Jakub Muzyka i Łukasz Simbiga

Wykorzystanie poznanych metod służących do analizy zależności liniowej dla wybranych danych rzeczywistych Raport nr 1

22 grudnia 2022

1. Wstęp - opis danych

Do analizy zależności liniowej wybraliśmy dane pochodzące ze strony: `https://www.kaggle.com/datasets/kolawale/focusing-on-mobile-app-or-website`. Dane te, przedstawiają informacje na temat kont użytkowników w konkretnym sklepie internetowym.
Możemy wyróżnić kolumny takie, jak:

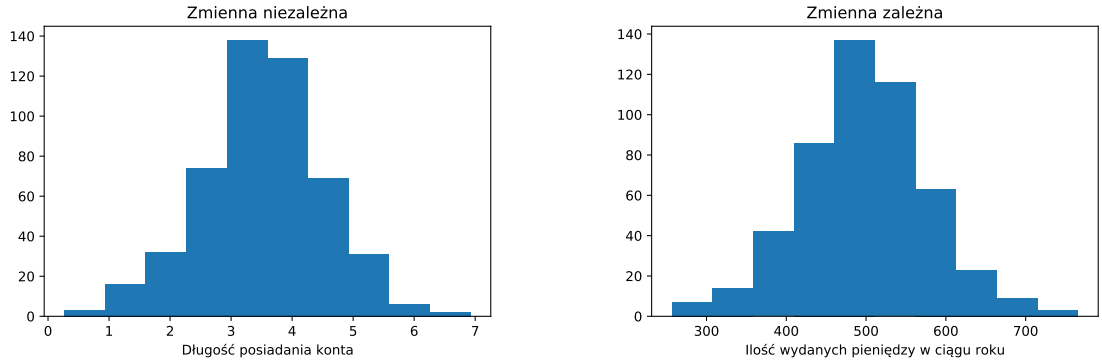
1. **Email** — przechowuje e-mail danego użytkownika.
2. **Address** — kolumna przechowująca adres klienta podawany przy rejestracji na stronie.
3. **Avatar** — Obrazek wypełniony kolorem, który klient ustawia przy rejestracji.
4. **Avg. Session Length** - Średni czas spędzony ze stylistą, przy wyborze ubrań w sklepie stacjonarnym. Ten sklep oferuje taką możliwość.
5. **Time on App** - czas spędzony na aplikacji danego sklepu.
6. **Length of Membership** - Długość posiadania konta na stronie internetowej sklepu. Nie jest jednoznacznie stwierdzone, czy podajemy ją w latach, czy miesiącach.
7. **Yearly Amount Spent** - ilość pieniędzy wydanych w tym sklepie. Prawdopodobnie wyrażona w dolarach.

Spośród wszystkich kolumn wybraliśmy dwie, które naszym zdaniem, idealnie nadają się do analizy zależności liniowej. W naszym raporcie **zmienna niezależna**, to dane przechowujące **długość posiadania konta na stronie internetowej sklepu**, a **zmienna zależna** to **ilość pieniędzy wydanych w tym sklepie**.

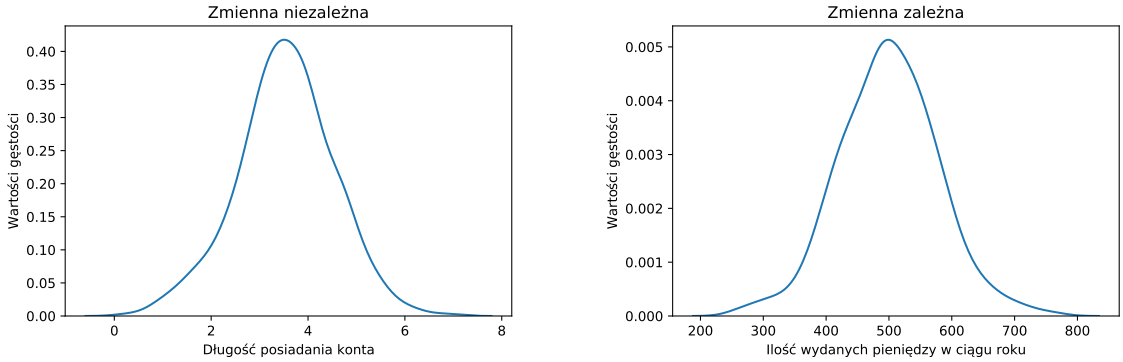
2. Analiza jednowymiarowa zmiennej zależnej oraz zmiennej niezależnej

Na początek przedstawimy nasze dane względem zmiennych na wykresach, a następnie podamy ich podstawowe miary.

Histogram oraz wykres gęstości wygląda następująco:

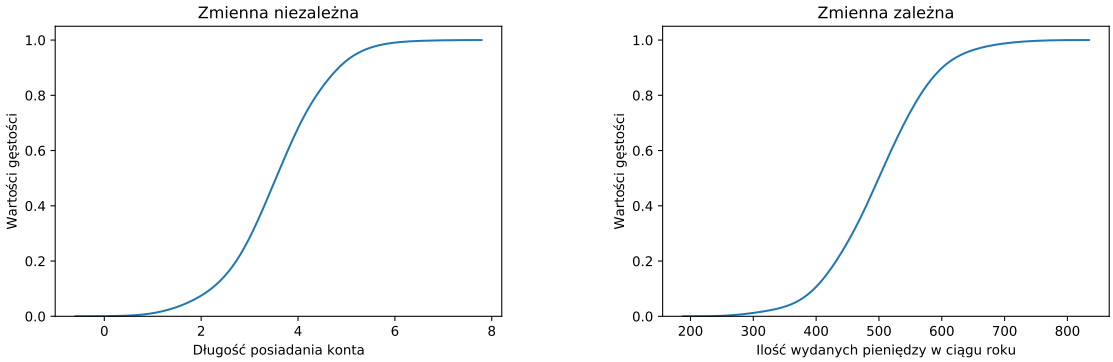


Rysunek 1. Histogramy zmiennej niezależnej i zależnej

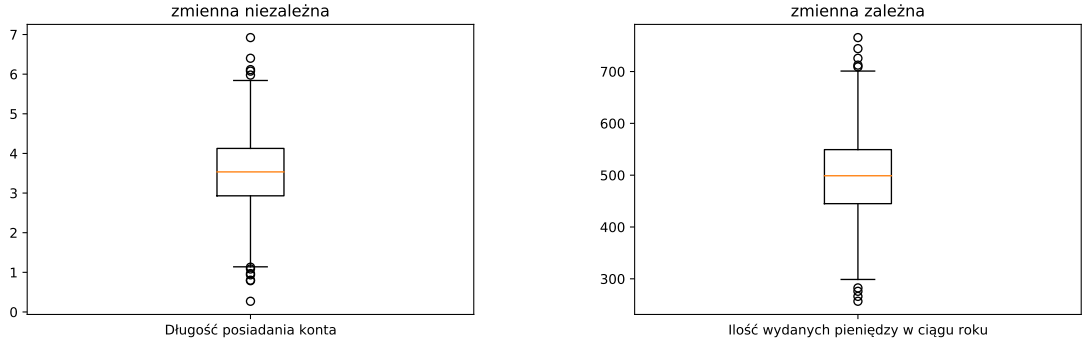


Rysunek 2. Gęstości zmiennej niezależnej i zależnej

Natomiast wykres dystrybuanty oraz pudełkowy:



Rysunek 3. Dystrybuanty zmiennej niezależnej i zależnej



Rysunek 4. Wykresy pudełkowe zmiennej niezależnej i zależnej

2.1. Podstawowe miary zmiennych

Wyróżniamy kilka podstawowych miar. W naszej analizie wyniki zaokrąglamy do części tysięcznych. Podajemy wzór oraz jego kod, który użyliśmy. Do obliczeń wykorzystaliśmy bibliotekę numpy oraz scipy.stats, które można zaimportować w Pythonie.

Użyte skróty:

x_i — każda kolejna obserwacja zmiennej

σ — wariancja

s — odchylenie standardowe

$\bar{x}$ — średnia arytmetyczna z obserwacji zmiennej

q_1 i q_3 — kwartyl pierwszy oraz trzeci, uzyskane przy wyznaczaniu podstawowych statystyk zmiennych komendą .describe() w Pythonie

μ_4 — czwarty moment centralny wyrażony wzorem: $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{n_i}$

1. Dla zmiennej niezależnej:

   a) **średnia arytmetyczna** (np.mean $= \frac{1}{n}\sum_{i=1}^{n} x_i$) $= 3.533$

   b) **średnia geometryczna** (scipy.stats.gmean $= \sqrt[n]{\sum_{i=1}^{n} x_i}$) $= 3.363$

   c) **średnia harmoniczna** (scipy.stats.hmean $= \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$) $= 3.113$

d) **Wariancja** (np.var = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$) = 0.997
e) **Odchylenie standardowe** (np.std = $\sqrt{\sigma}$) = 0.999
f) **wartość minimalna** (min = x_{min}) = 0.270
g) **wartość maksymalna** (max = x_{max}) = 6.923
h) **Rozstęp** (x_{max} - x_{min}) = 6.653
i) **Rozstęp międzykwartylowy** (scipy.stats.iqr = $q_3 - q_1$)= 1.196
j) **Współczynnik skośności** (scipy.stats.skew = $\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)(n-2)s^3}$)= -0.106
k) **Kurtoza** (scipy.stats.kurtosis = $\frac{\mu_4}{\sigma^4} - 3$) = 0.334

2. Dla zmiennej zależnej:

a) **średnia arytmetyczna** = 499.314
b) **średnia geometryczna** = 492.796
c) **średnia harmoniczna** = 485.93
d) **Wariancja** = 6278.253
e) **Odchylenie standardowe** = 79.315
f) **wartość minimalna** = 256.671
g) **wartość maksymalna** = 765.518
h) **Rozstęp** = 508.848
i) **Rozstęp międzykwartylowy** = 104.276
j) **Współczynnik skośności** = 0.035
k) **Kurtoza** = 0.447

2.2. Interpretacja otrzymanych wyników

Na podstawie histogramu oraz wykresu gęstości, czy dystrybuanty, możemy stwierdzić, że nasze dane są bardzo zbliżone do rozkładu normalnego. Aby potwierdzić tę tezę, policzyliśmy współczynnik skośności, który w obu przypadkach wyszedł dodatni, ale bliski zero. Oznacza to, że nasze zmienne są prawostronnie skośne, a nawet prawie symetryczne.
Najlepszym dowodem na potwierdzenie poprawności wyliczonej średniej jest wykres pudełkowy. Pomarańczowa linia wyznacza właśnie tą średnią, która zbliżona jest do wyliczonej analitycznie. Dodatkowo sprawdziliśmy jak zachowują się ogony naszych zmiennych w stosunku do rozkładu normalnego — w tym celu, policzyliśmy kurtozę. W obu przypadkach otrzymaliśmy wyniki większe od 0, ale bardzo do tej wartości zbliżone. Wynika z tego, że nasze zmienne mają ogony lekko węższe niż rozkład normalny, czyli więcej wartości skupionych jest wokół średnich.

3. Analiza zależności liniowej pomiędzy zmienną zależną a zmienną niezależną

3.1. Założenia

Naszą hipotezą jest występowanie zależności: $y_i = Ax_i + B + z_i$, gdzie y_i jest i-tą wartościom kolumny "Yearly Amount Spent" , x_i jest i-tą wartościom kolumny "Length of Membership" , z_i są niezależnymi zmiennymi losowymi z rozkładu $N(0, \sigma^2)$, oraz $A, B, \sigma \in R$.

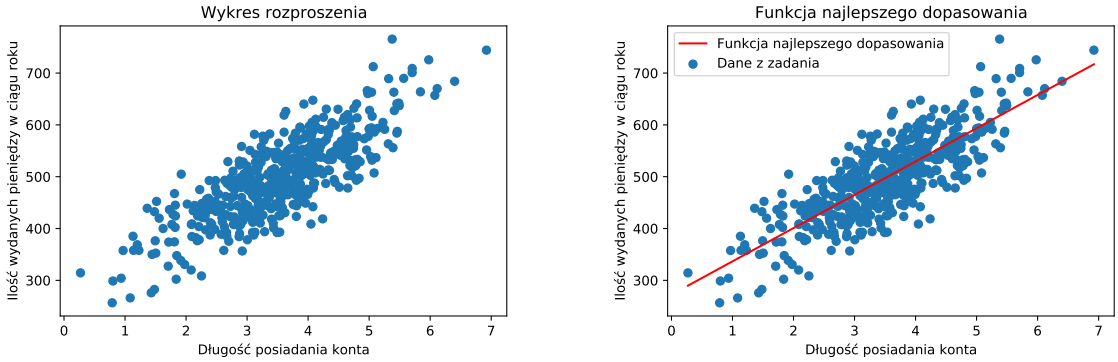
3.2. Estymacja parametrów

Naszym celem jest znalezienie wartości parametrów A i B minimalizujących sumę kwadratów błędów: $\sum_{i=1}^{n}(y_i - Ax_i - B)^2$. W tym celu przyrównujemy pochodne tego wyrażenia po parametrach do zera:

$$-\sum_{i=1}^{n} X_i(y_i - Ax_i - B) = 0 \quad , \quad -\sum_{i=1}^{n}(y_i - Ax_i - B) = 0.$$

Z powyższego układu równań otrzymaliśmy ,że wartościami minimalizującymi sumę kwadratów błędów jest: $\hat{A} = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{D}$, oraz $\hat{B} = \overline{y} - \hat{A}\overline{x}$. Dla uproszczenia notacji przyjęliśmy oznaczenia : $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $D = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

3.3. Wykres rozproszenia oraz funkcja najlepszego dopasowania



Rysunek 5. Wykres rozproszenia oraz funkcja najlepszego dopasowania do danych rzeczywistych

Za pomocą wbudowanej funkcji polyfit() w Pythonie otrzymaliśmy funkcję liniową najlepszego dopasowania wyrażoną wzorem: $64.22x + 272.4$. Do sprawdzenia poprawności metody, policzyliśmy współczynnik determinacji równy w przybliżeniu: 0.9998. Wynika z tego, że otrzymana prosta jest bardzo dobrze dopasowania, ponieważ im wynik bliższy 1, tym lepsze przybliżenie do danych.

3.4. Przypomnienie własności estymatorów

Na wykładzie dowiedliśmy, iż estymatory $\hat{A}$, oraz $\hat{B}$ są nieobciążonymi estymatorami o rozkładach normalnym i wariancji odpowiednio: $\frac{\sigma^2}{D}$, oraz $\frac{\sigma^2}{n} + \frac{\sigma^2\overline{x}^2}{D}$. Jednakże z powodu nieznajomości dokładnej wartości parametru σ^2 konieczne jest użycie w jego miejscu estymatora nieobciążonego wariancji: $S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{A}x_i - \hat{B})^2}{n-2}$. Statystyka $(\hat{A} - A)\sqrt{\frac{D}{S^2}}$ jest równoznaczna statystyce $(\hat{A} - A)\sqrt{\frac{D}{\sigma^2}} : \sqrt{\frac{\sigma^2}{S^2}}$. Ponieważ $(\hat{A} - A)\sqrt{\frac{D}{\sigma^2}}$ ma rozkład $N(0,1)$ a $\frac{(n-2)S^2}{\sigma^2}$ ma rozkład chi-kwadrat z $n-2$ stopniami swobody to nasza pierwotna statystyka $(\hat{A} - A)\sqrt{\frac{D}{S^2}}$ ma rozkład t-Studenta z $n-2$ stopniami swobody. Dokonując analogicznych przekształceń dla parametru $\hat{B}$ uzyskujemy statystyki o rozkładach t-Studenta z $n-2$ stopniami swobody dla obu parametrów.

Korzystając z symetryczności rozkładu t-Studenta , oraz działań na nierównościach otrzymujemy przedziały zawierające wartości A , oraz B z prawdopodobieństwem $1-\alpha$ dla $\alpha \in (0,1)$:

$$A \in (\hat{A} - t_{(1-\frac{\alpha}{2};n-2)}\sqrt{\frac{S^2}{D}}; \hat{A} + t_{(1-\frac{\alpha}{2};n-2)}\sqrt{\frac{S^2}{D}})$$

,

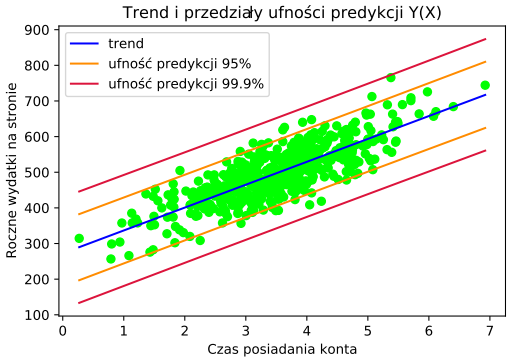
$$B \in (\hat{B} - t_{(1-\frac{\alpha}{2};n-2)}\sqrt{\frac{S^2}{n} + \frac{S^2\bar{x}^2}{D}}; \hat{B} + t_{(1-\frac{\alpha}{2};n-2)}\sqrt{\frac{S^2}{n} + \frac{S^2\bar{x}^2}{D}}) \quad .$$

Gdzie $t_{(1-\frac{\alpha}{2};n-2)}$ jest kwantylem rzędu $1-\frac{\alpha}{2}$ z rozkładu t-Studenta z $n-2$ stopniami swobody.

3.5. Predykcja

Do wyznaczania przedziału zawierającego zmienną objaśnianą z prawdopodobieństwem $1-\alpha$ dla nowej zmiennej objaśniającej wykorzystaliśmy wzór podany na laboratoriach:

$$Y_{(x_0)} \in (\hat{A}x_0 + \hat{B} - t_{(1-\frac{\alpha}{2};n-2)}\sqrt{\frac{(n+1)S^2}{n} + \frac{S^2(x_0-\bar{x})^2}{D}}; \hat{A}x_0 + \hat{B} + t_{(1-\frac{\alpha}{2};n-2)}\sqrt{\frac{(n+1)S^2}{n} + \frac{S^2(x_0-\bar{x})^2}{D}})$$



Rysunek 6. Wykres trendu i przedziały ufności predykcji Y(X)

W wizualizacji wykorzystaliśmy $\alpha_1 = 0.05$, dla którego wygenerowana pomarańczowa obramówka z prawdopodobieństwem 95% będzie zawierać nową parę zmiennych x_0, y_0 , oraz $\alpha_2 = 0.001$, dla którego wygenerowana czerwona obramówka z prawdopodobieństwem 99.9% będzie zawierać nową parę zmiennych x_0, y_0. Oznacza to, że generując kolejne 500 par zmiennych x_i, y_i z ponad 60-procentowym prawdopodobieństwem wszystkie znajdą się we wnętrzu czerwonego obrysu.

3.6. Ocena poziomu zależności

Do oceny zależności między zmiennymi wykorzystaliśmy następujące wzory:

1. **Współczynnik korelacji Pearsona** — Wartość współczynnika korelacji mieści się w przedziale domkniętym [-1, 1]. Im większa jest jego wartość bezwzględna, tym silniejsza jest zależność liniowa między zmiennymi. Wyraża się wzorem:

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

gdzie:
x_i i y_i — kolejne obserwacje zmiennej niezależnej i zależenej

$\bar{x}$ i $\bar{y}$ — średnie zmiennej niezależnej i zależnej

Otrzymaliśmy wynik równy: 0.809, który mogliśmy przewidzieć po wykresie rozproszenia.

2. **Suma kwadratów błędów SSE** — Jest to suma różnic wartości obserwacji zmiennej zależnej i estymatora tej zmiennej podniesionych do kwadratu. Wyraża się wzorem:

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

Otrzymany przez nas wynik, to 461.214.

3. **Suma kwadratów odchyleń regresyjnych SSR** — Suma różnic estymatora obserwacji zmiennej zależnej i wartości średniej tej zmiennej podniesionych do kwadratu. Wyraża się wzorem:

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

Otrzymany przez nas wynik, to 2058302.246.

4. **całkowita suma kwadratów SST i współczynnik determinacji** R^2 — Całkowita suma kwadratów to suma kwadratów błędów SSE oraz suma kwadratów odchyleń regresyjnych SSR:

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

Wyżej wspomniany współczynnik determinacji jest miarą stopnia, w jakim model pasuje do próby. Przyjmuje wartości z przedziału [0,1]. Najczęściej wyrażą się w procentach. Dopasowanie modelu jest tym lepsze, im wartość R^2 jest bliższa 1. Wyraża się wzorem:

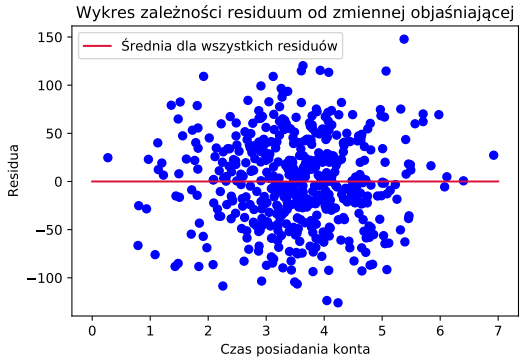
$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{SEE}{SST}$$

Otrzymaliśmy SST = 2058763.461

4. Analiza residuów

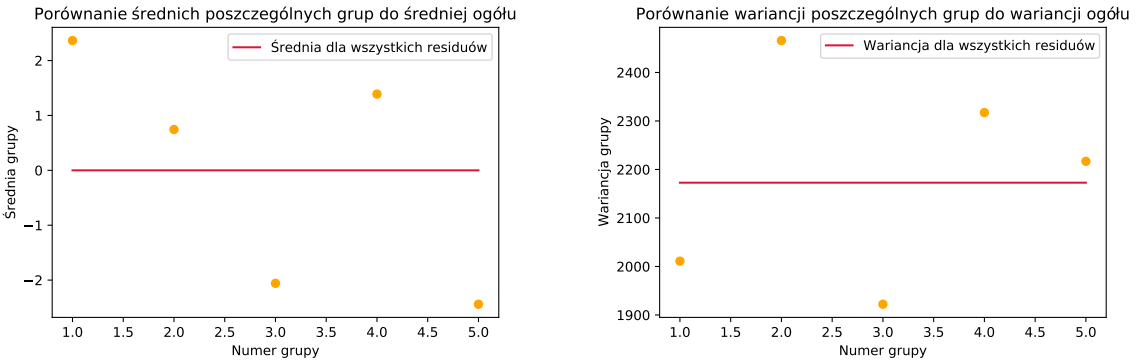
4.1. Średnia i stałość wariancji

Na podstawie poniższej wizualizacji zależności residuów od zmiennej objaśniającej dostrzec można oscylacje wokół zera ze stosunkowo stabilną amplitudą różnicy. Test wzrokowy nie wykazał błędności założeń początkowych, jednak nadal wymagają one sprawdzenia numerycznego.



Rysunek 7. Wykres zależności residuum od zmiennej objaśniającej

Wyznaczyliśmy empiryczną średnią oraz wariancje residuów na odpowiednio: -5.32×10^{-13} oraz 2172.75. Następnie posortowaliśmy zmienne objaśniające wraz z odpowiadającymi im residuami oraz podzieliliśmy je na pięć grup , po sto residuów każda i ponownie wyliczyliśmy dla każdej grupy jej empiryczną średnią i wariancję. Otrzymujemy w wyniku tej procedury poniższe wykresy wraz z zaznaczonymi poziomymi liniami wartości uzyskane dla wszystkich wartości.



Rysunek 8. Wykres porównania średnich i wariancji dla poszczególnych grup

Średnie w grupach odbiegają od średniej ogólnej na nie więcej, niż 3 jednostki, co przy residuach przekraczających 100 jest zadowalającym wynikiem. Natomiast z wykresu wariancji gdzie względna różnica sięga niemal 15% od wartości wariancji dla wszystkich danych. A największa różnica pomiędzy grupami przekracza 25% wariancji ogólnej. Zdaje się to podważać nasze założenie o stałości wariancji.

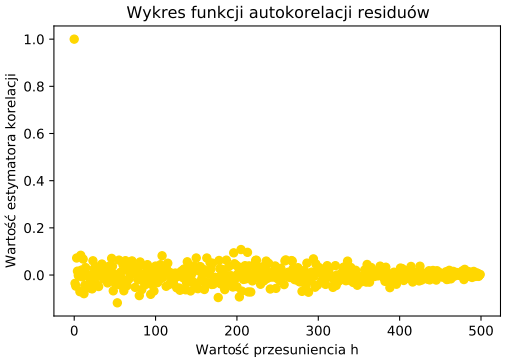
4.2. Niezależność i normalność

Do badania niezależności residuów wykorzystaliśmy empiryczny estymator kowariancji i korelacji:

$$Cov(h) = \frac{1}{n}\sum_{i=1}^{n-|h|}(x_{i+|h|} - \overline{x})(x_i - \overline{x})$$

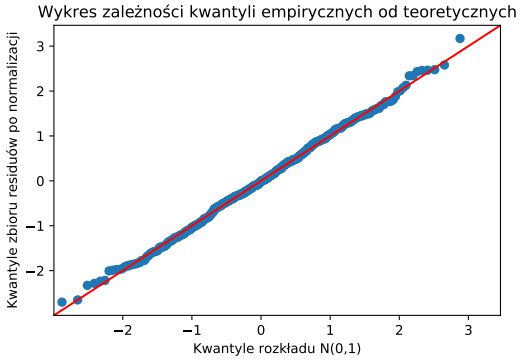
,

$$Cor(h) = \frac{Cov(h)}{Cov(0)}$$



Rysunek 9. Wykres funkcji autokorelacji residuów

.
Na powyższym wykresie szum oscylujący w pobliżu zera dla $h \neq 0$ sugeruje brak istotnej zależności liniowej pomiędzy kolejnymi residuami.
Następnie w celu sprawdzenia normalności rozkładu residuów sortujemy je w ciąg rosnący , odejmujemy od każdej wartości średnią residuów , a następnie wynik dzielimy przez pierwiastek z ich wcześniej wyliczonej wariancji. Tak uzyskany zbiór liczb traktujemy jako empiryczne kwantyle pewnego rozkładu. Z założenia o normalności residuów wynika, że po tych przekształceniach rozkładem tym powinien być rozkład $N(0,1)$, więc porównujemy nasze kwantyle empiryczne z kwantylami teoretycznymi rozkładu $N(0,1)$.



Rysunek 10. Wykres zależności kwantyli empirycznych od teoretycznych

Ułożenie się punktów wzdłuż prostej $x = y$ sugeruje brak przesłanek do odrzucenia naszego założenia o normalności residuów. Jednak wymagane jest jeszcze numeryczne potwierdzenie. W tym celu wykorzystaliśmy test Kołmogorowa-Smirnowa polegającego na analizie maksimum różnicy pomiędzy dystrybuantą teoretyczną (Hipotetyczną) a dystrybuantą empiryczną (wynikającą z danych). Zakładając poziom ufności 95% wartością krytyczną testu będzie 0.060821. Ponieważ statystyka używana w tym teście $sup_{x \in R}|F_{emp}(x) - F_N(x)|$ wyniosła 0.0208875, to brak nam przesłanek do odrzucenia hipotezy o normalności rozkładu residuów.

5. Wnioski

Pomimo prawdopodobnego braku niezależności wariancji residuów od zmiennej objaśniającej, oznaczającego tym samym niepoprawność zakładanego przez nas modelu , to ich względnie niewielkie fluktuacje umożliwiły utrzymanie użyteczności naszego modelu jako stosunkowo dobrego przybliżenia. Jednak dla ewentualnej dalszej analizy konieczne byłoby uwzględnienie tego faktu do konstrukcji modelu lepiej opisującego posiadane dane.