Analiza zależności liniowej wyniku kobiet w martwym ciągu i w przysiadzie w Trójboju Siłowym

Magdalena Szymkowiak, Adrian Sobczak

22 grudnia 2022

Spis treści

1 Wstęp

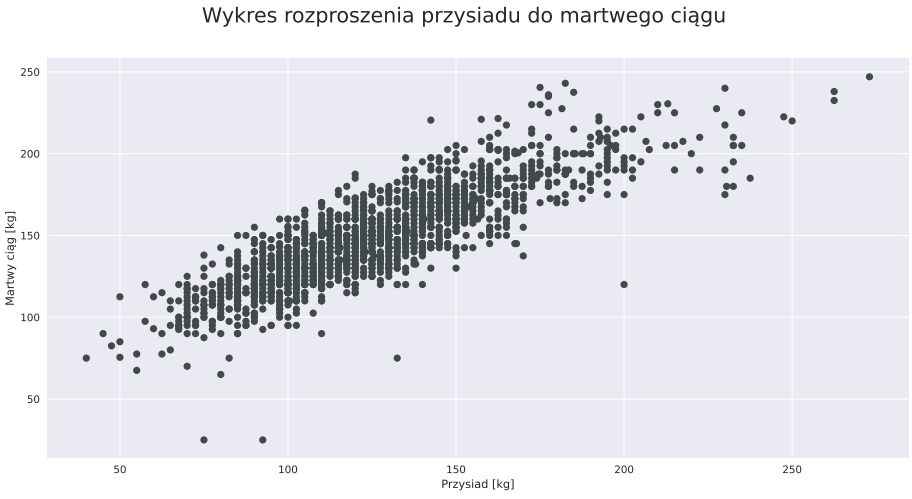
Zajmiemy się badaniem liniowej zależności zachodzącej pomiędzy dwoma z trzech boi zaliczających się do Trójboju Siłowego. Bojami tymi będą martwy ciąg oraz przysiad ze sztangą na plecach.

Ważnym punktem sprawozdania jest założenie a priori, o spełnieniu przez dane założeń wymaganych do skorzystania z klasycznego modelu regresji liniowej. W ostatniej części sprawdzimy czy faktycznie tak jest.

Będziemy analizować dane pochodzące ze strony *openpowerlifting.org*. Dane, które pobraliśmy znaleźliśmy na kaggle.[2] Zbiór zawiera wyniki wszystkich zawodników startujących od czerwca 2012 roku do stycznia 2019, na zawodach organizowanych przez różne federacje na całym świecie.

Zajmiemy się analizą wyłącznie wyników kobiet dla jednej z najbardziej rozpoznawalnych federacji IPF. Dodatkowo wybraliśmy te wyniki, które kobiety uzyskały w startowaniu RAW, tzn. bez specjalnych kostiumów pomagających w wykonaniu bojów oraz bez bandaży wiązanych na kolanach. Jedynym dopuszczalnym sprzętem, w ramach bezpieczeństwa zawodników są neopreny na kolana, owijki na nadgarstki oraz pas.[3]

Do analizy wybraliśmy najlepsze podejścia wykonane na zawodach z przysiadu oraz martwego ciągu dla poszczególnego zawodnika. Najlepszym podejściem jest jedno z trzech, w którym został podniesiony największy ciężar. W dalszej części analizy będziemy traktować słowa *martwy ciąg* i *przysiad* jako wynik z danego boju w najlepszym podejściu z trzech. Takich wyników mamy dokładnie 1699.



Rysunek 1: Wykres rozproszenia martwego ciągu od przysiadu.

2 Analiza jednowymiarowa zmiennych

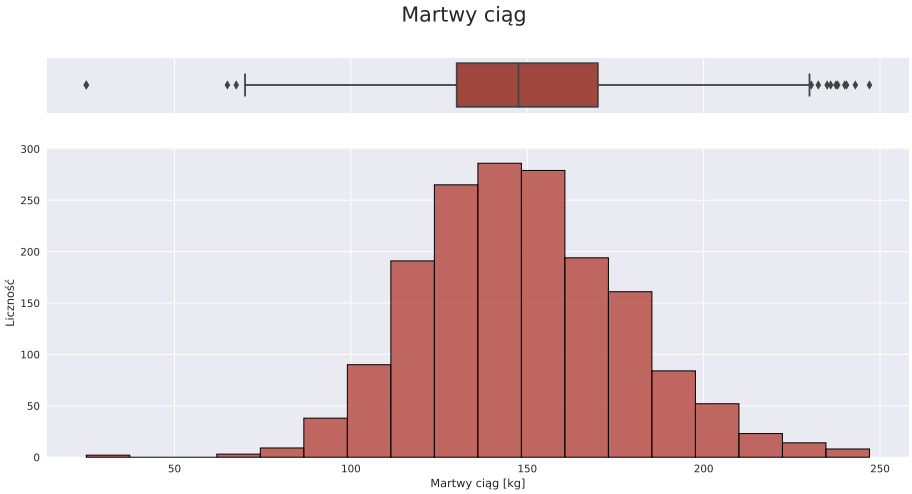
W tej części analizy przyjrzymy się jednowymiarowym rozkładom oraz podstawowym statystykom dla obu zmiennych.

	Martwy ciąg [kg]	**Przysiad [kg]**
$\bar{x}$	149	126.14
$x_{\mathbf{m}}$	147.5	122.5
Q_1	130	102.5
Q_3	170	145
IQR	40	42.5
R	222	233
s^2	859.14	1067.47
s	29.31	32.67
d_1	23.21	25.42
ν	0.2	0.26
α	0.23	0.68
K	0.3	0.98

Tabela 1: Podstawowe miary położenia, rozproszenia, skośności oraz spłaszczenia dla analizowanych danych

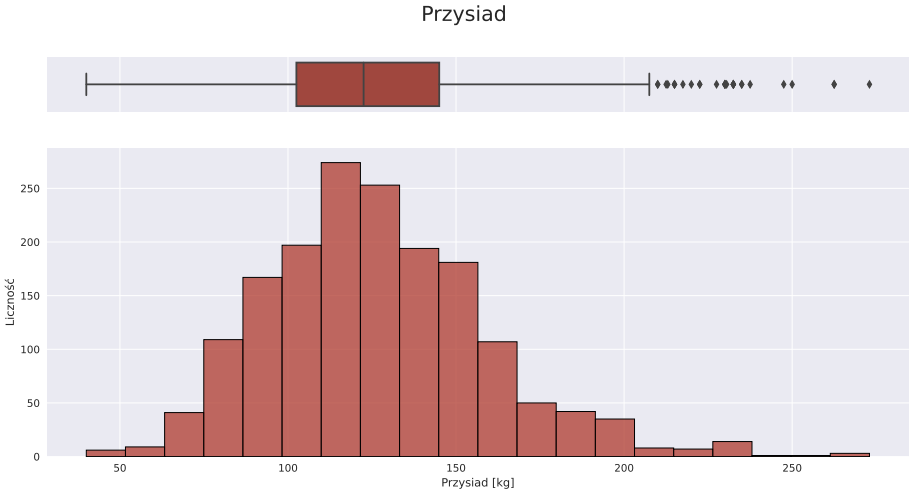
2.1 Opis i wizualizacja danych

Do wykonania przysiadu ze sztangą na plecach zawodnik układa sobie sztangę poziomo na barkach, trzymając gryf rękami po obu stronach. Następnie wykonuje przysiad, obniżając tułów do momentu, w którym górna część uda w stawie biodrowym będzie poniżej wierzchołka kolan. Podczas zawodów ważne jest aby pamiętać o komendach, które dają sędziowie. Niezbędne są one do zaliczenia podejścia.[3]



Rysunek 2: Histogram i boxplot dla rozkładu jednowymiarowego martwego ciągu.

Wykonanie martwego ciągu polega na podejściu do sztangi, która leży poziomo na ziemi oraz podniesienie jej obiema rękami, aż do wyprostowanej pozycji. Wyprostowana pozycja to taka, w której nogi są wyprostowane w kolanach a barki odciągnięte do tył. W tym boju również należy pamiętać o komendach. Ważnym aspektem jest także fakt, że nie można rzucać sztangi na ziemie. Sztangę można puścić dopiero w momencie, kiedy znajdzie się z powrotem na ziemi. [3]



Rysunek 3: Histogram i boxplot dla rozkładu jednowymiarowego przysiadu.

2.2 Miary rozkładu

W tabeli (1) przedstawiliśmy podstawowe miary położenia, rozproszenia, skośności oraz spłaszczenia dla każdego z rozpatrywanych boi. Poniżej objaśnimy każdą z miar:

Miary położenia

- Średnia arytmetyczna

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

- Mediana

  gdy n nieparzyste
  $x_{\mathbf{m}} = x_{((n+1)/2)}$,

  gdy n parzyste
  $x_{\mathbf{m}} = \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right)$.

- Pierwszy i trzeci kwartyl, kolejno Q_1 i Q_3, to odpowiednio mediana z prób $\{x_i : x_{(1)} \leqslant x_i \leqslant x_{\mathbf{m}}\}$ oraz $\{x_i : x_{\mathbf{m}} \leqslant x_i \leqslant x_{(n)}\}$.

Miary asymetrii

- Współczynnik skośności

  $$\alpha = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3.$$

Miary spłaszczenia

- Kurtoza (*Fishera*)

  $$K = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3.$$

Miary rozproszenia

- Rozstęp międzykwartylowy

  $\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$,

- Rozstęp

  $R = x_{(n)} - x_{(1)}$,
  gdzie $x_{(n)}, x_{(1)}$ to największa i najmniejsza wartość z próby.

- Wariancja

  $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- Odchylenie standardowe

  $s = \sqrt{s^2}$.

- Odchylenie przeciętne od wartości średniej

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

- Współczynnik zmienności

  $\nu = \frac{s}{\bar{x}} \quad (\cdot\, 100\%)$.

gdzie: n — długość próby, $x_{(i)}$ — i-ta realizacja z uporządkowanej próby.

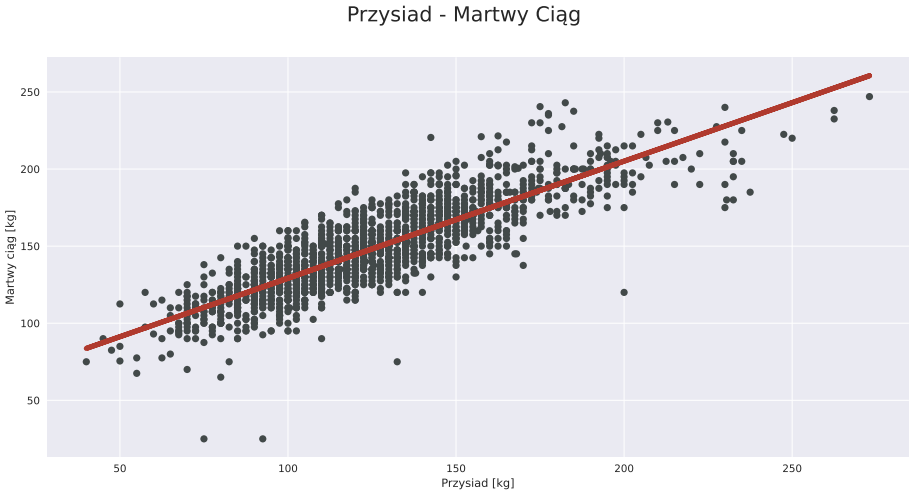
2.3 Wnioski

Rozkład wyników zarówno w martwym ciągu jak i w przysiadzie jest prawoskośny i platokurtyczny. W obu bojach występuje bardzo duże rozproszenie wyników. Podejrzewamy, że może wynikać z wagi ciała oraz wieku zawodniczek. Duży rozrzut wyników powoduje całkiem duże odchylenie przeciętne od wartości średniej. Zauważalna jest również różnica między średnimi arytmetycznymi a medianami. Wskazuje ona na to, że dane nasze mają więcej lepszych wyników zawyżających średnią.

3 Analiza zależności liniowej

W tej części zajęliśmy się badaniem zależności pomiędzy przysiadem a martwym ciągiem.

3.1 Prezentacja danych



Rysunek 4: Wykres rozproszenia z naniesioną prostą regresji liniowej dla martwego ciągu oraz przysiadu.

3.2 Statystyki zależności

W tabeli 2 przedstawiamy wszystkie istotne statystyki dla modelu regresji. Przedziały ufności C_{β_0} i C_{β_1}, zostały wyliczone na poziomie ufności $\alpha = 0.05$.

	Wartość statystyki
β_0	53.29
β_1	0.76
C_{β_0}	[47.73 , 58.85]
C_{β_1}	[0.72 , 0.8]
SSE	414490.83
SSR	1045185.27
SST	1459676.1
r^2	0.72
ρ	0.85

Tabela 2: Statystyki opisujące zależność liniową między danymi.

Poniżej, objaśniamy każdą ze statystyk.

Estymacja punktowa współczynników regresji

- Współczynnik kierunkowy

$$\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Współczynnik przesunięcia

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Estymacja przedziałowa współczynników regresji

- Przedział dla współczynnika kierunkowego

  $C_{\beta_1} = [\beta_1 + a, \beta_1 - a]$, gdzie

$$a = t \cdot \sqrt{\mathrm{Var}(\varepsilon) \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

- Przedział dla współczynnika przesunięcia

  $C_{\beta_0} = [\beta_0 + a, \beta_0 - a]$, gdzie

$$a = t \cdot \sqrt{\frac{\mathrm{Var}(\varepsilon))}{\sum_{i=1}^{n}(x_i - \bar{x}^2}}$$

  gdzie t to kwantyl rzedu $\alpha/2 = 0.025$ z rozkładu t-studenta o $n-1$ stopniach swobody,

  $\mathrm{Var}(\varepsilon)$ to wariancja próbkowa liczona z błędów regresji.

  Pozostałe oznaczenia tak samo jak w podsekcji 2.2.

Pozostałe miary zmienności

- Suma kwadratów błędów

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

- Suma kwadratów regresji

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

- Całkowita suma kwadratów

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

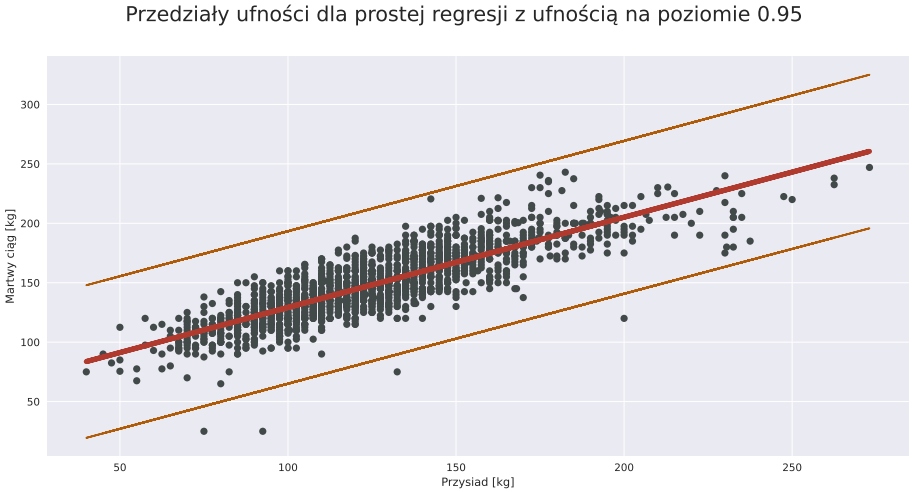
- Współczynnik determinacji

$$r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}$$

- Współczynnik korelacji Pearsona

$$\rho = \frac{\sum_{i=0}^{n}\left[(x_i - \bar{x})(y_i - \bar{y})\right]}{s_x s_y}$$

3.3 Przedziały ufności



Rysunek 5: Wykres rozproszenia z naniesioną prostą regresji liniowej oraz przedziałami ufności modelu dla poziomu ufności 0.95.

3.4 Wnioski

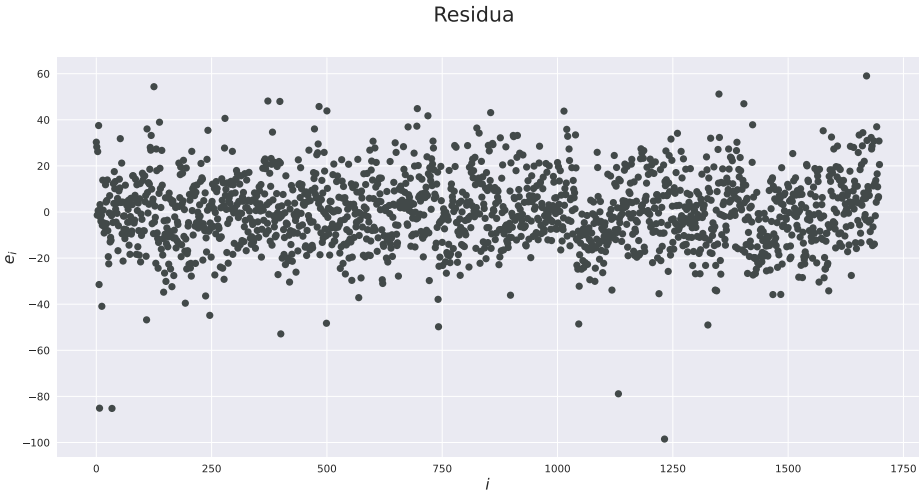
Spoglądając na statystyki zależności jesteśmy w stanie stwierdzić, że dopasowanie jest dobre. Korelacja pomiędzy zmiennymi jest dodatnia i duża. Z wykresu **??** widać, że większość wyników mieści się w przedziale ufności.

4 Analiza residuów

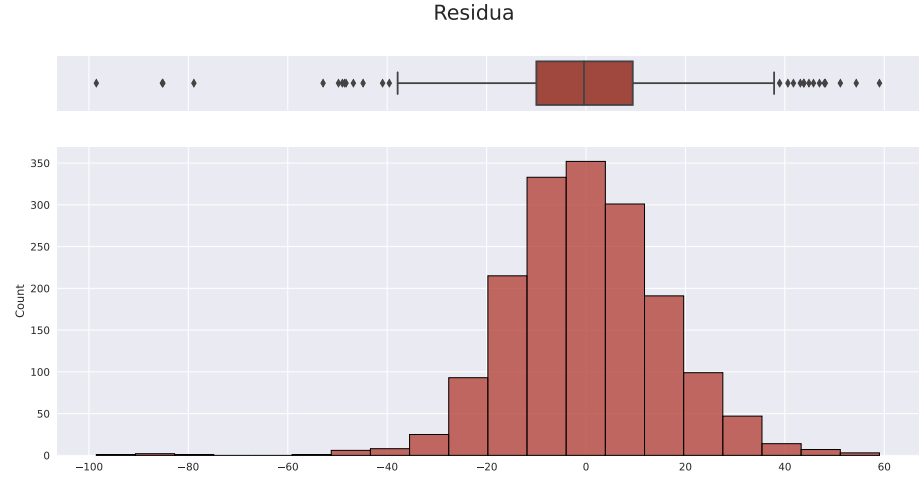
W tej części za pomocą analizy residuów sprawdzimy czy są spełnione założenia pozwalające nam na używanie pokazanych wcześniej estymatorów regresji liniowej.

4.1 Wizualizacja residuów

Na rysunku 4.1 przedstawiono wykres rozproszenia rediduów, a na rysunku 4.1 boxplot oraz histogram liczności.



Rysunek 6: Wykres residuów dla kolejnych próbek.



Rysunek 7: Histogram liczności i boxplot dla residuów.

4.2 Sprawdzenie założeń

Aby użycie modelu regresji było uzasadnione, zjawisko musi spełnić założenia potrzebne do jego wyprowadzenia. Poniżej sprawdzone jest każde z nich.

Średnia

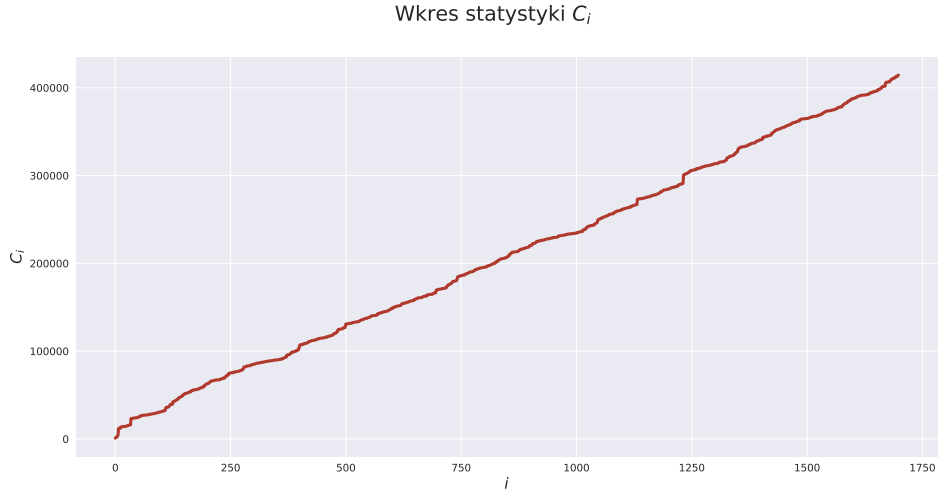
Estymator średniej dla residuów jest równy

$$\hat{\mu} = 2.71 \cdot 10^{-15},$$

co stanowi dobre numeryczne przybliżenie zera. Na wykresie 4.1 widzimy również, że dane są rozłożone równomiernie. Uznajemy więc nasze założenie za spełnione.

Stała wariancja

Korzystając z metody przedstawionej w artykule [1] sprawdzamy statystykę C_i i szukamy punktu zmiany reżimu $\hat{l}$. Dla naszych danych $\hat{l} = 5$. Patrząc na wykres 4.2 oraz na boxplot 4.1 nie wydaje się być dziwnym, że właśnie te obserwacje mają większą wariancję i zostały wykryte przez algorytm z artykułu. Przyglądając się wykresowi statystyki C_i widzimy, że wariancja tylko na krótkich odcinkach rośnie szybciej niż ogólna tendencja wzrostowa, która wygląda na całkowicie liniową. Zjawisko to możemy uzasadnić zatrzymywaniem się zawodników na trudnych psychicznie do pokonania obciążeniach jak np 200 kg. Zawodnik polepsza wtedy swoje wyniki w drugim ćwiczeniu, a w pierwszym nie może pokonać wyznaczonej przez siebie bariery. Dlatego, patrząc na ogólną tendencję, założenie o stałej wariancji uznajemy za spełnione.



Rysunek 8: Wykres stastyki C_i od indeksów obserwacji.

Obserwacje odstające

Na boxplocie który jest elementem wykresu (4.1) szczególną uwagę przykuwają trzy najmniejsze obserwacje odstające. Są to zawodnicy którzy osiągnęli znacznie lepszy wynik w przysiadzie niż w martwym ciągu. Co ciekawe, w drugą stronę to zjawisko nie zachodzi. Uwzględnimy te obserwacje w dalszej analizie.

Rozkład normalny

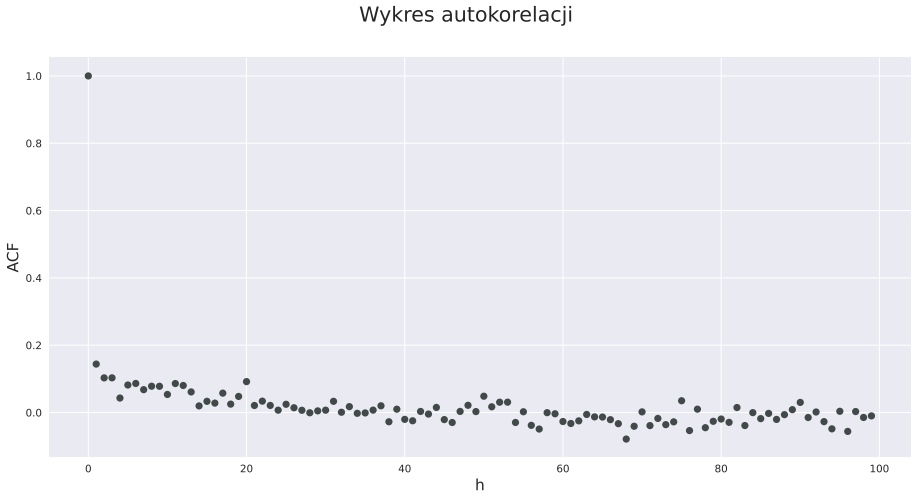
Przeprowadzone testy, statystyki testowe oraz p-wartości umieszczone są w tabeli 3. Jedynie test Kołmogorova-Smirnov'a nie daje nam podstaw do odrzucenia hipotezy zerowej o normalności rozkładu residuów. Testy Shapiro-Wilka oraz Jarque-Bera dają nam przeciwny wynik. Wykazujemy jednak większe zaufanie do dwóch ostatnich testów (zwłaszcza test Shapiro-Wilk'a który świetnie spisuje się dla niedużych próbek) wstępnie odrzucamy hipotezę o normalności residuów.

	statystyka testowa	p-wartość
Kolmogorov-Smirnov	0.28	0.14
Shapiro-Wilk	0.98	0.00
Jarque-Bera	391.52	0.00

Tabela 3: Statystyki opisujące zależności między danymi.

Autokorelacja

Odrzuciwszy wcześniej założenie o normalności residuów, nie posiadamy już niestety implikacji z nieskorelowania w niezależność. Patrząc jednak na funkcję autokorelacji widzimy, że z pozoru bliskie zero wartości dla $h \neq 0$ wraz z wzrostem h wykazują lekką tendencję spadkową co nie powinno być widocznie przy faktycznie nieskorelowanej próbce. Dlatego odrzucamy hipotezę o nieskorelowanych residuach.



Rysunek 9: Wykres autokorelacji dla residuów dopasowania prostej.

4.3 Wnioski

Niestety nie udało nam się spełnić wszystkich założeń potrzebnych do zasadnego skorzystania z klasycznego modelu regresji liniowej. Jednak wyniki testów na normalność i funkcja autokorelacji, które obaliły nasze założenia, nie wskazały tego w sposób absolutnie jednoznaczny. Pozbywając się obserwacji odstających wyniki zapewne byłyby jeszcze bardziej dyskusyjne. Z tego powodu nie uważamy, że dopasowanie regresji jest w tym przypadku niewłaściwym wyborem. Przetworzenie danych oraz inna interpretacja wyników, może przynieść rezultat na korzyść klasycznego modelu regresji liniowej.

5 Podsumowanie

Zastosowaliśmy klasyczny model regresji do danych opisujących zależność między najlepszymi podejściami w martwym ciągu i przysiadzie. Między tymi zmiennymi zależność w oczywisty sposób występuje co widać na wykresie 1. W sekcji 4 zbadaliśmy czy założenia wymagane do zastosowania klasycznego modelu regresji są spełnione. Według naszej interpretacji nie są, jednak nie wykluczamy tego jednoznacznie przez wzgląd na niejednoznaczność testów (3) i bliskie zeru wartości w funkcji autokorelacji (4.2).

Bibliografia

[1] Regime variance testing - a quantile approach. “Gajda Janusz, Sikora Grzegorz, Wyłomańska Agnieszka”. W: (1984).

[2] OPENPOWERLIFTING. *Powerlifting Database*. URL: https://www.kaggle.com/datasets/open-powerlifting/powerlifting-database.

[3] Kolegium Sędziów Trójboju Siłowego PZKFiTS. *Przepisy organizacji i sędziowania zawodów Trójboju Siłowego*. URL: https://www.pzkfits.pl/wp-content/uploads/2019/01/IPF_rulebook_2019_pl-v16.pdf.