Waga trójboisty a jego wyniki siłowe.

Komputerowa analiza szeregów czasowych
Raport 1.

Emil Olszewski, Artur Sadurski

22 grudnia 2023

Streszczenie

Przedmiotem analizy są dane ze zbioru zawierającego informacje na temat męskich trójboistów zrzeszonych w ramach federacji IPF. Dane zostały udostępnione na warunkach licencji GNU AGPLv3. Przeanalizowano zależność pomiędzy masą ciała zawodnika a jego wynikiem w kateogrii RAW w poszczególnych bojach tj. *wyciskaniu na ławce*, *przysiadzie ze sztangą* oraz *martwym ciągu* jak i wyniku *total* będącego sumą rezultatów z trzech wcześniej wymienionych bojów. Przedstawiono oraz opisano metody statystyczne jak i te z dziedziny regresji użyte do określenia zależności pomiędzy danymi. W wyniku analizy zaobserwowano dodatnią korelację pomiędzy tymi zmiennymi, jednakże *dobre* dopasowanie modelu liniowego uzyskano dopiero po transformacji logarytmicznej zmiennej niezależnej.

1 Aparatura

Do analizy danych użyto języka *Julia* w wersji *1.9.3* wraz z następującymi bibliotekami:

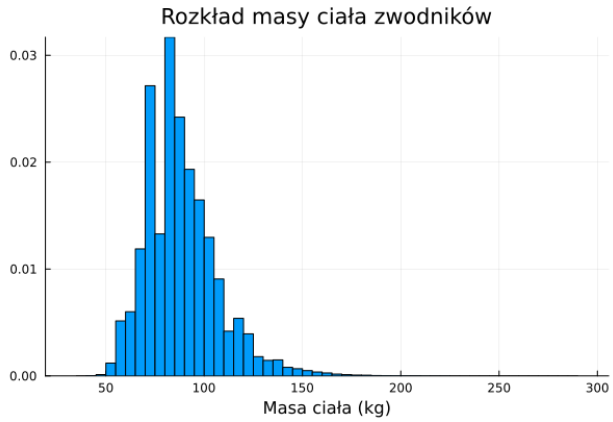
- **DataFrames.jl, Statistics.jl** - analiza danych.
- **Plots.jl** - wykresy i wizualizacja.
- **GLM.jl** - model regresji liniowej.
- **HypothesisTests.jl** - testy statystyczne.

2 Opis danych

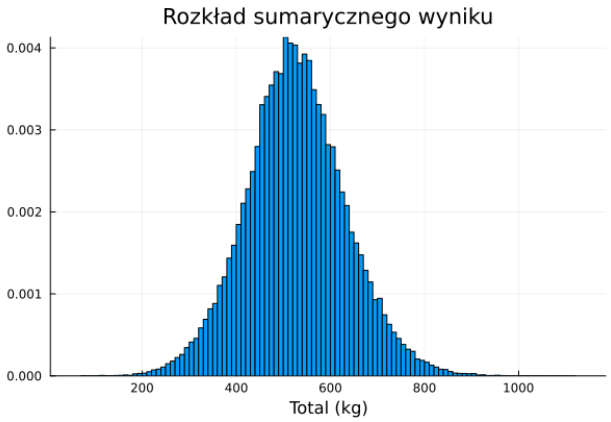
Pod uwagę wzięto tylko zawodników płci męskiej, dla których dostępny był pełen zestaw danych dotyczący wyników uzyskanych w każdym z trzech bojów. Ograniczono się dodatkowo do cenzusu wiekowego w przedziale od 16 do 40 lat oraz rozpatrywano tylko wyniki uzyskane w kategorii RAW (kategoria, która zabrania używania sprzętu dającego przewagę mechaniczną np. koszulek do wyciskania, kaftanów itd. Jest to klasyczna kategoria trójboju siłowego).
Skoncentrowano się na dwóch kluczowych zmiennych:

- **BodyweightKg** (**Masa ciała**): Ta zmienna niezależna reprezentuje masę ciała zawodnika w kilogramach. Masa ciała jest istotnym parametrem w trójboju siłowym, ponieważ klasyfikuje zawodników w odpowiednie kategorie wagowe i może wpływać na ich wydajność w zawodach.
- **TotalKg** (**Całkowity wynik**): Jako zmienna zależna, całkowity wynik odnosi się do sumy maksymalnych ciężarów, które zawodnik podniósł w trzech próbach: przysiadzie, wyciskaniu leżąc i martwym ciągu. Jest to główny wskaźnik wydajności w trójboju siłowym, odzwierciedlający siłę i umiejętności zawodnika. W dalszej części raportu będziemy używać określeń takich jak **Wynik sumaryczny**, **Wynik total** czy po prostu **total**.

Ze względu na występujące powszechnie duplikaty zmiennej niezależnej wpierw uśredniono wartości zmiennej zależnej dla takiej samej wagi zawodnika tak, aby uzyskać jednoznaczność pomiędzy zbiorem zmiennych niezależnych i niezależnych. Następnie, aby pozbyć się szumu wygładzono dane metodą średniej ruchomej o oknie szerokości 9. Tym samym z początkowego zestawu danych o długości 107416 uzyskano próbkę długości 704.



(a) Zmienna niezależna



(b) Zmienna zależna

Rysunek 1: Histogramy przedstawiające rozkłady zmiennych

Zmienna		Niezależna	Zależna
		Masa Zawodnika	**Wynik total**
Średnia		$88,4$	$525,6$
Mediana		$86,4$	$522,5$
Q1		$74,6$	$457,5$
Q3	[kg]	$98,3$	$592,5$
Minimum		$37,9$	$75,0$
Maksimum		$285,0$	$1110,0$
Odchylenie stand.		$18,5$	$106,1$
Skośność		1.11	0.12
Kurtoza (nadwyżkowa)		2.65	0.36

Tabela 1: Statystyki dla Masy Zawodnika i Wyniku Total

3 Analiza jednowymiarowa zmiennej zależnej i niezależnej

Na rysunku 1 przedstawiono histogramy, zaś w tabeli 1 podstawowe statystyki dla obu zmiennych.

Przy analizie powyższych statystyk warto zwrócić szczególną uwagę na dwie ostatnie, czyli **skośność** i **kurtozę**. Prawostronna skośność zmiennej niezależnej ma proste wyjaśnienie. W zawodach trójbojowych celem wyrównania szans zawodników o różnej budowie ciała i predyspozycjach stosuje się kategorie wagowe. W większości tego typu zawodów ostatnią kategorią wagową jest kategoria typu *120+*. Zawodnicy kwalifikujący się do tej kategorii będą dążyli do zwiększania swojej masy ciała (zakładając dodatnie skorelowanie masy z wynikami), gdyż nie grozi im wpadnięcie do wyższej kategorii wagowej. Tym samym obserwujemy prawostronną skośność w rozkładzie masy ciała zawodników.

4 Analiza zależności liniowej pomiędzy zmienną zależną i niezależną

4.1 Metoda średniej ruchomej (Moving Average)

Średnia ruchoma jest techniką wygładzania danych, która polega na obliczaniu średniej z określonej liczby kolejnych wartości z serii czasowej. Jest to powszechnie stosowana metoda do usuwania krótkoterminowych fluktuacji i uwydatniania długoterminowych trendów lub cykli.

4.1.1 Prosta średnia ruchoma (SMA)

$$SMA_t = \frac{1}{n} \sum_{i=t-n+1}^{t} \hat{y}_i$$

gdzie SMA_t to średnia ruchoma w czasie t, n to rozmiar okna, a $\hat{y}_i$ to wartości serii.

4.1.2 Zastosowanie w analizie

Średnia ruchoma może być używana do wygładzenia szeregów czasowych przed przeprowadzeniem dalszej analizy, takiej jak estymacja trendów czy sezonowości. Pomaga również w redukcji efektu przypadkowych wahań danych i w wizualizacji ogólnego kierunku zmian w danych.

4.1.3 Wybór rozmiaru okna

Rozmiar okna n ma znaczący wpływ na wyniki wygładzania. Zbyt małe okno może nie wyeliminować wszystkich niepożądanych fluktuacji, podczas gdy zbyt duże może spowodować zbytnie zatarcie użytecznych informacji. Wybór optymalnego rozmiaru okna zależy od charakterystyki danych oraz celu analizy.

4.2 Klasyczny model regresji liniowej

Model regresji liniowej to prosty, lecz potężny model statystyczny służący do przewidywania wartości zmiennej zależnej Y na podstawie wartości zmiennej niezależnej X. Matematycznie, klasyczny model regresji liniowej wyraża się wzorem:

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

gdzie:

- Y_i to wartość zmiennej zależnej dla i-tej obserwacji.
- x_i to wartość zmiennej niezależnej dla i-tej obserwacji.
- β_0 to wyraz wolny (intercept).
- β_1 to współczynnik kierunkowy (slope).
- ϵ_i to błąd losowy dla i-tej obserwacji.

Założenia tego modelu to:

- Liniowość względem parametrów.
- $E(\epsilon_i) = 0$: średnia wartość błędu losowego jest równa 0 dla wszystkich i.
- $Var(\epsilon_i) = \sigma^2$: stała wariancja błędów dla wszystkich obserwacji.
- $Cov(\epsilon_i, \epsilon_j) = 0$ dla $i \neq j$: błędy są niezależne między obserwacjami.
- Opcjonalnie, ϵ_i ma rozkład normalny, co nie jest konieczne, ale często przyjmowane dla uproszczenia.

4.3 Metoda najmniejszych kwadratów (MNK)

Metoda najmniejszych kwadratów (MNK) to technika estymacji parametrów w modelu regresji, która minimalizuje sumę kwadratów różnic między obserwowanymi a modelowanymi wartościami zmiennej zależnej. Funkcja kosztu, którą minimalizujemy, wyraża się wzorem:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

gdzie $\hat{y}_i = \beta_0 + \beta_1 x_i$ to wartości przewidywane przez model.

4.4 Estymacja punktowa

W estymacji punktowej chcemy znaleźć konkretne wartości $\hat{\beta}_0$ i $\hat{\beta}_1$, które najlepiej pasują do naszych danych:

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

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

gdzie $\bar{x}$ i $\bar{y}$ to średnie wartości odpowiednio zmiennej niezależnej i zależnej.

4.5 Estymacja przedziałowa

W kontekście klasycznego modelu regresji liniowej, estymacja przedziałowa jest wykorzystywana do określenia przedziałów ufności dla estymatorów $\hat{\beta}_0$ i $\hat{\beta}_1$. Dodatkowo, zakładając normalność składników losowych ϵ_i z średnią 0 i wariancją σ^2, możemy wykorzystać te założenia do konstrukcji przedziałów ufności.

4.5.1 Rozkład estymatorów

Załóżmy, że $\epsilon_i \sim N(0, \sigma^2)$. Wówczas estymatory $\hat{\beta}_0$ i $\hat{\beta}_1$, obliczone metodą najmniejszych kwadratów, również mają rozkłady normalne:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$

4.5.2 Estymator wariancji składnika losowego

Nie znając prawdziwej wartości σ^2, estymujemy ją przy pomocy:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$$

gdzie s^2 jest estymatorem nieobciążonym wariancji σ^2, a $\hat{Y}_i$ to wartości przewidywane przez model.

4.5.3 Statystyki do wyznaczenia przedziałów ufności

Dla konstrukcji przedziałów ufności dla $\hat{\beta}_1$ i $\hat{\beta}_0$, korzystamy ze statystyk:

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t(n-2)$$

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t(n-2)$$

gdzie $SE(\hat{\beta}_1)$ i $SE(\hat{\beta}_0)$ to standardowe błędy estymatorów, a $t(n-2)$ to rozkład t-Studenta z $n-2$ stopniami swobody.

4.5.4 Przedziały ufności

Przedziały ufności dla β_1 i β_0 na poziomie ufności $1-\alpha$ są dane jako:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1)$$

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_0)$$

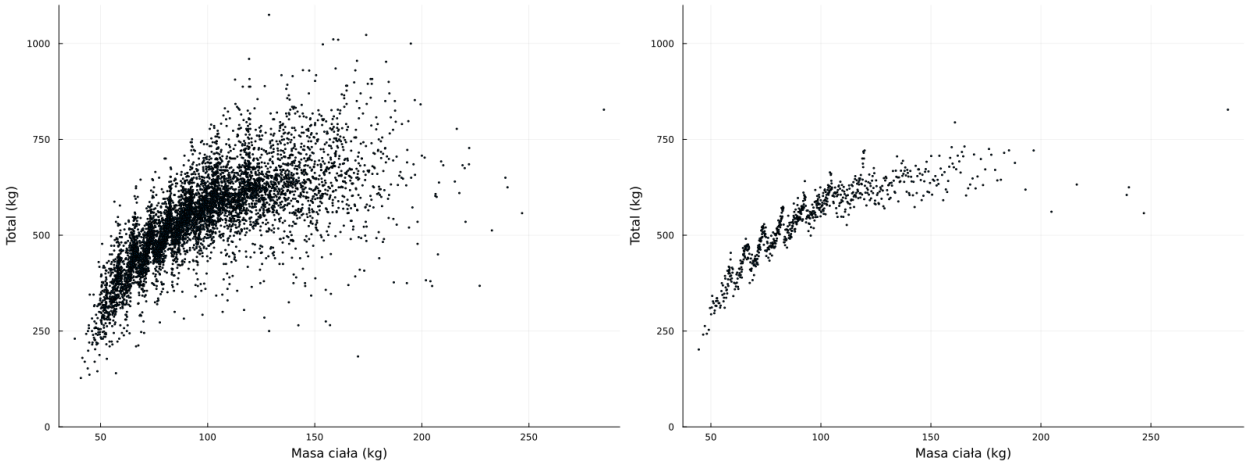
gdzie $t_{\alpha/2,n-2}$ to wartość krytyczna z rozkładu t-Studenta odpowiadająca poziomowi ufności $1-\alpha$.

Przy pomocy tych przedziałów możemy stwierdzić, z określonym poziomem ufności, gdzie spodziewamy się znaleźć rzeczywiste wartości β_1 i β_0. Te przedziały dają nam lepszy wgląd w niepewność związaną z naszymi estymacjami i są kluczowe w statystycznej interpretacji wyników modelu regresji liniowej.

4.6 Zastosowanie do zestawu danych

Na rysunku 2a przestawiony został wykres rozproszenia danych przed wygładzeniu średnią ruchomą. Na rysunku 2b zaś po. Jak widać zależność pomiędzy badanymi zmiennymi jest wklęsła, co świadczy o tym, że nie powinniśmy dopasowywać do niej modelu liniowego. Można zakładać logarytmiczną bądź potęgową zależność pomiędzy danymi. Dla obu hipotez wykonano odpowiednie wkyresy przedstawione na rysunku 3. Na wykresie 3a zakładamy, że pomiędzy danymi występuje zależność logarytmiczna, a więc celem *wyprostowania* danych stosujemy transformację logarytmiczną (konkretnie logarytm dziesiętny) tylko na zmiennej niezależnej. Na wykresie 3b zakładmy zaś, że między zmiennymi występuje zależność potęgowa. Tym samym logarytmujemy obie zmienne.

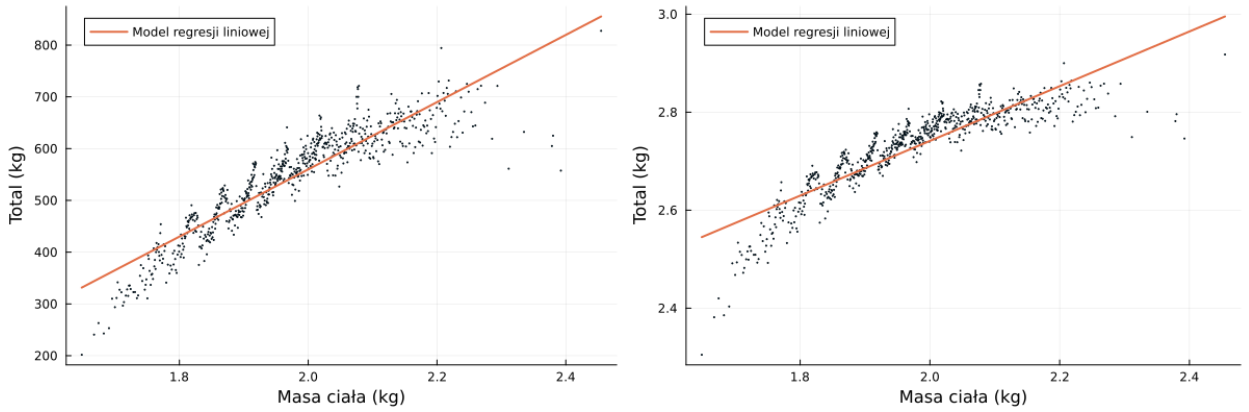
Współczynniki determinacji (opisane szczegółowo w 4.8.5) dla 3a oraz 3b wynoszą odpowiednio $0,83$ oraz $0,78$. Tym samym pozostaniemy przy transformacji logarytmicznej tylko zmiennej niezależnej.



(a) Przed wygładzeniem MA

(b) Po wygładzeniu MA

Rysunek 2: Wykresy rozproszenia danych. Wraz ze wzrostem wagi obserwujemy coraz większe rozproszenie oraz wypłaszczenie się wykresu.



(a) Transformacja logaryt. zmiennej niezależnej.

(b) Transformacja logaryt. obu zmiennych.

Rysunek 3: Wykresy rozproszeń wraz z dopasowanymi modelami regresji liniowej metodą najmniejszych kwadratów.

4.7 Usunięcie wartości odstających

Aby jeszcze lepiej dopasować model do danych dokonamy usunięcia wartości odstających. W tym celu dokonamy **standaryzacji Z**. Polega ona na obliczeniu statystyki

$$z = \frac{x - \mu}{\sigma}$$

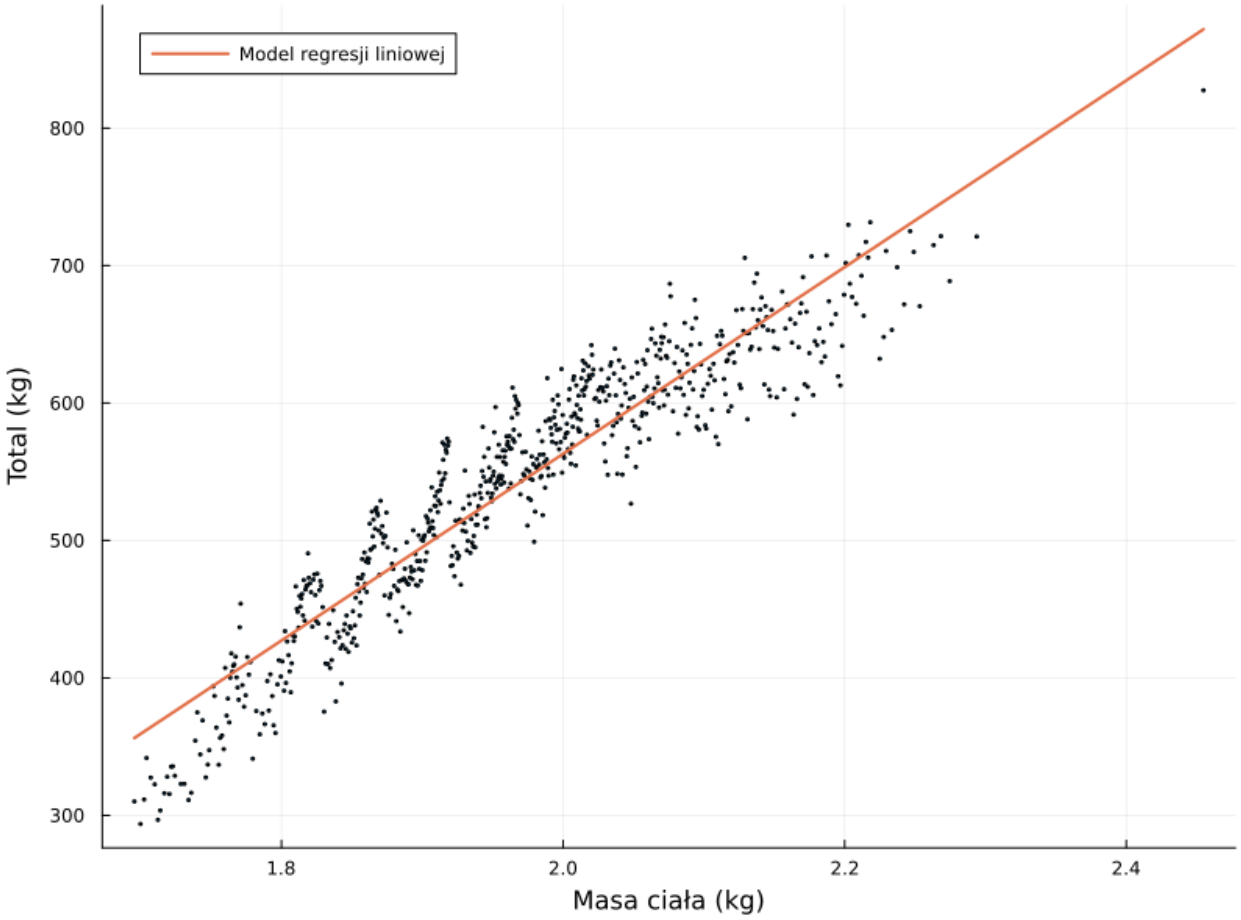
, gdzie μ jest średnią z populacji, a σ odchyleniem standardowym. Standaryzacje tą wykonamy dla residuów. Zgodnie z wynikami z sekcji 5 residua mają rozkład normalny, więc statystyka z ma standardowy rozkład normalny. Odrzucimy wszelkie obserwacje, dla których $|z| > 3$ kierując się zasadą 3 sigm. Dopasowanie modelu po odrzuceniu obserwacji odstających zostało przedstawione na wykresie 4.

4.8 Ocena poziomu zależności

4.8.1 Współczynnik korelacji Pearsona

Współczynnik korelacji Pearsona (r) jest miarą siły i kierunku związku liniowego między dwoma zmiennymi. Wartość r mieści się w przedziale od -1 do 1, gdzie 1 oznacza idealną korelację dodatnią, -1 idealną korelację ujemną, a 0 brak liniowej zależności. Wzór na r wygląda następująco:

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



Rysunek 4: Wykres rozproszenia danych po transformacji i usunięciu wartości odstających.

gdzie: - x_i i y_i to indywidualne wartości dwóch zmiennych. - $\bar{x}$ i $\bar{y}$ to średnie wartości odpowiednich zmiennych.

4.8.2 Suma kwadratów całkowita (SST)

SST (Total Sum of Squares) mierzy całkowitą zmienność w danych związanych z zmienną zależną Y. Jest to suma kwadratów różnic między obserwowanymi wartościami zmiennej zależnej a ich średnią wartością. Można to zapisać jako:

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

SST jest używana jako punkt odniesienia do oceny, jak dużo zmienności w zmiennej zależnej jest wyjaśnione przez model.

4.8.3 Suma kwadratów błędów (SSE)

SSE (Sum of Squares due to Error) mierzy ilość zmienności w danych, która nie jest wyjaśniona przez model. Jest to suma kwadratów różnic między obserwowanymi wartościami zmiennej zależnej a wartościami przewidywanymi przez model:

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

gdzie $\hat{y}_i$ to wartości przewidywane przez model.

4.8.4 Suma kwadratów regresji (SSR)

SSR (Sum of Squares due to Regression) mierzy ilość zmienności w zmiennej zależnej, która jest wyjaśniona przez model. Matematycznie jest to różnica między SST a SSE:

$$SSR = SST - SSE$$

lub równoważnie, jest to suma kwadratów różnic między wartościami przewidywanymi przez model a średnią wartością zmiennej zależnej:

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

Te trzy statystyki - SST, SSE i SSR - są kluczowe do oceny dopasowania modelu regresji, umożliwiając zrozumienie, ile zmienności w danych jest wyjaśnione przez model, a ile pozostaje niewyjaśnione.

4.8.5 Współczynnik determinacji (R^2)

Aby ująć trzy powyższe statystyki w jedną stosuje się współczynnik R^2 dany wzorem

$$R^2 = \frac{SSR}{SSE} = 1 - \frac{SSE}{SST}$$

.

Współczynnik determinacji przyjmuje wartości z zakresu $[0, 1]$ oraz mówi jak dobrze dobrany model opisuje zależność między danymi. Wartości bliskie 0 świadczą o tym, że rozrzut residuów jest taki sam co danych, a tym samym model nic nie wnosi. W wyidealizowanym przypadku, gdy $R^2 = 1$, model przedstawia dokładną zależność między danymi.

4.9 Miary dopasowania dla zestawu danych

Dla modelu z rysunku 4 otrzymujemy następujące miary dopasowania.

Współczynnik korelacji Pearsona R_{XY}	**SSE**	**SST**	**Współczynnik determinacji** R^2
$0,94$	676892	5871410	0.88

Tabela 2: Miary dopasowania modelu

Jak widać dzięki współczynnikowi R_{XY} po wszystkich transformacjach uzyskujemy ewidentną zależność liniową między danymi, zaś współczynnik R^2 mówi nam, że udało się dopasować do nich *dobry* model.

4.10 Przedziały ufności

Dopasowanie modelu do danych metodą najmniejszych kwadratów zwraca estymatory

$$\hat{\beta}_0 = -795,27 \quad \hat{\beta}_1 = 679,15$$

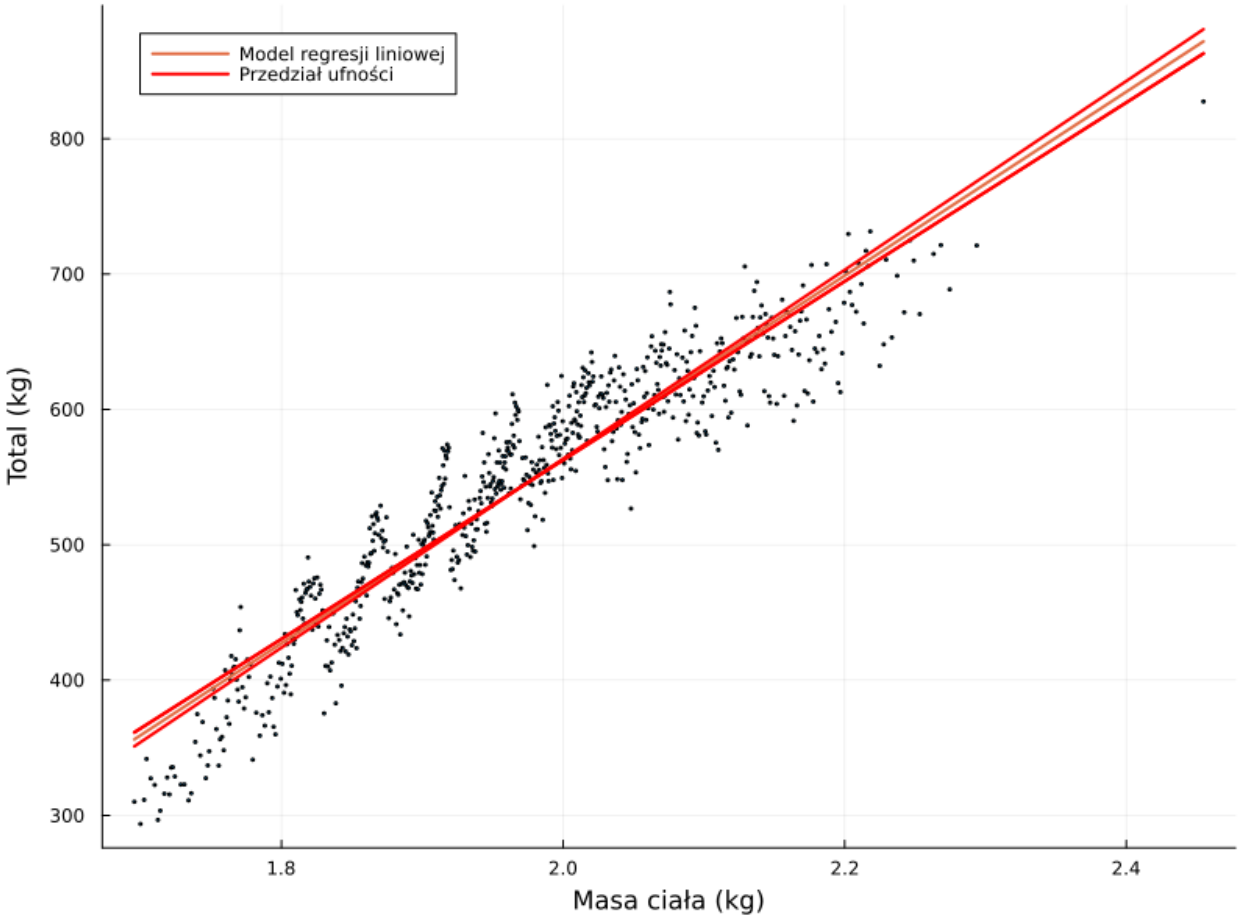
. Korzystając ze wzorów wyznaczonych w 4.5.4 otrzymujemy przedziały ufności

$$\hat{\beta}_0 \in [-831,91; -758,63] \quad \hat{\beta}_1 \in [660,58; 697,72]$$

. Model regresji liniowej wraz z przedziałem ufności został przedstawiony na wykresie 5.

5 Analiza residuów

Analiza residuów jest kluczowym elementem oceny dopasowania modelu regresji liniowej. Residua, czyli różnice między obserwowanymi wartościami zmiennej zależnej a wartościami przewidywanymi przez model, dostarczają informacji o adekwatności modelu oraz o założeniach leżących u jego podstaw. Sprawdzenie następujących założeń jest istotne:



Rysunek 5: Wykres rozproszenia danych wraz z modelem regresji liniowej i przedziałem ufności.

5.1 Średnia równa zero

Założenie to mówi, że średnia residuów powinna być równa zero ($E(\epsilon_i) = 0$). Oznacza to, że model nie systematycznie nie przewiduje ani nie niedoszacowuje wartości zmiennej zależnej. Z metody najmniejszych kwadratów wynika, że suma residuów jest równa zero, co prowadzi do wniosku, że ich średnia też jest równa zero, pod warunkiem, że model zawiera wyraz wolny.

5.2 Stała wariancja (Homoskedastyczność)

Stała wariancja residuów, znana jako homoskedastyczność, oznacza, że wariancja błędów jest stała dla wszystkich poziomów wartości zmiennej niezależnej. W praktyce oznacza to, że rozrzut residuów wokół linii regresji jest mniej więcej jednakowy niezależnie od wartości zmiennej niezależnej. Niestety, metoda najmniejszych kwadratów nie gwarantuje homoskedastyczności.

5.3 Niezależność residuów

Niezależność residuów od siebie jest kluczowa dla wiarygodności modelu regresji. W szczególności oznacza to, że wartości błędu dla jednej obserwacji nie są zależne od wartości błędu dla innej obserwacji. Niezależność residuów można sprawdzić poprzez analizę funkcji autokorelacji.

5.4 Rozkład normalny

Ostatnie założenie dotyczy rozkładu residuów, które powinny mieć rozkład normalny, szczególnie w małych próbach. Dzięki temu można stosować różnego rodzaju testy statystyczne, które zakładają normalność. Rozkład normalny residuów można zweryfikować za pomocą różnych testów statystycznych czy analizy wykresów kwantylowych (Q-Q plot).

5.4.1 Teoretyczne uzasadnienie zerowej średniej residuów w estymacji metodą najmniejszych kwadratów

W kontekście prostego modelu regresji liniowej, nasz model przedstawia się następująco:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Tutaj Y_i to obserwowane wartości, X_i to wartości zmiennej niezależnej, β_0 i β_1 to parametry, a ϵ_i reprezentuje reszty, czyli różnice pomiędzy wartościami obserwowanymi a wartościami przewidywanymi przez nasz model:

$$\epsilon_i = Y_i - \hat{Y}_i$$
$$\epsilon_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

Metoda najmniejszych kwadratów (MNK) ma na celu znalezienie estymatora parametrów ($\hat{\beta}_0$ i $\hat{\beta}_1$), które minimalizują sumę kwadratów residuów:

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Aby znaleźć minimum S, obliczamy pochodne cząstkowe względem $\hat{\beta}_0$ i $\hat{\beta}_1$ i przyrównujemy je do zera.

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

Z pierwszego równania:

$$\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \bar{X}$$

Wstawiając $\hat{\beta}_0$ z powrotem do definicji residuów:

$$\epsilon_i = Y_i - \left(\frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i \right)$$

Sumując wszystkie reszty po wszystkich i:

$$\sum_{i=1}^{n} \epsilon_i = \sum_{i=1}^{n} Y_i - n \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) + \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})$$

Zauważając, że $\sum_{i=1}^{n} (X_i - \bar{X}) = 0$, termin z $\hat{\beta}_1$ znika:

$$\sum_{i=1}^{n} \epsilon_i = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Y_i = 0$$

Co oznacza, że suma residuów jest równa zero, co implikuje, że ich średnia również wynosi zero:

$$\frac{1}{n} \sum_{i=1}^{n} \epsilon_i = 0$$

To wyprowadzenie pokazuje, że z natury metody najmniejszych kwadratów wynika, iż średnia residuów jest koniecznie równa zero, co umacnia założenie, że nasz model nie przewiduje systematycznie ani nie niedoszacowuje zmiennej zależnej.

5.5 Test Andersona-Darlinga

Test Andersona-Darlinga jest jednym z testów statystycznych używanych do oceny, czy próbka danych pochodzi z populacji o określonym rozkładzie, najczęściej normalnym. Test ten jest modyfikacją testu Kołmogorowa-Smirnowa i jest bardziej czuły na obserwacje znajdujące się w ogonach rozkładu.

5.5.1 Założenia i formuła

Założenie podstawowe testu to hipoteza zerowa, która mówi, że dane pochodzą z populacji o rozkładzie normalnym. Test porównuje empiryczną funkcję dystrybuanty (ECDF) z teoretyczną funkcją dystrybuanty rozkładu normalnego. Jego statystyka testowa wyraża się wzorem:

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$

gdzie $F_n(x)$ to empiryczna funkcja dystrybuanty z próbki, $F(x)$ to teoretyczna funkcja dystrybuanty rozkładu normalnego, a n to rozmiar próbki.

5.5.2 Interpretacja

Wartość statystyki testowej A^2 jest następnie porównywana z wartościami krytycznymi dla danego poziomu istotności. Jeśli statystyka testowa przekracza wartość krytyczną, hipoteza zerowa jest odrzucana, co sugeruje, że dane nie pochodzą z rozkładu normalnego. Test Andersona-Darlinga jest szczególnie wrażliwy na odchylenia od normalności w ogonach rozkładu, co jest jedną z jego głównych zalet.

5.6 Test Jarque-Bera

Test Jarque-Bera to test statystyczny używany do sprawdzania, czy dane mają skośność i kurtozę odpowiadające rozkładowi normalnemu. Test ten jest popularny w ekonometrii i innych dziedzinach, które zakładają normalność residuów w modelach regresji.

5.6.1 Założenia i formuła

Test Jarque-Bera opiera się na skośności (ang. skewness) i kurtozie (ang. kurtosis) danych. Skośność jest miarą asymetrii rozkładu, a kurtoza mierzy „ostrość" rozkładu lub grubość jego ogonów. Hipoteza zerowa testu Jarque-Bera mówi, że dane mają skośność równą 0 i kurtozę równą 3 (charakterystyczną dla rozkładu normalnego). Statystyka testowa JB wyraża się wzorem:

$$JB = \frac{n}{6}(S^2 + \frac{(K-3)^2}{4})$$

gdzie n to liczba obserwacji, S to skośność próbki, a K to kurtoza próbki.

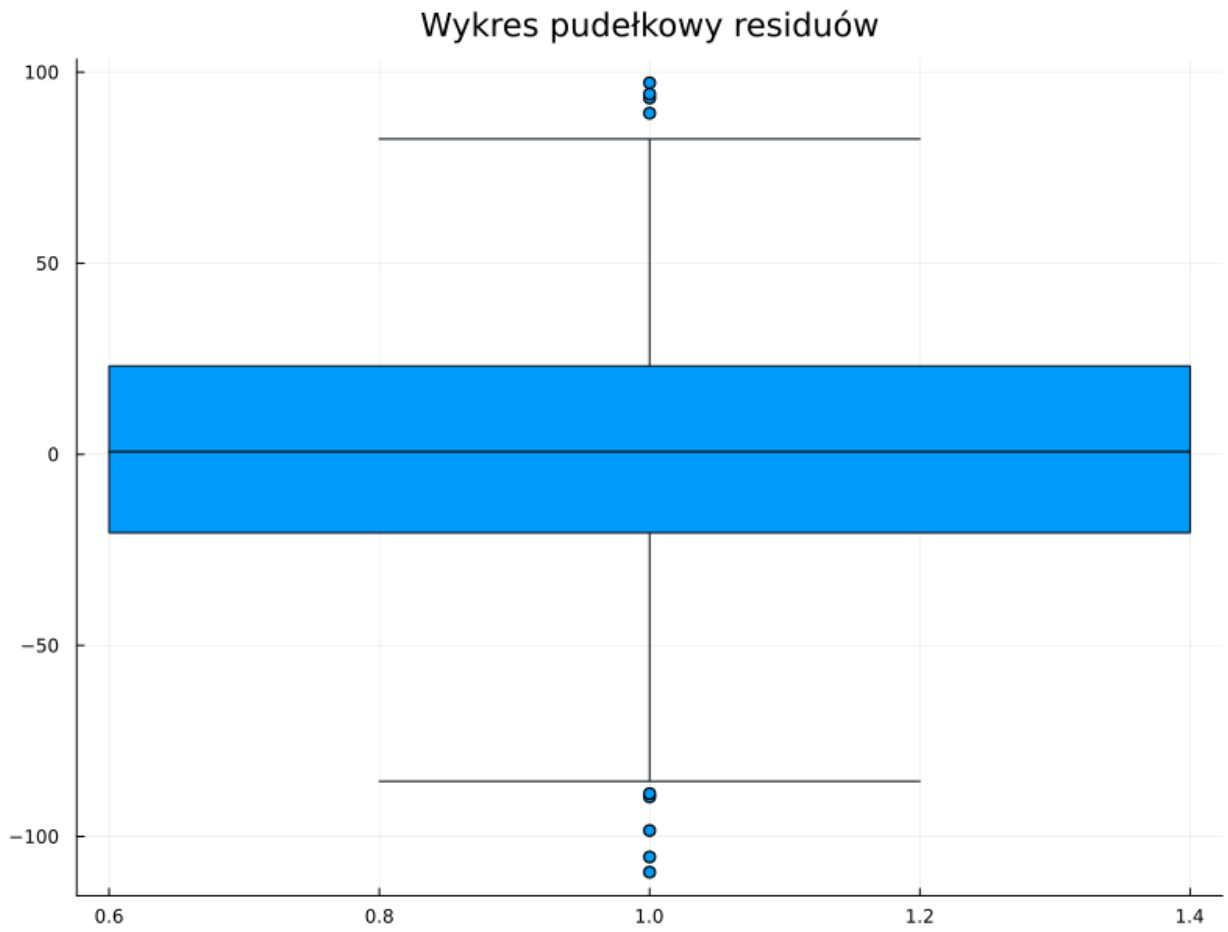
5.6.2 Interpretacja

Wartość statystyki JB jest następnie porównywana z wartościami krytycznymi chi-kwadrat z dwoma stopniami swobody. Jeżeli wartość statystyki JB jest znacząco wysoka, hipoteza zerowa o normalności rozkładu jest odrzucana. Test Jarque-Bera jest szeroko stosowany do badania normalności residuów w modelach regresji liniowej i jest szczególnie użyteczny w dużych próbach, gdzie jego moc jest wysoka.

Oba testy, Andersona-Darlinga i Jarque-Bera, są istotnymi narzędziami w statystycznej analizie danych, pozwalającymi ocenić zgodność danych z rozkładem normalnym, co jest kluczowym założeniem w wielu modelach statystycznych i ekonometrycznych.

5.7 Analiza residuów dla zestawu danych

Rozkład residuów dla rozpatrywanych danych został przedstawiony na wykresie 6, zaś statystyki opisujące kluczowe cechy tego rozkładu w tabeli 3. Jak widać mamy doczynienia z zerową średnią oraz bliską zera skośnością oraz kurtozą. Tym samym można wysunąć hipotezę o normalności rozkładu. W celu przetestowania tej hipotezy przeciwko hipotezie alternatywnej (residua nie pochodzą z rozkładu normalnego) wykonamy test Jarque-Bera oraz Andersona-Darlinga. Wyniki testów przedstawione są na wykresie 7. Oba testy nie odrzuciły hipotezy alternatywnej, więc mamy przesłanki aby sądzić, że residua mają rozkład normalny.



Rysunek 6: Rozkład residuów.

Średnia	Odchylenie standardowe	Skośność	Kurtoza
0.00	32.97	−0.18	0.26

Tabela 3: Statystyki rozkładu residuów

```
Jarque-Bera normality test
--------------------------
Population details:
    parameter of interest:   skewness and kurtosis
    value under h_0:         "0 and 3"
    point estimate:          "-0.1617457596836042 and 3.2683540951304106"

Test summary:
    outcome with 95% confidence: fail to reject h_0
    one-sided p-value:           0.0931

Details:
    number of observations:         645
    JB statistic:                   4.74776
```

(a) Jarque-Bara

```
One sample Anderson-Darling test
--------------------------------
Population details:
    parameter of interest:   not implemented yet
    value under h_0:         NaN
    point estimate:          NaN

Test summary:
    outcome with 95% confidence: fail to reject h_0
    one-sided p-value:           0.7994

Details:
    number of observations:   645
    sample mean:              -9.760323471061221e-14
    sample SD:                32.613224119213776
    A² statistic:             0.44869202312316925
```

(b) Anderson-Darling

Rysunek 7: Wyniki testów statystycznych Jarque-Bara i Andersona-Darlinga wykonanych przy pomocy pakietu *HypothesisTests.jl* języka *Julia*

6 Podsumowanie i wnioski

W wyniku analizy udało nam się odpasować zależność logarytmiczną między danymi daną wzorem

$$\hat{y} = -795,27 + 679,15 \log_{10}(x)$$

oraz ustalić przedziały ufności dla odpowiednich współczynników. Dodatkowo nie znaleźliśmy przesłanek ku odrzuceniu hipotezy o normalności residuów co jest kolejną oznaką słuszności dopasowania. Powyższe wyniki prowadzą do konkluzji, że istnieje dodatnia korelacja pomiędzy wynikami w trójboju siłowym a masą zawodnika.