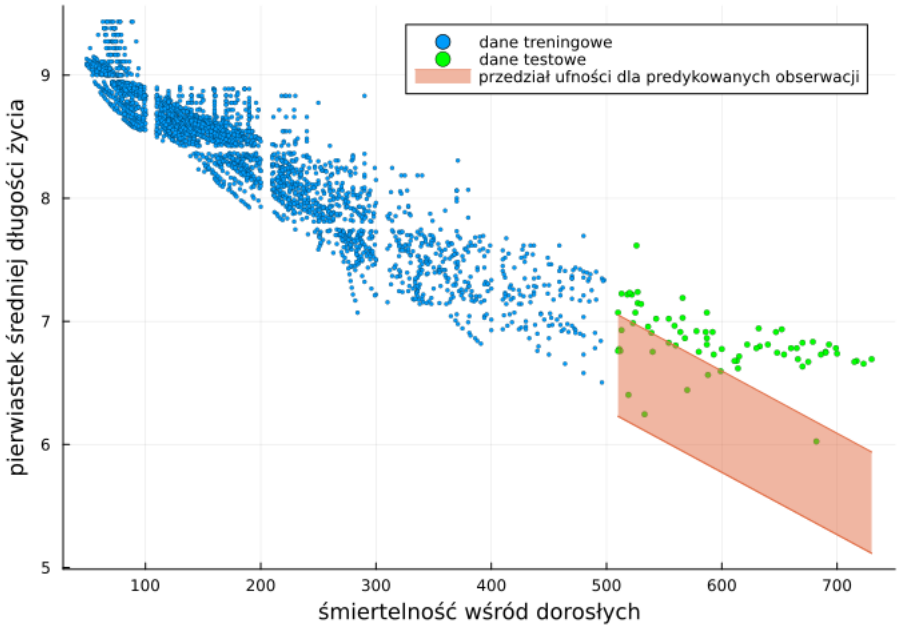
gdzie:
Y_0 - predykowana zmienna losowa dotycząca danych testowych
x_0 - punkt w którym predykujemy wartość zmiennej losowej
$\hat{Y}_0 = \beta_0 + \beta_1 x_0$
$S, \bar{x}, n$ - wartości wyliczone na podstawie tylko danych treningowych.

Korzystając z takich statystyk możemy wyznaczyć przedział ufności na poziomie istotności $\alpha \in (0, 1)$ dla danych testowych:

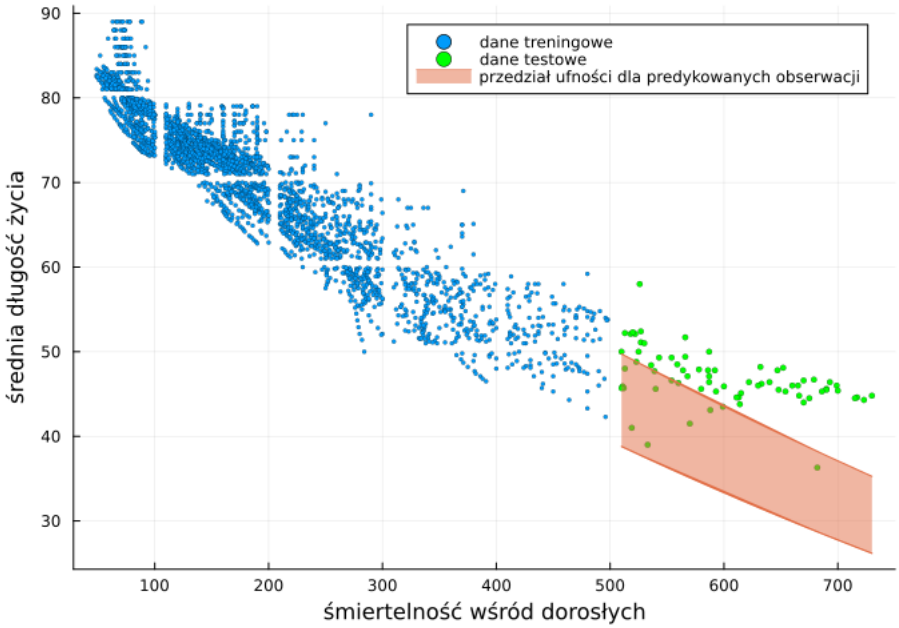
$$Y_0 \in \left[\hat{Y}_0 - t_{1-\alpha/2} S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{Y}_0 + t_{1-\alpha/2} S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

Wyliczono przedziały ufności na poziomie istotności $\alpha = 0{,}05$ dla wcześniej oddzielonych danych testowych. Na wykresie poniżej (Rys. 19) widać przedziały ufności predykowanych zmiennych dla przekształconych danych, natomiast na rysunku 20 widać te przedziały dla nieprzekształconych danych. Analizując wykresy można stwierdzić, że dane testowe w dosyć małej ilości wpadają do

wyznaczonych przedziałów ufności. Może to oznaczać, że model regresji (przynajmniej dla tej części danych) nie jest dobrze dobrany.



Rysunek 19: Przedziały ufności dla predykowanych, przekształconych danych

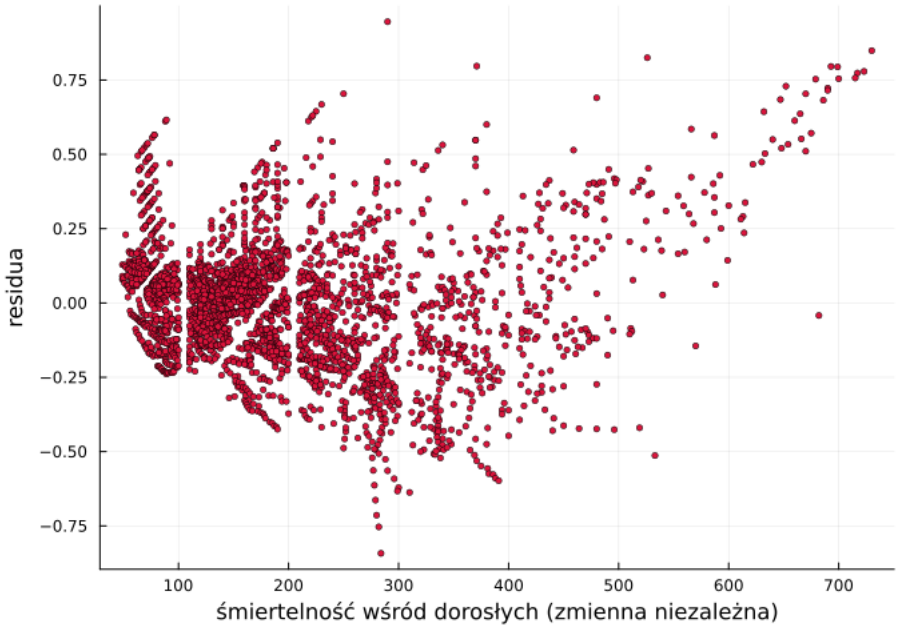


Rysunek 20: Przedziały ufności dla predykowanych, nieprzekształconych danych

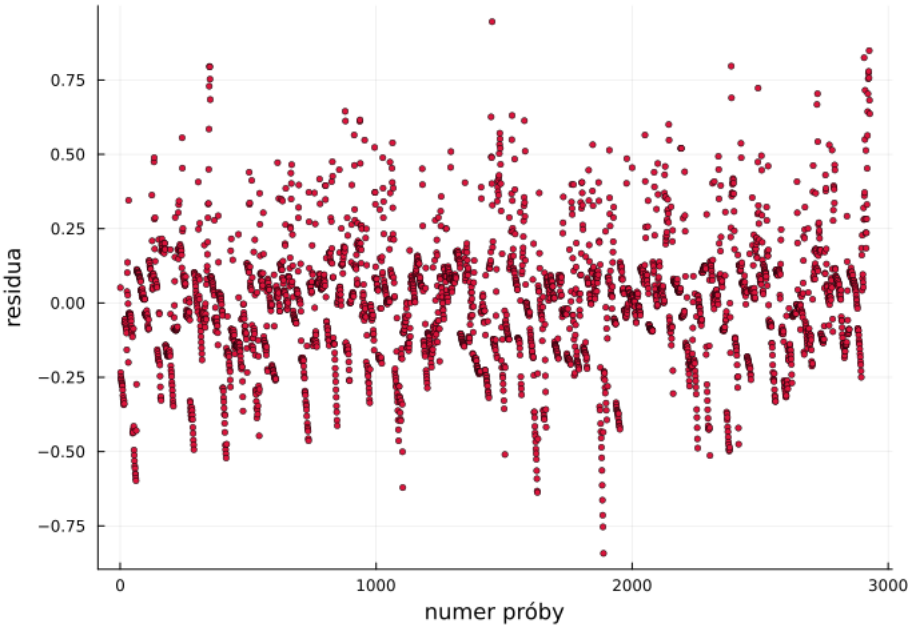
6 Analiza residuów

Analizą residuów jest badanie (sprawdzanie czy spełniają założenia modelu) realizacji zmiennych losowych $\mathcal{E}_i$, zdefiniowanych jako $e_i := y_i - \hat{y}_i$, gdzie $\hat{y}_i = \beta_0 + \beta_1 x_i$, dla wyestymowanych wartości współczynników β_1 i β_0.

Na poniższym wykresie (Rys. 21) narysowano residua dla badanych przekształconych danych i modelu regresji współczynników wyestymowanych w sekcji 4.2.1. Residua narysowano w zależności od zmiennej zależnej, żeby potencjalnie zobaczyć jakąś zależność od badanych danych. Nie jest to całkiem losowy kształt, ale nie widać żadego wyraźnego trendu. Narysowano również (Rys. 22) residua w zależności od numeru próby. Ten wykres będzie podstawą do większości dalszych rozważań o residuach. W dalszej części zostaną sprawdzone założenia dotyczące tych residuów.



Rysunek 21: Residua w zależności od zmiennej niezależnej



Rysunek 22: Residua w zależności od numeru próby

6.1 Średnia

W przyjętym modelu zakładamy, że średnia zmiennych $\mathcal{E}_i$ wynosi zero dla każdego $i = 1,2,\ldots,n$. Do badania średniej dodatkowo założona zostanie normalność residuów. Wykorzystany zostanie t-test, dla którego będzie testowana hipoteza zerowa $H_0 :\ \mu = \mu_0$, przeciwko hipotezie alternatywnej $H_1 :\ \mu \neq \mu_0$, dla μ będącą średnią badanych residuów. Statystyka testowa, gdy zachodzi hipoteza zerowa jest postaci:

$$\frac{\bar{\mathcal{E}} - \mu_0}{S_{\mathcal{E}}/\sqrt{n}} \sim \mathcal{T}(n-1)$$

gdzie:
$S_{\mathcal{E}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathcal{E}_i - \bar{\mathcal{E}})^2$ - nieobciążony estymator $\sigma^2 = Var(\mathcal{E}_i)$,
$\mu_0 = 0$ - oczekiwana średnia zmiennych.

Test z wysoką p-wartością (prawie 1) nie odrzuca hipotezy zerowej. Licząc średnią arytmetyczną z zaobserwowanych residuów dostajemy $\mu \approx 1{,}398 \cdot 10^{-15}$, więc wynik testu jest zupełnie zrozumiały.

Żeby sprawdzić czy średnia jest stale równa zero wykonano ten sam test, ale dla pierwszej (obserwacje od 1 do 1464) i drugiej (obserwacje od 1465 do 2928) połowy danych osobno. W obu przypadkach hipoteza zerowa została odrzucona, ale w obu przypadkach p-wartość wynosiła około 0,02, co nie jest aż tak małą wartością. Licząc wartośći średniej z tych obserwacji dla pierwszej połowy otrzymano $\mu_1 \approx -0{,}013$, a dla drugiej połowy $\mu_2 \approx 0{,}013$. Nie wydaje się to aż tak duża różnica, ale zważając, że próba jest dosyć duża test odrzuca hipotezy zerowe.

Zważywszy na wyniki testów ciężko stwierdzić, czy średnie rzeczywiście jest stale równa zero. Testy mogą działać nie do końca poprawnie z uwagi na założenie o normalności rozkładu. Dodatkowo patrząc na rysunek 22, można wizualnie spostrzec, że residua są dosyć równomiernie rozłożone

wokół zera. Ostatecznie nie można jednoznacznie zaprzeczyć twierdzeniu, że residua mają średnią stale równą zero.

6.2 Wariancja

W przyjętym modelu zakładamy, że wariancja zmiennych $\mathcal{E}_i$ jest stałą dla każdego $i = 1,2,\ldots,n$. Do badania wariancji ponownie założona zostanie normalność residuów. Wykorzystany zostanie F-test[4] służący do sprawdzania równości wariancji. Testowana będzie hipoteza $H_0 : \sigma_1^2 = \sigma_2^2$ przeciwko hipotezie alternatywnej $H_1 : \sigma_1^2 \neq \sigma_2^2$. Statystyka testowa jest postaci:

$$\frac{S_X^2}{S_Y^2} \sim \mathcal{F}(n-1, m-1)$$

gdzie:
$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ - nieobciążony estymator $\sigma_1^2 = Var(X_i)$ dla $i = 1,2,\ldots,n$,
$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2$ - nieobciążony estymator $\sigma_2^2 = Var(Y_i)$ dla $i = 1,2,\ldots,m$,
$\mathcal{F}(n-1, m-1)$ - rozkład F Snedecora z $n-1$ i $m-1$ stopniami swobody.

Test wykorzystany zostanie do sprawdzenia równości wariancji między obserwacjami residuów podzielonymi na dwie równe grupy, tak jak podczas badania średniej. Test odrzuca hipotezę zerową (p-wartość na poziomie 0,0004). Licząc wariancję próbkową obu grup otrzymujemy dla pierwszej połowy $\sigma_1^2 \approx 0{,}0401$ i dla drugiej połowy $\sigma_2^2 \approx 0{,}0483$. Różnica jest zauważalna zważając na ilość obserwacji.

Wyniki testu sugerują niezgodność założenia o stałej wariancji, ale ponownie test może nie działać do końca poprawnie z uwagin na założenie o normalności. Patrząc na wykres residuów (Rys. 22), rzeczywiście można zobaczyć nierównomierne rozproszenie wartości residuów. Najprawdopodobniej residua modelu rzeczywiście nie mają stałej wariancji.

6.3 Niezależność

W ogólnym modelu regresji liniowej zakładamy nieskorelowanie zmiennych $\mathcal{E}_i$ dla $i = 1,2\ldots,n$, a w niektórych przypadkach dodatkowo zakładamy ich niezależność. Przy dodatkowym założeniu o normalności rozkładu residuów, można badając korelacje zbadać niezależność.

Do sprawdzenia korelacji między danymi wykorzystana zostanie funkcja autokorelacji próbkowej. Jeśli obserwacje są rzeczywiście nieskorelowane to teoretyczna funkcja autokorelacji powinna być postaci:

$$\rho(h) = \frac{Cov(\mathcal{E}_1, \mathcal{E}_{1+h})}{\sqrt{Var(\mathcal{E}_1)Var(\mathcal{E}_{1+h})}} = \begin{cases} 1, & \text{dla } h = 0 \\ 0, & \text{dla } h \neq 0 \end{cases}$$

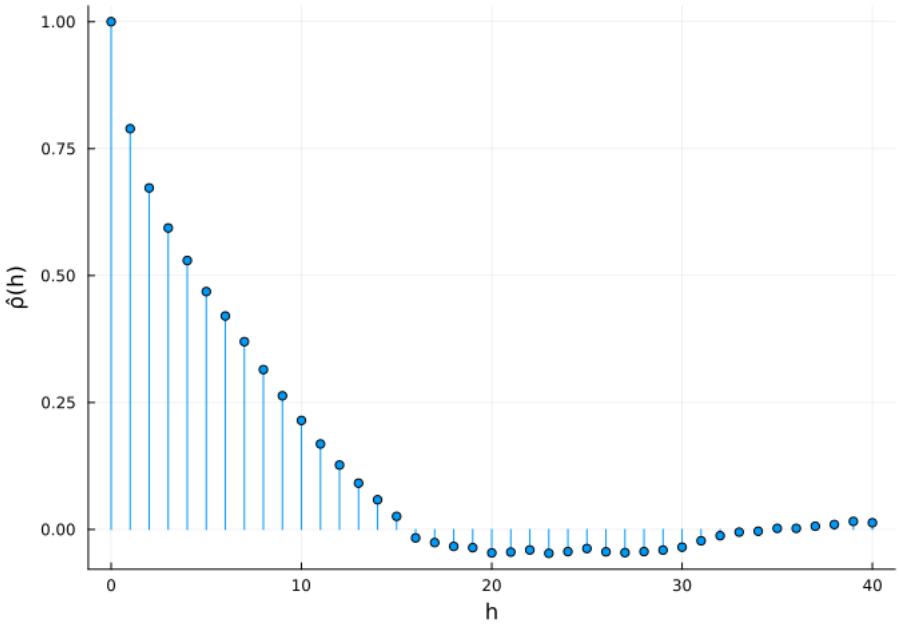
gdzie:
h - lag, tzn. względne przesunięcie między numerami obserwacji których korelacja jest liczona,

[4] George E. P. Box, "Non-Normality and Tests on Variances", Biometrika 40 (3/4): 318–335, 1953.

Funckja autokorelacji próbkowej jest postaci:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad \text{dla } \hat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n}(e_{1+h} - \bar{e})(e_1 - \bar{e})$$

Na wykresie poniżej (Rys. 23) narysowano funkcję autokorelacji próbkowej dla badanych obserwacji residuów dla lagów h od 0 do 40. Widać że dla pierwszych kilkunastu wartości funkcja nie przybliża zachowania funkcji teoretycznej. Świadczy to najprawdopodobniej o tym, że zaobserwowane residua są w jakimś stopniu skorelowane. Zatem założenie o nieskorelowaniu (a tym bardziej o niezależności) residuów nie jest spełnione.

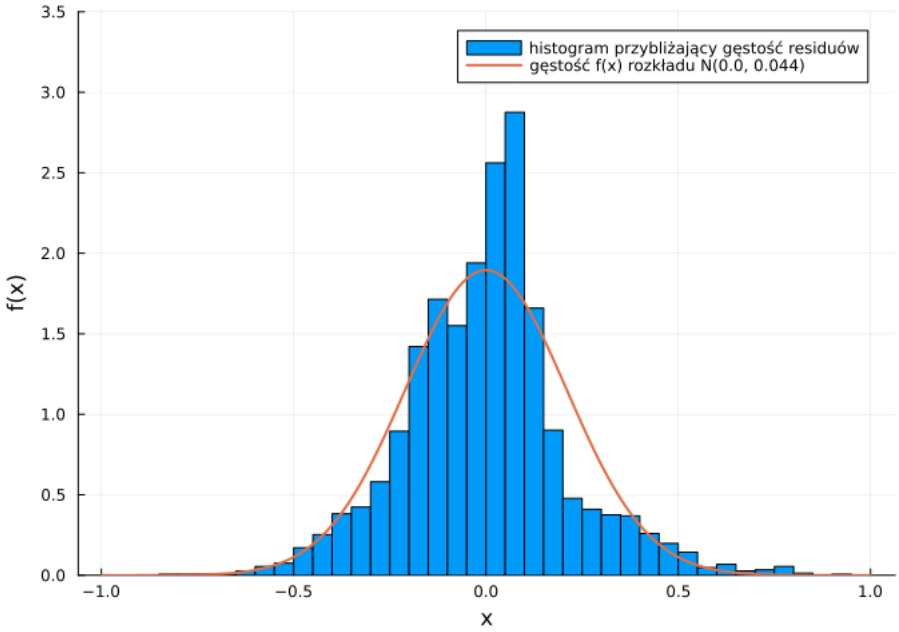


Rysunek 23: Autokorelacja próbkowa residuów

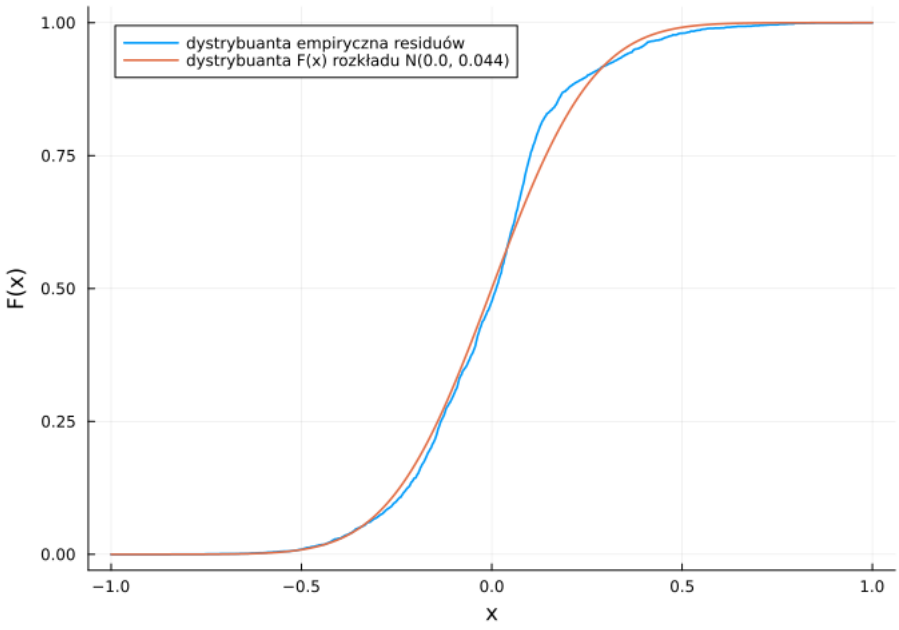
6.4 Normalność

Ogólny model regresji liniowej nie wymaga założenia o normalności residuów, ale w niektórych częściach sprawozdania (przykładowo w estymacji przedziałowej, sekcja 4.2.2) była ona dodatkowo zakładana.

Wykonano wykres (Rys. 24) przedstawiający unormowany histogram residuów w porównaniu do gęstości rozkładu normalnego z parametrem średniej i wariancji wyliczonym z obserwacji residuów. Wyraźnie widać na nim brak normalności danych. Narysowano również (Rys. 25) porównanie dystrybuanty empirycznej z residuów z dystrybuantą odpowiedniego rozkladu normalnego. Również widać na nim różnicę, jednak nie aż tak wyraźnie.



Rysunek 24: Histogram przybliżający gęstość residuów



Rysunek 25: Dystrybuanta empiryczna residuów

W celu zweryfikowania obserwacji o braku normalności wykonany zostanie test normalności Shapiro-Wilka z poprawką na wielkość próby[5]. Badana będzie hipotezą zerowa:

$$H_0: \text{ dane pochodzą z rozkładu normalnego.}$$

przeciwko hipotezie alternatywnej:

$$H_1: \text{ dane nie pochodzą z rozkładu normalnego}$$

Statystyka testowa wygląda następująco:

$$W = \frac{(\sum_{i=1}^{n} a_i \mathcal{E}_{(i)})^2}{\sum_{i=1}^{n} (\mathcal{E}_i - \bar{\mathcal{E}})^2}$$

gdzie:
$\mathcal{E}_{(i)}$ - i-ta statystyka pozycyjna residuów,
a_i - stabularyzowane wartości,
W - statystyka której dokładny rozkład nie jest nazwany, ale jest stabularyzowany na podstawie metod Monte Carlo.

Jak oczekiwano test odrzuca hipotezę zerową z p-wartością rzędu 10^{-20}. Wynik testu wraz z wizualną weryfikacją normalności świadczy o tym, że residua nie pochodzą z rozkładu normalnego.

7 Podumowanie

Celami sprawozdania była w szczególności analiza zależności liniowej między danymi oraz dopasowywanie modelu regresji liniowej. Zbadano residua dopasowania modelu regresji w celu weryfikacji jego poprawności. Zrobiono również predykcję odpowiednio oddzielonych danych testowych na bazie danych treningowych. Dodatkowo wykonano analizę jednowymiarową zmiennej zależnej i zmiennej niezależnej.

Dane dotyczyły średniej długości (zmienna zależna) oraz śmiertelności wśród dorosłych (zmienna niezależna). Do obserwacji (po odpowiednich przekształceniach) udało się dopasować prostą regresji liniowej. Współczynniki mieżące zależność liniową danych wskazywały na wysoką zależność. Podczas próby dokonania predykcji części danych zauważone zostały problemy modelu. Sprawdzając założenia dotyczące residuów odrzucono wszystkie z nich poza założeniem dotyczącym średniej. Ostatecznym wnioskiem z analizy jest słabe dopasowanie modelu do badanych danych.

[5]Rahman und Govidarajulu (1997). "A modification of the test of Shapiro and Wilk for normality"