Politechnika Wrocławska

Wydział Matematyki
Kierunek studiów: Matematyka Stosowana
Specjalność: –

Praca dyplomowa – inżynierska

ANALIZA KORELACJI, REGRESJA LINIOWA I LOGISTYCZNA W BADANIACH SPORTOWYCH

Martyna Maria Boniatowska

słowa kluczowe:
Analiza, dane, zmienna, próba, model, dopasowanie, regresja, liniowa, logistyczna, korelacja, zależność, współczynnik, Pearson, Spearman, Kendall.

krótkie streszczenie:
W pracy przeanalizowano istnienie zależności między parametrami fizycznymi zawodników i liczbą punktów zdobywanych przez nich w meczach siatkówki na Igrzyskach Olimpijskich 2020. Następnie dopasowano model regresji liniowej dla liczby punktów zdobytych przez zawodników w czwartym i piątym spotkaniu oraz sprawdzono, czy dodanie kolejnych zmiennych objaśniających, wpływa na lepsze dopasowanie modelu. Zastosowano też model regresji logistycznej, aby sprawdzić, czy to jak punktuje zawodnik, ma wpływ na jego szansę gry w kolejnym spotkaniu.

Opiekun pracy dyplomowej	dr hab. inż. Maciej Wilczyński	.....	.....
	Tytuł/stopień naukowy/imię i nazwisko	ocena	podpis

*Do celów archiwalnych pracę dyplomową zakwalifikowano do:**

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

** niepotrzebne skreślić*

pieczątka wydziałowa

Wrocław, rok 2022

Wrocław University
of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Applied Mathematics
Specialty: –

Bachelor Thesis

CORRELATION ANALYSIS, LINEAR AND LOGISTIC REGRESSION IN SPORTS RESEARCH.

Martyna Maria Boniatowska

keywords:
Analysis, data, variable, sample, model, fit, regression, linear, logistic, correlation, relationship, coefficient, Pearson, Spearman, Kendall.

short summary:
The study analyzed the existence of a relationship between the physical parameters of players and the number of points scored by them in volleyball matches at the 2020 Olympics. Then the linear regression model was adjusted for the number of points scored by players in the fourth and fifth games, and it was checked whether adding further explanatory variables resulted in a better fit model. Also a logistic regression model was used to see if how a player scores has an impact on his chance to play in the next game.

Supervisor	dr hab. inż. Maciej Wilczyński	.....	.....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław, 2022

Spis treści

Oznaczenia **3**

Wstęp **5**

1 Analiza korelacji **7**
1.1 Korelacja Pearsona 7
1.2 Korelacja Spearmana 9
1.3 Korelacja Kendalla 10
1.4 Podsumowanie analizy korelacji 11

2 Regresja liniowa **13**
2.1 Model regresji liniowej z jedną zmienną objaśniającą 14
2.1.1 Predykcja i przedziały ufności 15
2.2 Model liniowy regresji wielokrotnej 16
2.2.1 Predykcja i przedziały ufności 17
2.3 Wartości resztowe 17
2.4 Ocena zależności liniowej 17
2.5 Diagnostyka modelu 18
2.6 Testy statystyczne wykorzystane do sprawdzenia założeń odnośnie residuów 19

3 Regresja logistyczna **21**
3.1 Predykcja i przedziały ufności 22
3.2 Diagnostyka modelu 22

4 Wyniki własne **25**
4.1 Przedstawienie danych poddanych analizie 25
4.2 Analiza korelacji 32
4.2.1 Korelacja Pearsona 32
4.2.2 Korelacja Spearmana 34
4.2.3 Korelacja Kendalla 36
4.3 Model regresji liniowej dla dwóch zmiennych 37
4.3.1 Dopasowanie modelu 37
4.3.2 Diagnostyka modelu — analiza residuum 41
4.4 Model regresji wielokrotnej 44
4.4.1 Dopasowanie modelu 44
4.4.2 Diagnostyka modelu – analiza residuum 47
4.5 Regresja logistyczna 50
4.6 Podsumowanie otrzymanych wyników własnych 53

Bibliografia **54**

Oznaczenia

1. n – długość próby.
2. X, Y – zmienne losowe.
3. $\overline{x}, \overline{y}$ – średnie arytmetyczne z próby.
4. S_x, S_y – odchylenia standardowe z próby.
5. r_{XY} – współczynnik korelacji Pearsona.
6. ρ – rangowy współczynnik korelacji Spearmana.
7. τ – rangowy współczynnik korelacji Kendalla.
8. x_i, y_i – rzeczywiste wartości pochodzące z próby.
9. $\hat{y}_i, \hat{Y}_i$ – wartości estymowane wzorem.
10. b_0, β_0 – wyrazy wolne w modelu.
11. b_1, β_1 – współczynniki kierunkowe w modelu.
12. ϵ – niezależne zmienne losowe (będące błędami losowymi).
13. $t_{1-\frac{\alpha}{2},n-2}$ – kwantyl z rozkładu t-studenta z $n-2$ stopniami swobody.
14. $t_{1-\frac{\alpha}{2},n-p}$ – kwantyl z rozkładu t-studenta z $n-p$ stopniami swobody.
15. e_i – wartości resztowe (residua).
16. $\mathbf{Y}'$ – wektor zmiennych zależnych.
17. $\mathbf{X}$ – macierz wartości zmiennych objaśnianych.
18. β' – wektor współczynników modelu.
19. $\sum_{\mathbf{Y}}$ – macierz kowariancji.
20. $\mathbf{b}$ – wektor estymowanych współczynników.
21. $\pi(x)$ – prawdopodobieństwo sukcesu.
22. $logit$ – funkcja logitowa.
23. R^2 – współczynnik determinacji.

Wstęp

Analiza danych sportowych jest praktykowana od wielu lat. Jednak dzięki ostatnim postępom w tej dziedzinie zakres i możliwości analizy znacząco się powiększyły. W obecnych czasach większość profesjonalnych drużyn sportowych z różnych dyscyplin sportu zbiera dane i wykorzystuje je do analizy w celu uzyskiwania coraz lepszych rezultatów. Wyniki analiz mogą być przydatne trenerom, w celu ustalenia strategii na mecz z przeciwnikiem, ponieważ z ich pomocą można odkryć mocne i słabe strony swojej, a także przeciwnej drużyny. Mogą być pomocne także podczas doboru składu drużyny, aby sprawdzić aktualną dyspozycję zawodników. Warto zaznaczyć także, że większość klubów sportowych funkcjonuje również jako firmy. W tym kontekście analiza danych może pomóc poprawieniu zysków i zmniejszeniu wydatków całej organizacji. Umiejętność modelowania i analizy danych może być również przydatna podczas obstawiania wyników wydarzeń sportowych w zakładach bukmacherskich. Intuicja, czyli znajomość sportu i drużyn czasem może być niewystarczająca i dobrze jest się oprzeć na pewnym modelu matematycznym.

Obecnie pojawia się coraz więcej badań pokazujących zastosowanie modeli matematycznych w różnych dziedzinach sportowych. Ta praca będzie poświęcona analizie korelacji i próbie dopasowania modeli regresji liniowej i logistycznej do siatkarskich danych. Zdecydowałam się na wykorzystanie obserwacji dotyczących tego sportu, ponieważ od dziecka jestem wielką fanką siatkówki. Dlatego podczas wyboru tematu zależało mi na tym, żeby praca koniecznie dotyczyła tej dziedziny sportu. Jeśli chodzi o siatkówkę, modelowanie danych nie jest jeszcze tak popularne. Nie ma jeszcze zbyt wielu artykułów naukowych, które pokazują jak je modelować, lub też które modele do ich analizy są najbardziej odpowiednie. Podczas studiów poznałam kilka modeli regresyjnych i w swojej pracy za cel postawiłam sobie sprawdzenie możliwości ich zastosowania do przewidywania wyników rozgrywek siatkarskich. Chciałam zbadać zależności między danymi, które udało mi się zebrać. Sprawdzić, czy w przypadku tego sportu, gdy mamy do czynienia z profesjonalnymi sportowcami, ich parametry fizyczne mogą mieć wpływ na to jak punktują zawodnicy oraz, czy wyniki z poprzednich spotkań mają wpływ na to jak będą punktować gracze w kolejnych. Kolejnym celem pracy było sprawdzenie, czy liczba zdobywanych punktów przez zawodnika może mieć wpływ na to, czy trener zdecyduje się wystawić go w kolejnym spotkaniu. W pracy przedstawiona zostanie zarówno wspomniana analiza jak i wszystkie potrzebne zagadnienia teoretyczne do jej przeprowadzenia.

Rozdział 1

Analiza korelacji

Analiza korelacji to metoda wnioskowania statystycznego, za pomocą której sprawdza się, czy dwie zmienne są ze sobą istotnie statystycznie powiązane. Innymi słowy, jak bardzo jedna ze zmiennych wpływa na drugą. Wykorzystuje się ją, badając różne zjawiska np. ekonomiczne, biologiczne czy tak jak w tej pracy sportowe. Pomaga ona w znalezieniu zależności i związku, jednak należy pamiętać, że nie jest to zależność przyczynowo-skutkowa. Czyli nie zawsze wartości jednej cechy wpływają na drugą, a jedynie zachowują one podobny trend. Analiza korelacji bywa więc przedmiotem wielu badań i podstawą do stosowania modeli wykorzystywanych do predykcji. Istnieje wiele sposobów badania korelacji, a do najpopularniejszych należą: korelacja liniowa Pearsona oraz rangowe współczynniki badające ogólną zależność monotoniczną takie jak Rho Spearmana (ρ) czy Tau Kendalla (τ).

Na początku należy zaznaczyć, że próba, dla której wyznaczany jest współczynnik korelacji we wszystkich omówionych poniżej przypadkach ma postać: $(x_1, y_1), \ldots, (x_n, y_n)$ i zawiera wartości przyjęte przez niezależne wektory losowe $(X_1, Y_1), \ldots, (X_n, Y_n)$, z których każdy ma ten sam rozkład co wektor (X, Y).

1.1 Korelacja Pearsona

Definicja 1.1. [14] Współczynnik korelacji Pearsona jest to miara zależności liniowej między dwiema zmiennymi losowymi opisana wzorem:

$$r_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{1.1}$$

$Cov(X, Y)$ we wzorze (1.1) to kowariancja między X i Y, która odzwierciedla stopień, w jakim obserwacje dwóch zmiennych różnią się od ich odpowiednich średnich w tym samym stopniu i kierunku. Dzielenie przez σ_X i σ_Y, czyli odchylenia standardowe sprawia, że współczynnik nie jest wrażliwy na zmianę skali.

Własności współczynnika r_{XY} dla dowolnych zmiennych losowych X i Y:

1. $|r_{XY}| \leq 1$.
2. Jeśli zmienne X i Y są niezależne, to $r_{XY} = 0$.
3. Jeśli $|r_{XY}| = 1$ to będą istniały liczby a, b takie, że $P(Y = aX + b) = 1$.
4. Wartość $r_{XY} = 0$ nie oznacza niezależności zmiennych losowych X i Y.

Estymatorem współczynnika r_{XY} jest **próbkowy współczynnik korelacji Pearsona** [13]:

$$\hat{r}_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}$$

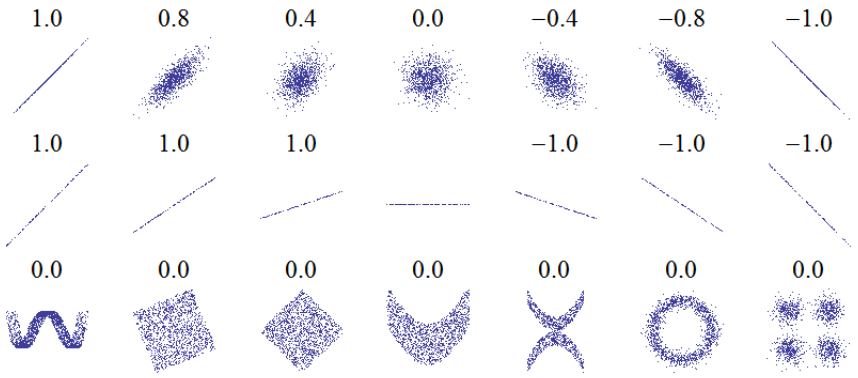
gdzie:

- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

Analogicznie wyznacza się estymatory średniej i wariancji dla zmiennej losowej Y.

Własności próbkowego współczynnika korelacji Pearsona:

1. Przyjmuje wartości z przedziału $[-1, 1]$ i nie zmienia się w zależności od jednostek miary użytych dla żadnej zmiennej.
2. Wartości r_{XY} bliskie zeru oznaczają brak powiązania liniowego.
3. Wartości bliskie $r_{XY} = 1$ oznaczają silną dodatnią zależność liniową.
4. Wartości bliskie $r_{XY} = -1$ oznaczają silną ujemną zależność liniową.



Rysunek 1.1: Wykresy rozproszenia XY i odpowiadające im współczynniki korelacji [11].

Wykresy przedstawione na rysunku (1.1) obrazują zachowanie się współczynnika, w zależności od tego jak dane prezentują się na wykresie rozproszenia. Zależność liniową widoczna jest na pierwszy rzut oka.

Istnieją jednak pewne ograniczenia stosowalności tego współczynnika. Mianowicie jest on podatny na występowanie skrajnych obserwacji odstających, a także zakłada, że próby, pomiędzy którymi badana jest korelacja, pochodzą z rozkładu normalnego lub liczność próby n jest bardzo duża. Tylko przy spełnieniu tego założenia interpretacja wyniku jest oczywista. Dlatego należy zastosować testy normalności próby przed zinterpretowaniem wartości tego współczynnika.

1.2 Korelacja Spearmana

Współczynnik korelacji Spearmana jest ogólniejszym i nieparametrycznym odpowiednikiem współczynnika korelacji Pearsona. Bada ogólną monotoniczną zależność, która nie musi być liniowa. Wartość współczynnika można interpretować niezależnie od tego czy zmienne losowe mają rozkład normalny, ponieważ dane zostają poddane przekształceniu. Parę (x_i, y_i) zastępuje się parą (q_i, r_i), gdzie q_i jest rangą obserwacji x_i w próbie X i r_i jest rangą obserwacji y_i w próbie Y. Współczynnik wylicza się dla ciągu par rang [8]. Zakłada się jedynie, że rozkład pary (X, Y) jest ciągły i istnieje możliwość uporządkowania danych.

Definicja 1.2. [8] Współczynnik korelacji rangowej Spearmana między zmiennymi X, Y ma postać

$$\rho = 3\,[P(\,(X_1 - X_2)(Y_1 - Y_3\,) > 0\,) - P(\,(X_1 - X_2)(Y_1 - Y_3\,) < 0\,)]$$

gdzie $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$ są niezależnymi kopiami (X, Y).

Dla danych rzeczywistych można wyznaczyć **estymator współczynnika korelacji Spearmana.** Wygląda on następująco [8]:

$$\begin{aligned}\hat{\rho} &= \frac{\sum\limits_{i=1}^{n}\left(r_i - \frac{n+1}{2}\right)\left(q_i - \frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(r_i - \frac{n+1}{2}\right)^2}\sqrt{\sum\limits_{i=1}^{n}\left(q_i - \frac{n+1}{2}\right)^2}} \\ &= \frac{12}{n^3 - n}\sum_{i=1}^{n}\left(r_i - \frac{n+1}{2}\right)\left(q_i - \frac{n+1}{2}\right),\end{aligned}$$

Własności rangowego współczynnika korelacji Spearmana [13]:

1. Współczynnik ρ przyjmuje wartości z przedziału $[-1, 1]$. Wartość $+1$ oznacza ściśle dodatnią monotoniczność, wartość -1 ściśle ujemną monotoniczność.
2. Jeśli $\rho > 0$, to wzrost (spadek) jednej zmiennej jest związany ze wzrostem (spadkiem) drugiej zmiennej. Jeśli $\rho < 0$, to wzrost (spadek) jednej zmiennej wiąże się ze spadkiem (wzrostem) drugiej zmiennej.
3. Dla dwóch zmiennych niezależnych współczynnik korelacji wynosi $\rho = 0$.

Etapy postępowania podczas wyliczania współczynnika korelacji rang Spearmana:

1. Uporządkowanie wartości x_i i y_i rosnąco bądź malejąco.
2. Nadanie im numerów kolejnych liczb naturalnych q_i i r_i(rangowanie).
3. Dopasowanie odpowiednich rang dla par x_i i y_i.
4. Wyliczenie różnicy rang.
5. Podstawienie do wzoru na współczynnik.

Współczynnik ten jest mniej wrażliwy na obserwacje odstające od poprzedniego, ze względu na zastosowane rangowanie. Wyliczona zależność pomiędzy zmiennymi losowymi nie musi jednak świadczyć o istnieniu związku przyczynowo-skutkowego, a jedynie o występującej zależności monotonicznej, która niekoniecznie musi być liniowa.

1.3 Korelacja Kendalla

Współczynnik korelacji Kendalla jest kolejną z rangowych metod badania zależności monotonicznej, a nie tylko liniowej między dwiema zmiennymi. Aby otrzymać rangowy współczynnik korelacji Kendalla, wylicza się różnicę między prawdopodobieństwem, że porównywane zmienne będą układały się w tym samym porządku dla dwóch obserwacji, a prawdopodobieństwem, że ułożą się w przeciwnym porządku [9]. Współczynnik ten jest odporny na obserwacje odstające i z racji tego, że jest to metoda rangowa. Nie posiada żadnych założeń odnośnie rozkładu, z którego muszą pochodzić badane zmienne.

Definicja 1.3. [6] Współczynnik korelacji rangowej Kendalla między zmiennymi X, Y ma postać:

$$\tau = P(\,(X_1 - X_2)(Y_1 - Y_2\,) > 0\,) - P(\,(X_1 - X_2)(Y_1 - Y_2\,) < 0\,)$$

gdzie $(X_1, Y_1), (X_2, Y_2)$ są niezależnymi kopiami (X, Y).

Dla dwóch cech ilościowych X i Y w n-elementowej zbiorowości jej elementy łączy się w dwuelementowe podzbiory. Dla takiej zbiorowości można utworzyć $N = n \cdot (n-1)$ takich podzbiorów, czyli uporządkowanych par. Współczynnik korelacji Kendalla wyznacza się więc na podstawie dwuelementowych podzbiorów, utworzonych z elementów zbioru wyjściowego [13].

Estymator współczynnika korelacji próbkowej Kendalla τ wyraża się wzorem:

$$\hat{\tau} = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} sgn(r_i - r_j) sgn(q_i - q_j).$$

gdzie $sgn(r_i - r_j)$ to wartość funkcji signum dla pary pochodzącej ze zbioru wyjściowego X, a $sgn(q_i - q_j)$ to wartość funkcji signum dla pary pochodzącej ze zbioru wyjściowego Y. Znak jest dodatni jeśli wartość cechy dla pierwszego elementu w i-tej parze jest większa niż dla drugiego elementu. Natomiast znak jest ujemny gdy wartość cechy dla pierwszego elementu jest mniejsza niż dla drugiego elementu [13].

Funkcja sgn to funkcja signum zdefiniowana w następujący sposób [3]:

$$sgn(x) = \begin{cases} -1 & \text{gdy } x < 0 \\ 0 & \text{gdy } x = 0 \\ 1 & \text{gdy } x > 0 \end{cases}$$

Własności rangowego współczynnika korelacji Kendalla [13]:

1. Współczynnik τ przyjmuje wartości z przedziału $[-1, 1]$. Wartość $+1$ oznacza silną dodatnią monotoniczność, wartość -1 silną ujemną monotoniczność.

2. Jeśli $\tau > 0$, to wzrost (spadek) jednej zmiennej jest związany ze wzrostem (spadkiem) drugiej zmiennej. Jeśli $\tau < 0$, to wzrost (spadek) jednej zmiennej wiąże się ze spadkiem (wzrostem) drugiej zmiennej.

1.4 Podsumowanie analizy korelacji

Nie należy bezpośrednio porównywać wartości opisanych wyżej współczynników, ponieważ każdy z nich mierzy coś innego. Współczynnik Pearsona mówi tylko o zależności linowej. Współczynniki Spearmana i Kendalla mierzą zaś ogólną zależność monotoniczną. W celu sprawdzenia, czy zaobserwowane podczas estymacji współczynników zależności w próbach mogły wystąpić przypadkowo, czyli na skutek zmienności losowej, można zastosować t-test, czyli test istotności współczynnika korelacji.

Test istotności współczynnika korelacji. Hipoteza zerowa tego testu mówi, że zmienne X i Y są niezależne, czyli korelacja między badanymi cechami nie występuje. Natomiast hipoteza alternatywna mówi, że zmienne losowe X i Y są zależne. Test dla współczynnika Pearsona wykorzystuje statystykę testową [7]:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

gdzie r to wartość współczynnika korelacji, a n liczebność próby.

Przy pomocy tej statystyki można wyznaczyć zbiór krytyczny, czyli przedział, dla którego hipoteza zerowa jest odrzucana.

$$(-\infty; -t_{1-\frac{\alpha}{2},n-2}) \cup (t_{1-\frac{\alpha}{2},n-2}, +\infty)$$

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

Test przeważnie wykonuje się na poziomie istotności $\alpha = 0.05$, jest to z góry przyjęty dopuszczalny poziom ryzyka popełnienia błędu, wynoszący 5%. Jeśli statystyka znajduje się w zbiorze krytycznym, hipoteza zerowa jest odrzucana na rzecz hipotezy alternatywnej. Czyli korelacja jest istotna statystycznie. Jeśli statystyka nie należy do wyznaczonego zbioru, wtedy nie odrzuca się hipotezy zerowej i korelacja jest nieistotna statystycznie [7].

Rozdział 2

Regresja liniowa

Regresja jest metodą statystyczną pozwalającą określić współzmienność kilku zmiennych poprzez dopasowanie do nich funkcji regresji. Taki zabieg pozwala na przewidywanie wartości jednych zmiennych na podstawie drugich. Regresja liniowa polega więc na dopasowaniu funkcji liniowej.

Celem regresji jest ustalenie wpływu p zmiennych objaśniających $X_1, \ldots, X_p$ na zmienną objaśnianą Y. Przykładowo:

1. wpływ wzrostu matki i wzrostu ojca (zmienne objaśniające X_1 i X_2) na wzrost ich dziecka (zmienna objaśniana Y).

2. wpływ wzrostu i wagi skoczka (zmienna objaśniająca X_1 i X_2) na długość oddanego przez niego skoku (zmienna objaśniana Y).

Model statystyczny, opisujący tę zależność, tworzy się na podstawie danych, znając wyniki n pomiarów zmiennej objaśnianej i odpowiadających im n pomiarów zmiennych objaśniających:

$$\begin{array}{ccccc} y_1 & x_{11} & x_{12} & \ldots & x_{1p} \\ y_2 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n1} & x_{n2} & \ldots & x_{np} \end{array}$$

W modelu regresji liniowej przyjmuje się, że zależność między $X_1, \ldots, X_{p-1}$ a Y ma postać [15]:

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_{p-1} X_{p-1} + \epsilon.$$

gdzie:

- $\beta_0, \ldots, \beta_{p-1}$ – współczynniki modelu, będące nieznanymi stałymi.

- ϵ – błąd losowy, odzwierciedlający wpływ czynników losowych na zmienną Y, które są nieznane. Zakłada się, że $E(\epsilon_i) = 0$ oraz $Var(\epsilon_i) = \sigma^2$, gdzie σ^2 jest pewną liczbą. Zazwyczaj przyjmuje się, że rozkład ϵ jest normalny [17].

Współczynniki modelu $\beta_0, \ldots, \beta_{p-1}$ i wariancję σ^2 można estymować na podstawie próby.

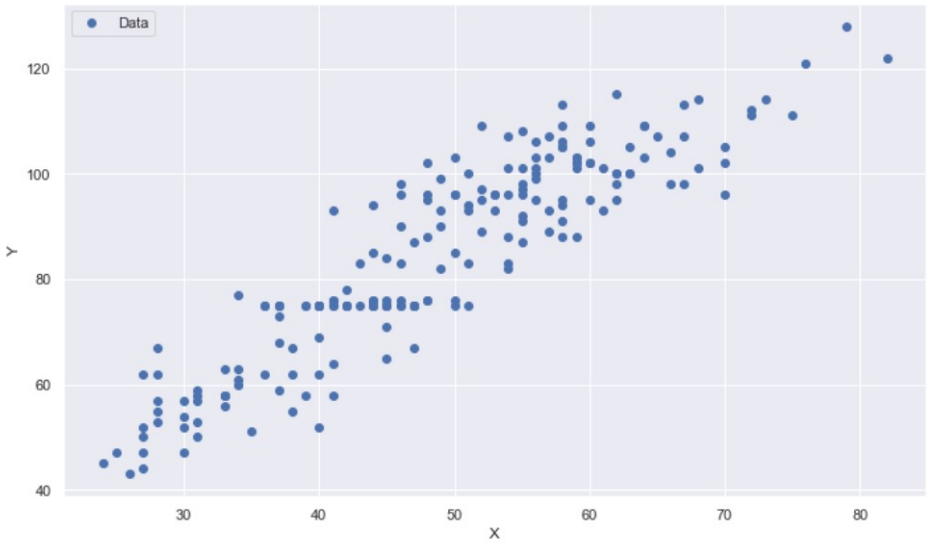
2.1 Model regresji liniowej z jedną zmienną objaśniającą

Na początku rozważany będzie przypadek uwzględniający tylko jedną zmienną objaśniającą. W tym przypadku zakłada się, że zależność między zmienną objaśnianą Y a zmienną objaśniającą X ma postać [14]:

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Nieznane współczynniki modelu regresji liniowej β_0 i β_1 estymuje się korzystając z próby $(x_1, y_1), \ldots, (x_n, y_n)$.

Konstruowanie modelu zaczyna się od wykonania wykresu rozproszenia, czyli umieszczenia na płaszczyźnie punktów $(x_1, y_1), \ldots, (x_n, y_n)$. Wykorzystując ten wykres, można sprawdzić, czy między zmiennymi Y i X istnieje zależność liniowa. Sprawdza się także, czy w próbie znajdują się obserwacje odstające, czyli takie, które nie pasują do ogólnego trendu, który odpowiada większości obserwacji. Poniższy rysunek (2.1) przedstawia przykładowy wykres rozproszenia, na którym obserwowana jest silna zależność liniowa. Ponieważ wraz ze wzrostem wartości jednej zmiennej, rosną wartości drugiej w sposób liniowy.



Rysunek 2.1: Przykładowy wykres rozproszenia XY. *Źródło: opracowanie własne.*

Jeśli stworzony wykres rozproszenia dla próby oraz wartość próbkowego współczynnika r_{XY} potwierdzają, że zmienna objaśniająca X wpływa liniowo na zmienną Y przyjmuje się, że model regresji liniowej jest odpowiedni. Jeśli tak jest, zależność między wartościami zmiennej Y a wartościami zmiennej objaśniającej ma postać [8]:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ &\vdots \qquad\qquad \vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n, \end{aligned}$$

gdzie $\epsilon_1, \ldots, \epsilon_n$ to nieskorelowanymi zmiennymi losowymi ze średnią zero i nieznaną wariancją σ^2, będącą liczbową wartością.

Szacując wartości nieznanych parametrów β_0, β_1 wykorzystuje się metodę najmniejszych kwadratów.
Niech

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2$$

Definicja 2.1. [8] Prostą regresji opartą na metodzie najmniejszych kwadratów nazywa się prostą:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

dla której funkcja $S(\beta_0, \beta_1)$ osiąga wartość najmniejszą względem β_0, β_1.

Szukane **estymatory nieznanych parametrów modelu** β_0 **i** β_1, wyznaczone metodą najmniejszych kwadratów przyjmują następującą postać [15]:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i-1}^{n}(x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

W celu sprawdzenia poprawności przeprowadzonej estymacji należy wyznaczyć błędy standardowe estymatorów. Aby to uczynić, warto zdefiniować estymator wariancji σ^2, zwany błędem średniokwadratowym S^2 [17]:

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

Dla tak zdefiniowanego estymatora σ^2 można już wyznaczyć błędy standardowe, które określają na, ile dany parametr może się zmieniać w różnych badaniach tego samego zjawiska [16]:

$$std(\hat{\beta}_1) = S\sqrt{\frac{1}{n}\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$std(\hat{\beta}_0) = S\sqrt{\frac{1}{n}\frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

2.1.1 Predykcja i przedziały ufności

Dopasowywanie modelu do danych ma na celu umożliwienie przewidywania, jaką wartość przyjmie zmienna objaśniana Y, gdy zmienna objaśniająca X przyjmie ustaloną wartość x_0. Można też skonstruować przedział ufności dla tej przyszłej wartości zmiennej objaśnianej.

Wspomniany wyżej przedział ufności dla predykowanych danych ma postać [16]:

$$\hat{Y}(x_0) \pm t_{1-\frac{\alpha}{2},n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{2.1}$$

gdzie:

- x_0 - obserwacja, dla której będzie przeprowadzona predykcja
- $\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- $t_{1-\frac{\alpha}{2},n-2}$ oznacza kwantyl rzędu $1 - \frac{\alpha}{2}$ rozkładu t-Studenta z $n-2$ stopniami swobody.

2.2 Model liniowy regresji wielokrotnej

Rzadko zdarza się, że na wartość zmiennej objaśnianej ma wpływ tylko jeden czynnik. W sytuacji, w której więcej zmiennych objaśniających ma wpływ na zmienną zależną, mówi się o regresji wielokrotnej. Dla takiego przypadku należy korzystać z ogólnego wzoru [8]:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon_i \tag{2.2}$$

gdzie:

- $\beta_0, \ldots, \beta_{p-1}$ – nieznane parametry.
- $X_1, \ldots, x_{p-1}$ – zmienne objaśniające.

Można zastąpić n powyższych równości (2.2) przechodząc na zapis macierzowy modelu regresji wielokrotnej.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

gdzie: $\mathbf{Y}' = (Y_1, \ldots, Y_n)$ to wektor zmiennych zależnych, $\epsilon' = (\epsilon_1, \ldots, \epsilon_n)$ to wektor błędów. Natomiast $\beta' = (\beta_0, \ldots, \beta_{p-1})$ to wektor współczynników.Należy jeszcze zdefiniować macierz X, której i-ty wiersz odpowiada wartościom zmiennych objaśniających dla i-tej zmiennej objaśnianej.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \ldots & x_{n,p-1} \\ 1 & x_{2,1} & \ldots & x_{n,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,p-1} \end{bmatrix}$$

W pierwszej kolumnie wszędzie występują jedynki, ponieważ sztucznie dodaje się wartość x_0, która odpowiada parametrowi β_0 i zawsze jest równa 1.

Aby przejść do wyznaczenia estymatorów dla wektora współczynników β, potrzebne jest wprowadzenie jeszcze kilku oznaczeń. Z założenia, że zmienne Y_i, Y_j dla $i \neq j$ są nieskorelowane wynika, że $Cov(Y_i, Y_j) = 0$. Wprowadzając więc macierz kowariancji $\sum_{\mathbf{Y}}$ wektora $\mathbf{Y}$ otrzymuje się, że $\sum_{\mathbf{Y}} = \sigma^2\mathbf{I}$. Gdzie $\mathbf{I}$, jest macierzą jednostkową, czyli taką, która posiada jedynki na przekątnej i zera poza nią [8]. Ponieważ Y z reguły nie jest prostą próbą losową, należy poddać go pewnym przekształceniom. Niech $\mathbf{Xb}$ będzie rzutem prostokątnym wektora Y na płaszczyznę.

Podobnie jak dla jednej zmiennej, wybierając estymatory, należy znaleźć minimum funkcji będącej sumą kwadratów odległości rzeczywistych wartości zmiennej Y od wartości prognozowanych przez model [8]:

$$Q(\mathbf{b}) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 x_{i,1} + \ldots + b_{p-1} x_{i,p-1}))^2 = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$$

Szukanie minimum takiej funkcji odpowiada poszukiwaniu wektora $\mathbf{b}$, dla którego odległość wektora $\mathbf{Y}$ od wektora $\mathbf{Xb}$ jest równa odległości wektora $\mathbf{Y}$ od zbioru wektorów $\mathbf{z}$ tworzących płaszczyznę, na którą rzutowany jest wektor $\mathbf{Y}$. Dlatego należy wyznaczyć wektor $\frac{\mathbf{d}}{\mathbf{db}(Q)}$, którego współrzędne będą cząstkowymi pochodnymi funkcji Q.

Wektor **b**, który minimalizuje funkcję Q, jest estymatorem wyznaczonym metodą najmniejszych kwadratów i przyjmuje postać [8]:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Jest to jawna postać estymatora, jednak aby można było zapisać go w takiej postaci, spełniony musi być jeden warunek. Mianowicie kolumny macierzy $\mathbf{X}'\mathbf{X}$ muszą być liniowo niezależne, aby można było odwrócić macierz.

2.2.1 Predykcja i przedziały ufności

Cel prognozy jak i metody, które wykorzystuje się do predykcji danych, są analogiczne jak w przypadku regresji jednokrotnej. Wartość $\hat{\mathbf{Y}}(\mathbf{x_0})$ będzie wyestymowaną wartością z dopasowanego modelu dla wektora $\mathbf{x_0}$. Wyznaczone zostaną przedziały ufności, a na koniec dokonane będzie sprawdzenie, czy rzeczywiste wartości zmieściłyby się w wyznaczonym przedziale ufności [15]. Do wyznaczenia przedziałów ufności potrzebne będzie wprowadzenie estymatora wariancji dla modelu regresji wielokrotnej. Do obliczenia, którego wykorzystuje się kwadrat różnicy pomiędzy rzeczywistą wartością zmiennej losowej a wartością predykowaną przy pomocy modelu. Estymator ten ma postać [17]:

$$S^2 = \frac{1}{(n-p)}\sum_{i=1}^{n} e_i^2$$

gdzie n to liczność próby, a p to liczba predykatorów.

Dla regresji wielokrotnej przedziały ufności wyraża się wzorem [8]:

$$Y(\hat{x}_0) \pm t_{1-\alpha/2,n-p} SE_{Y(\hat{x}_0)-Y(x_0)}$$

gdzie:

- $SE^2_{\hat{Y}(x_0)-Y(x_0)} = S^2(1 + x_0'(\boldsymbol{X}^T\boldsymbol{X})^{-1})$

2.3 Wartości resztowe

Residua mówią o tym, o ile estymowana wartość różni się od rzeczywistej wartości zmiennej losowej. Można je zapisać przy pomocy poniższego wzoru [14]:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

gdzie $\hat{\mathbf{Y}} = \mathbf{x_i b}$ jest wektorem predykowanych wartości.

2.4 Ocena zależności liniowej

Dopasowując do danych model liniowy, ważne jest, aby dane rzeczywiście miały charakter liniowy. Aby to sprawdzić, można skorzystać ze wskaźników, które oceniają zależność liniową, takich jak [16]:

- SSE – suma kwadratów błędów, Można ją interpretować jako indeks zmienności residuów wokół swojej średniej równej 0.

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

- SSR – suma regresyjna kwadratów. Można ją interpretować jako indeks zmienności wartości przewidywanych$\hat{y}_i$ wokół swojej średniej $\overline{\hat{y}_i}$.

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

- SST – całkowita suma kwadratów. Stanowi miarę całkowitej zmienności wartości y. Mówi się, że SST jest funkcją wariancji zmiennej y. Można pokazać, że:

$$SST = SSE + SSR$$

- Współczynnik korelacji Pearsona opisany w rozdziale pierwszym.

- Współczynnik determinacji R^2 - miara jakości dopasowania modelu. Mówi o tym, jak zmienność jednej zmiennej wyjaśnia zmienność drugiej. Przyjmuje on wartości od 0 do 1. Wartość maksymalna tego współczynnika jest osiągana, gdy regresja idealnie pasuje do danych.Wówczas nie ma błędów oszacowania, a więc wartości resztowe (residua) wynoszą 0, a więc SSE=0.

$$R^2 = \frac{SSR}{SST}$$

2.5 Diagnostyka modelu

Na koniec należy sprawdzić, czy model został dobrze dopasowany i czy spełnione są wszystkie założenia modelu. W tym celu wykonuje się F-test i analizę wartości resztowych.

F-test jest to test badający poprawność zastosowania modelu. Hipoteza zerowa tego testu mówi, że wszystkie współczynniki modelu regresji są równe 0. Czyli model nie może być wykorzystywany do predykcji danych rzeczywistych. Hipoteza alternatywna mówi zaś, że istnieje chociaż jeden niezerowy współczynnik. Aby zinterpretować wynik F-testu, należy sprawdzić p-wartość, czyli prawdopodobieństwo, tego, że przy prawdziwości hipotezy zerowej statystyka testowa przyjmie wartość taką jak w próbie lub większą. Jeśli p-wartość jest mniejsza niż założony poziom istotności α można odrzucić hipotezę zerową [7].

Analiza residuów. Jeśli spełnione są założenia modelu regresji liniowej, residua $e_1, \ldots, e_n$ niewiele się różnią od błedów $\epsilon_1, \ldots, \epsilon_n$. Ponieważ błędy powinny mieć rozkład normalny, adekwatność modelu sprawdza się za pomocą wykresu kwantylowego, badającego zgodność rezyduów z rozkładem normalnym. Poprawność modelu bada się także za pomocą wykresów rozproszenia dla prób $(\hat{y}_1, e_1), \ldots, (\hat{y}_n, e_n)$ oraz $(x_{1j}, e_1), \ldots, (x_{nj}, e_n)$. Na każdym z tych wykresów, punkty powinny oscylować wokół osi OX, bez żadnej zauważalnej tendencji [16].

Jednak to, co różni przypadek regresji liniowej jednokrotnej od regresji liniowej wielokrotnej to fakt, że dla tej drugiej należy zbadać współliniowość, a także dokonać wyboru zmiennych objaśniających w modelu.

Współliniowość. Jest to sytuacja, w której wektory zmiennych objaśniających są do siebie równoległe. Taka sytuacja może powodować dużą zmienność estymatorów, dlatego warto ją zbadać. Współliniowość można wykryć, badając współczynnik korelacji liniowej Pearsona, jednak można badać go tylko dla dwóch zmiennych. Dla przypadku z większą ilością zmiennych można zastosować współczynnik determinacji dla wielu zmiennych [1].

Wybór zmiennych objaśniających. Najlepszy model opisujący dane to ten, który jest dobrze dopasowany do danych i najprostszy. Z tego powodu dąży się do minimalizacji liczby zmiennych niezależnych, tak aby uwzględnić tylko te, które mają faktyczny wpływ na zmienną zależną. W tym celu można zastosować metodę eliminacji, która polega na stworzeniu modelu z największą możliwą liczbą zmiennych objaśniających, następnie testuje się istotność poszczególnych zmiennych w modelu i usuwa się tą zmienną, dla której p-wartość jest największa i przekracza ustalony poziom istotności α. Procedurę powtarza się aż do uzyskania optymalnego modelu [1].

2.6 Testy statystyczne wykorzystane do sprawdzenia założeń odnośnie residuów

Aby zbadać wspomniane w poprzedniej części założenia dotyczące wartości resztowych, można zastosować następujące testy statystyczne:

1. **Arch test** – Autoregresywna warunkowa heteroskedastyczność to model statystyczny używany do analizy zmienności. Wykorzystany zostanie do zbadania stałości wariancji. Jego hipoteza zerowa zakłada, brak heteroskedastyczności, czyli tego, że przynajmniej jedna wartość będzie różniła się wariancją od pozostałych. Jeśli hipoteza zerowa zostanie przyjęta, będzie można wnioskować, że wariancja wartości resztowych jest stała, w przypadku przyjęcia hipotezy alternatywnej, głoszącej, że w wektorze znajduje się przynajmniej jedna wartość, dla której wariancja odstaje od pozostałych, należy przyjąć brak stałej wariancji [10].

2. **Test Ljung'a-Box'a** – używany jest do oceny zależności między danymi. Za hipotezę zerową test ten przyjmuje, że korelacja między obserwacjami równa jest 0. Hipoteza alternatywna natomiast mówi, że są one zależne [10].

3. **Test Kołomogorowa-Smirnowa** – najczęściej stosowany test statystyczny sprawdzający, czy próba pochodzi z rozkładu normalnego o średniej 0 i wariancji 1. Bazuje on na różnicy pomiędzy dystrybuantą teoretyczną a empiryczną. Hipoteza zerowa zakłada właśnieże próba pochodzi z rozkładu normalnego, natomiast alternatywna odrzuca to założenie [7].

4. **Test Anderson-Darling** – test bazuje na różnicy pomiędzy dystrybuantą empiryczną a teoretyczną. Jest jednak bardziej wrażliwy na różnice w ogonach rozkładu niż kstest, a za hipotezę zerową również przyjmuje, że próba pochodzi z rozkładu normalnego. Hipoteza alternatywna odrzuca to założenie [10].

Rozdział 3

Regresja logistyczna

Regresja logistyczna to najpopularniejszy model wykorzystywany do opisywania zależności między dychotomiczną zmienną objaśnianą y, czyli taką, która przyjmuje tylko dwie wartości 0 albo 1, a zmiennymi objaśniającymi $X_1, X_2, \ldots, X_p$.

Niech $\boldsymbol{x} = (x_1, \ldots, x_p)$, z $x_1 = 1$, będzie wartością przyjętą przez wektor $\boldsymbol{X} = (X_1, \ldots, X_p)$, który został utworzony ze zmiennych objaśniających.

Modelu regresji logistycznej zakłada, że prawdpodododobieństwo warunkowe $\pi(\boldsymbol{x})$ przyjęcia przez zmienną Y wartości 1, pod warunkiem, że zmienne objaśniające $\boldsymbol{X}$ przyjęły wartość $\boldsymbol{x} = (x_1, \ldots, x_p)$, wyraża się wzorem [2]:

$$\pi(\boldsymbol{x}) = P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\mathrm{e}^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1 + \mathrm{e}^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}.$$

Gdzie symbole $\alpha, \beta_1, \ldots, \beta_p$ są nieznanymi parametrami modelu wyznaczonymi z wykorzystaniem danych.

Nieznane parametry w modelu regresji logistycznej estymuje się, wykorzystując metodę największej wiarygodności. Polega ona na wyznaczeniu punktu $(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p)$, w którym funkcja wiarygodności osiąga maksimum globalne. W rozważanym przypadku funkcja wiarogodności ma jednak tak skomplikowaną postać, że szukany punkt $(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ można wyznaczyć jedynie numerycznie, wykorzystując algorytm Newtona-Raphsona.

W dalszej części pracy wykorzystany będzie model regresji logistycznej z jedną zmienną objaśniającą, w celu uproszczenia rozważań.

Regresja logistyczna odchodzi od standardowego pojęcia prawdopodobieństwa klasycznego, które polega na wyliczeniu stosunku sukcesów do wszystkich prób. W tym modelu chodzi o to, aby obliczyć szansę (ang. odds), czyli stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki [2]:

$$Odds = \frac{\pi(x)}{1 - \pi(x)} = e^{\alpha} e^{\beta x}$$

Aby przekształcić prawdopodobieństwo na logarytm, stosuje się tzw. funkcję logitową. Dzięki temu można zauważyć, że zlogarytmowane szanse mają omawiany we wcześniejszym rozdziale charakter liniowy.

Po drobnych przekształceniach w celu wyznaczenia funkcji $\pi(x)$ model regresji logistycznej przyjmuje więc postać [2]:

$$\hat{\pi}(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{3.1}$$

Współczynnik β odpowiada za to, czy funkcja $\pi(x)$ rośnie wraz ze wzrostem x, czy też maleje. Jeżeli $\beta = 0$ to Y jest niezależny od zmian X. Z racji tego, że funkcja określa prawdopodobieństwo, zbiorem wartości funkcji $\pi(x)$ jest przedział $[0, 1]$.

3.1 Predykcja i przedziały ufności

Dla modelu z jedną zmienną objaśniającą można zapisać go w następujący sposób [1]:

$$logit[\pi(x)] = \alpha + \beta x$$

Aby móc stosować model regresji logistycznej, ważne jest, aby $\beta \neq 0$ ponieważ wtedy nie ma zależności pomiędzy X i Y. Należy więc przetestować hipotezę o niezależności. W tym celu wykorzystany zostanie test Wolda, w którym używa się estymatora $\hat{\beta}$ wyznaczonego metodą największej wiarygodności i statystyki testowej $z = \hat{\beta}/SE$, gdzie SE to błąd standardowy. Statystyka ta ma w przybliżeniu rozkład chi-kwadrat, a więc aby ocenić istotność współczynnika, wykorzystuje się kwantyle rozkładu chi-kwadrat. Test zakłada, że wartość $\mathcal{X}^2(\alpha)$ nie może być większe niż $z^2_{\frac{\alpha}{2}}$. Parametr β zawiera się więc w przedziale $\hat{\beta} \pm z_{\frac{\alpha}{2}}(SE)$ [2].

W kontekście wyznaczania przedziałów ufności inne wartości mogą mieć większe znaczenie niż szacowanie β. Model ma za zadanie wyznaczać prawdopodobieństwo, więc warto byłoby wyznaczyć przedziały ufności dla $\pi(x)$.

Na poziomie ufności $1 - \alpha = 0.95$ przedział ufności dla $logit[\pi(x)] = (\hat{\alpha} + \hat{\beta}x) \pm 1.96SE$. Gdzie SE dla konkretnego $x = x_0$ estymuje się, korzystając z pierwiastka kwadratowego wyciągniętego z poniższego wzoru [4]:

$$Var(\hat{\alpha} + \hat{\beta}x_0) = Var(\hat{\alpha}) + x_0 Var(\hat{\beta}) + 2x_0 Cov(\hat{\alpha}, \hat{\beta})$$

3.2 Diagnostyka modelu

Ostatnim krokiem po dopasowaniu modelu jest sprawdzenie poprawności jego zastosowania. Regresja logistyczna nie posiada zbyt wielu założeń odnośnie wartości resztowych, tak jak to było w przypadku regresji liniowej. Dlatego w tym celu wykorzystywać można różnego rodzaju testy i współczynniki m.in. współczynnik determinacji R^2, oraz jego zmodyfikowaną wersję pseudo R^2 [4].

Współczynnik determinacji, którego wartość reprezentuje procentową zmienność danych uwzględnionych przez model. Wynik znajduje się w zakresie od 0 do 1, im wyższa

wartość R^2, tym lepsze dopasowanie modelu. Współczynnik ten został opisany w podrozdziale (2.4).

Pseudo R^2 jest miarą dopasowania modelu dla regresji logistycznej. Polega na proporcjonalnej redukcji logarytmu funkcji wiarygodności. Przyjmuje wartości z przedziału $[0, 1]$ i jak przy poprzednich współczynnikach 0 odpowiada brakowi dopasowania, natomiast 1 bardzo dobre dopasowanie. Wyraża się on wzorem [5]:

$$R_p^2 = 1 - \frac{\ln L_p}{lnL_0}$$

gdzie L_p jest funkcją wiarygodności dla pełnego modelu, a L_0 funkcją wiarygodności dla modelu zawierającego jedynie wyraz wolny.

Rozdział 4

Wyniki własne

4.1 Przedstawienie danych poddanych analizie

Do analizy wykorzystane zostały dane dotyczące siatkówki. Zostały one zebrane w trakcie Igrzysk Olimpijskich 2020 w Tokio, odbywających się między 23 lipca 2021 a 8 sierpnia 2021. Analizowany zbiór pochodzi z fazy eliminacji grupowej i składa się z punktów zdobytych przez poszczególnych zawodników w kolejnych spotkaniach oraz ich fizycznych parametrów. Wybrani zostali zawodnicy grający na pozycjach przyjmującego i atakującego, ponieważ statystycznie rozgrywający najczęściej zagrywa do zawodników grających na tych pozycjach. Przeważnie atakują oni ze skrzydeł boiska. Szanse na to, że zawodnik grający na jednej z tych pozycji dostanie możliwość ataku, są więc zbliżone. Analizując tabele punktów zdobytych przez zawodników grających na wszystkich pozycjach ewidentnie widać, że siatkarze grający jako przyjmujący bądź atakujący punktują najlepiej. W fazie eliminacji grupowej wszystkie zespoły zostały podzielone na dwie grupy. W każdej z grup było po 6 zespołów, w związku z czym każdy zespół rozegrał w tej fazie 5 spotkań. Z każdego zespołu wybrano dwóch lub trzech najlepiej punktujących zawodników grających na rozważanych pozycjach. Dane odnośnie zdobytych punktów pochodzą ze oficjalnej strony FIVB [1], czyli międzynarodowej federacji piłki siatkowej. Informacje o parametrach zawodnika, takich jak wzrost, waga i zasięg ataku pochodzą natomiast z Wikipedii – Wolnej Encyklopedii, po wyszukaniu w niej każdego zawodnika z osobna. Analizowana próbka składa się z 36 obserwacji. Każdemu z zawodników przyporządkowane zostały kolejno numery od 0 do 35, a kolejność siatkarzy w tabeli jest przypadkowa. Wszystkie przedstawione w tej części wykresy zostały wykonane przy użyciu języka programowania Python.

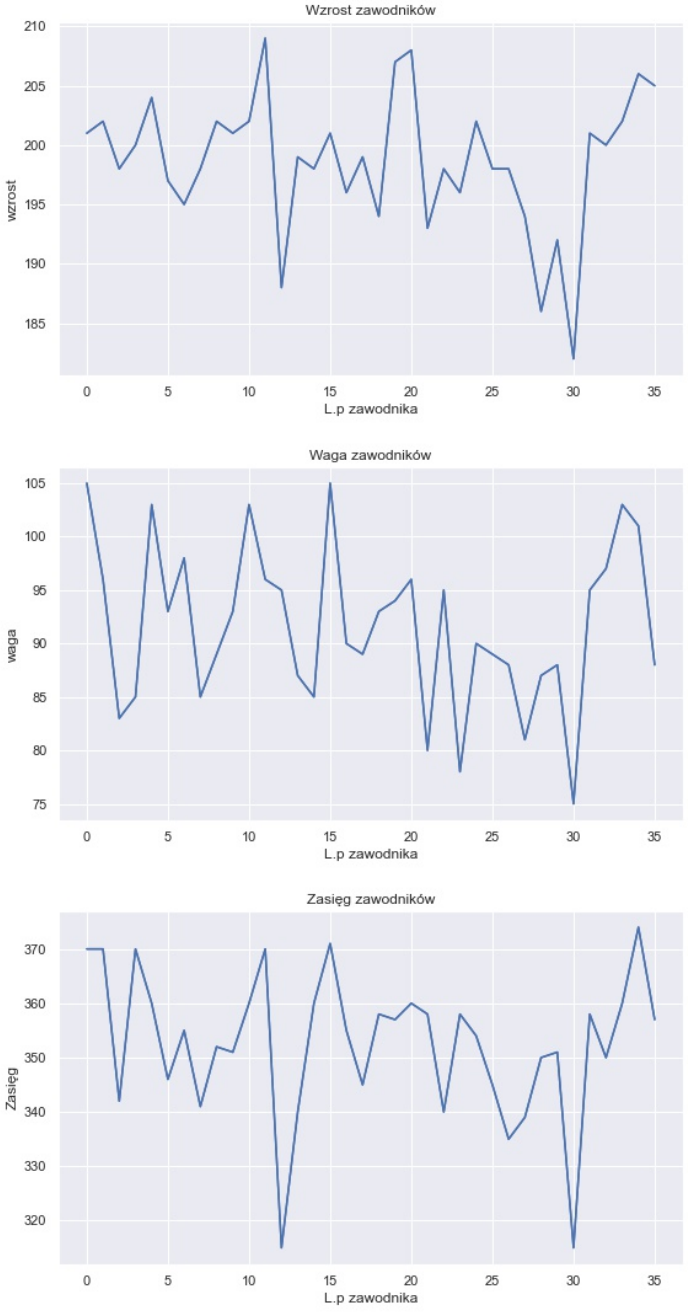
Tabela 4.1: Tabela przedstawiająca fragment analizowanych danych.

L.p	N	P	H	W	S	M1	M2	M3	M4	M5
0	Kurek Bartosz	O	201	105	370	20	14	1	11	17
1	Leon Venero Wilfredo	OH	202	96	370	23	18	0	17	16
2	Sliwka Aleksander	OH	198	83	342	13	9	0	13	0
3	Juantorena Osmany	OH	200	85	370	21	9	22	20	17
4	Zaytsev Ivan	OH	204	103	360	4	7	18	12	0

Źródło: opracowanie własne

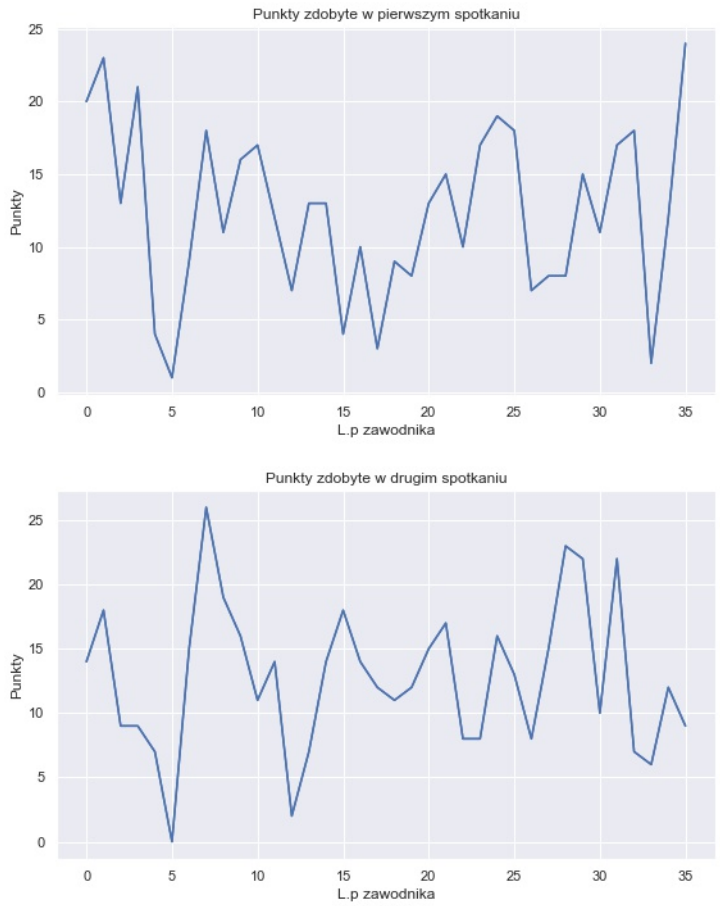
[1] `https://www.fivb.com/en/volleyball/competitions`

W tabeli (4.1) przyjęto oznaczenia w celu zwiększenia jej czytelności. *L.p* to liczba porządkowa zawodnika w tabeli, przydzielona losowo. *N* (*ang. name*) oznacza imię i nazwisko zawodnika, natomiast *P* pozycję, na której gra siatkarz. Litera *O* (*ang. Outside Hitter*) odpowiada pozycji atakującego, natomiast *OH* (*ang. Opposite Hitter*) odpowiada pozycji przyjmującego. Pod literą *H* (*ang. height*) znajduje się wzrost zawodnika w centymetrach, a jako *W* (*ang. weight*) została oznaczona jego waga w kilogramach. *S* (*ang. spike*) to zasięg w ataku zawodnika wyrażony w centymetrach. Natomiast wartości M_i dla $i \in \{1, 2, 3, 4, 5\}$ to punkty zdobyte w kolejno rozgrywanych spotkaniach.



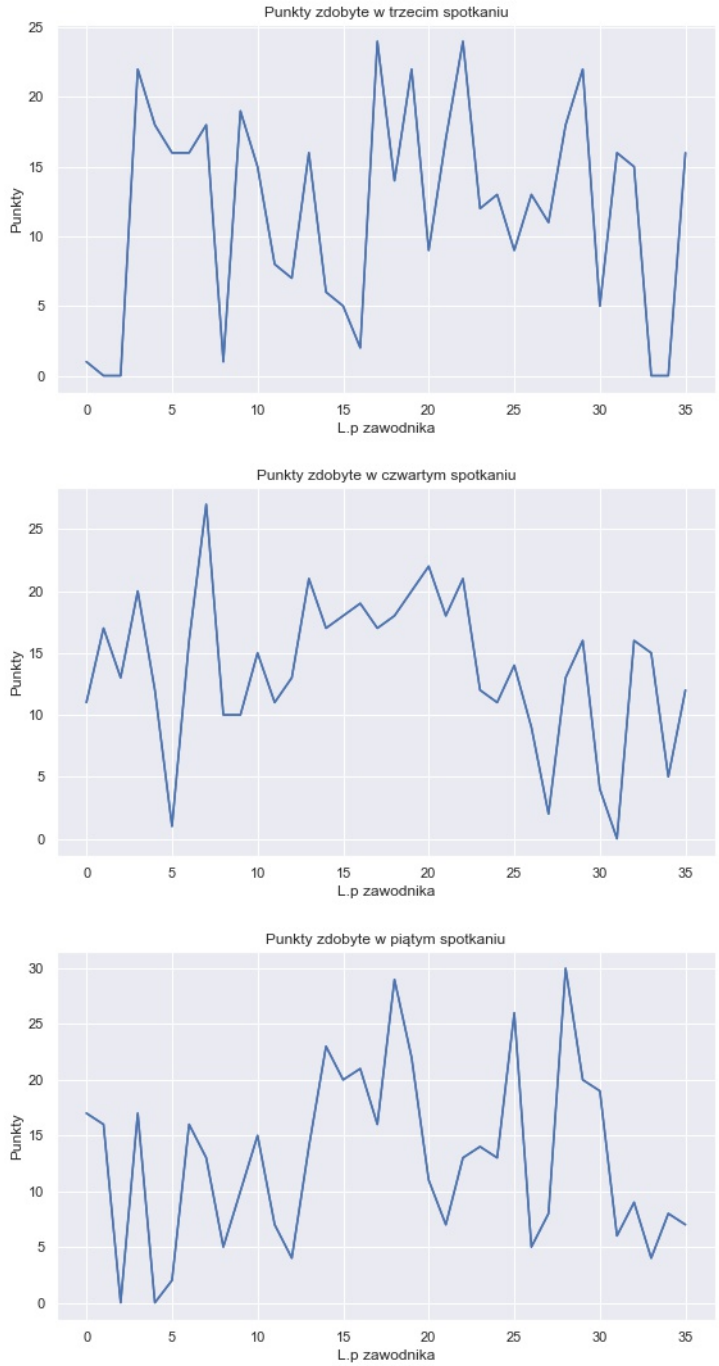
Rysunek 4.1: Wykresy przedstawiające parametry fizyczne zawodników. *Źródło: opracowanie własne.*

Wykresy przedstawione na rysunku (4.1) przedstawiają, jak w analizowanym zbiorze zachowują się poszczególne parametry fizyczne zawodników. Na osi OX znajduje się liczba porządkowa zawodnika z tabeli (4.1) natomiast na osi OY opisane parametry fizyczne: wzrost, waga i zasięg. Widać, że wszystkie trzy trajektorie zachowują się w podobny sposób. Można było zakładać, że wyżsi zawodnicy będą osiągać większy zasięg w ataku i tak rzeczywiście jest. Kolejną rzeczą, którą można zauważyć, jest to, że wyżsi siatkarze osiągają również większą wagę, co jest dość naturalnym zjawiskiem. W związku, z czym można też wnioskować, że istnieje zależność między zasięgiem a wagą zawodnika. Gdy na jednym wykresie obserwowane są spadki lub wzrosty, kolejne dwa zachowują się analogicznie. Może to świadczyć o istniejącej korelacji między parametrami fizycznymi zawodników.



Rysunek 4.2: Wykresy przedstawiające liczbę punktów zdobytą przez poszczególnych zawodników w pierwszych dwóch spotkaniach. *Źródło: opracowanie własne.*

Na rysunkach (4.2) i (4.3) przedstawione zostały wykresy obrazujące liczbę punktów zdobywaną przez poszczególnych zawodników w kolejnych spotkaniach. Na osi OX znajduje się liczba porządkowa zawodnika, natomiast na osi OY liczba punktów zdobyta przez gracza w danym spotkaniu. Wykresy te są bardziej zróżnicowanie niż w przypadku parametrów fizycznych zawodników. Zdarza się, że siatkarz, który punktował w jednym spotkaniu bardzo dobrze, nie dostał szansy zagrania w kolejnym spotkaniu. Może to świadczyć o tym, że we wcześniejszym spotkaniu zespół grał z teoretycznie słabszym zespołem i wystawiony został zawodnik niebędący graczem wyjściowej szóstki, przez co miał szansę zdobyć więcej punktów. Natomiast w kolejnym spotkaniu trener mógł postawić na gracza z podstawowego składu, na mecz z silniejszym zespołem, a z kolei ten miał trudniejsze zadanie, aby w tym



Rysunek 4.3: Wykresy przedstawiające liczbę punktów zdobytą przez poszczególnych zawodników w ostatnich trzech spotkaniach. *Źródło: opracowanie własne.*

spotkaniu dobrze punktować. Jednak przyglądając się wykresom, można zauważyć pewne występujące miejscami analogiczne zachowania. Nie można więc wykluczyć, że dane są w pewien sposób ze sobą skorelowane. Jednak zostanie to sprawdzone w dalszej części pracy.

W celu wstępnego zapoznania się z danymi wyznaczone zostały podstawowe charakterystyki liczbowe. Wyniki przedstawiono w tabeli (4.2). Wyznaczono średnią arytmetyczna dla każdej analizowanej kategorii danych. Patrząc na wartości wariancji dla parametrów fizycznych zawodników, widać, że dane nie są mocno zróżnicowane. Świadczy to zapewne

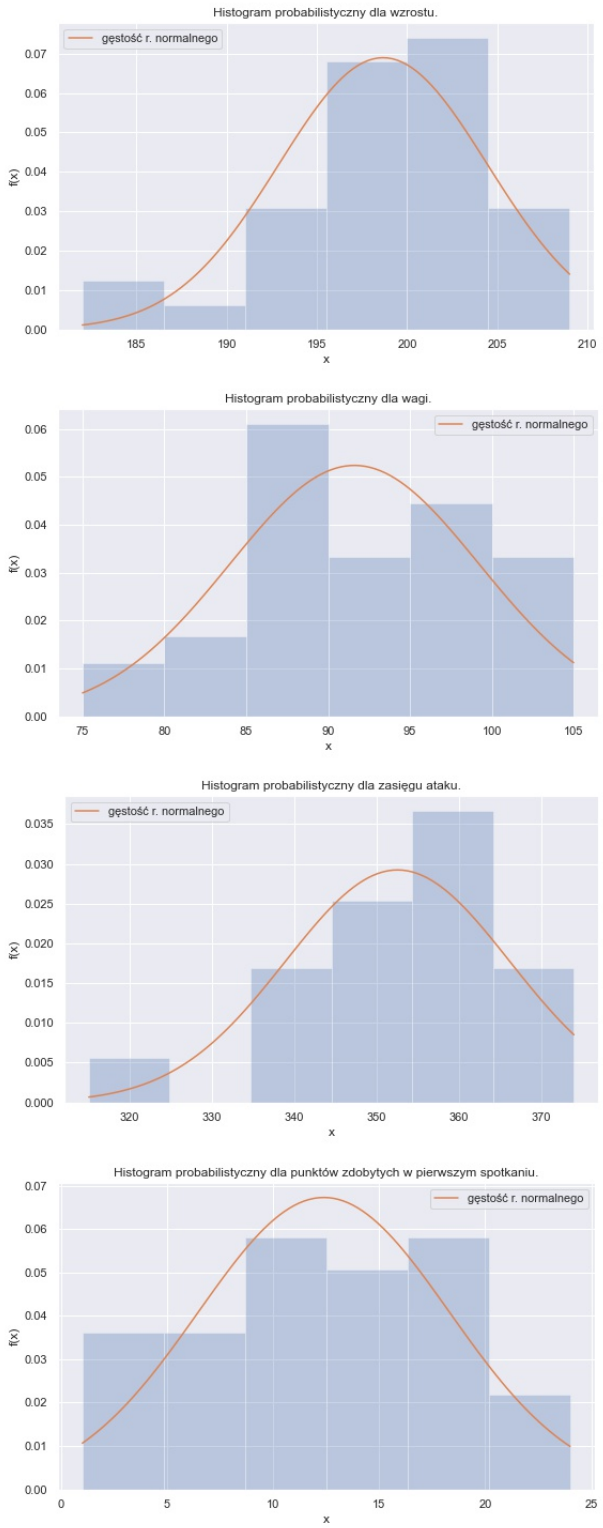
o tym, że analizowany zbiór dotyczy profesjonalnych, zawodowych sportowców, którzy muszą posiadać pewne warunki fizyczne. Rozstęp próby dla wzrostu jest dość duży i wynosi 27 cm, jednak minimalna wartość może zostać uznana za wartość odstającą, patrząc na wykresy przedstawione na rysunku (4.1) widać, że tylko trzech zawodników osiągnęło wzrost poniżej 185 cm. Reszta rozkłada się całkiem równomiernie między 195 cm a 205 cm. Podobnie wygląda sytuacja dla wagi i zasięgu. Z Tabeli można odczytać, że we wszystkich spotkaniach średnia punktów zdobytych przez zawodników jest zbliżona. Oscyluje między 12 a 13 punktami. Patrząc na wartości minimalne punktów w kolejnych spotkaniach widać, że w każdym spotkaniu znalazł się zawodnik, który nie miał szansy wystąpić na boisku. W każdym spotkaniu maksymalna liczba zdobytych punktów przekracza wartość 20. W związku z tym, że w prawie każdym zbiorze minimalna wartość to 0, rozstęp próby jest dość duży. Mediana, czyli wartość środkowa, mówi o tym, że ponad tą wartością znajduje się połowa obserwacji ze zbioru. W przypadku punktów mediana jest zbliżona do średniej, przyjmuje wartości w między 12.5 a 14.5 punktu. Świadczy to o tym, że połowa znajdujących się w zbiorze zawodników zdobyła więcej niż ta liczba punktów. Wariancja mówiąca o zróżnicowaniu wartości w zbiorze przyjmuje jednak wysokie wartości. Jest to całkiem zrozumiałe, ponieważ przyjmowane są zarówno ekstremalnie duże wartości takie jak 30 jak i wartości równe 0.

Tabela 4.2: Podstawowe charakterystyki analizowanego zbioru danych.

Charakterystyka	H	W	S	M1	M2	M3	M4	M5
Średnia	198.67	91.61	352.56	12.39	12.75	11.94	13.78	12.97
Wariancja	33.49	58.02	186.14	35.16	32.99	56.74	37.26	61.23
Odchylenie std.	5.79	7.62	13.64	5.93	5.74	7.53	6.10	7.82
Mediana	199.0	91.50	355.0	12.5	12.5	13.5	14.5	13.0
Minimum	182.0	75.0	315	1	0	0	0	0
Maksimum	209.0	105.0	374	24	26	24	27	30
Rozstęp	27.0	30.0	59	23	26	24	27	30

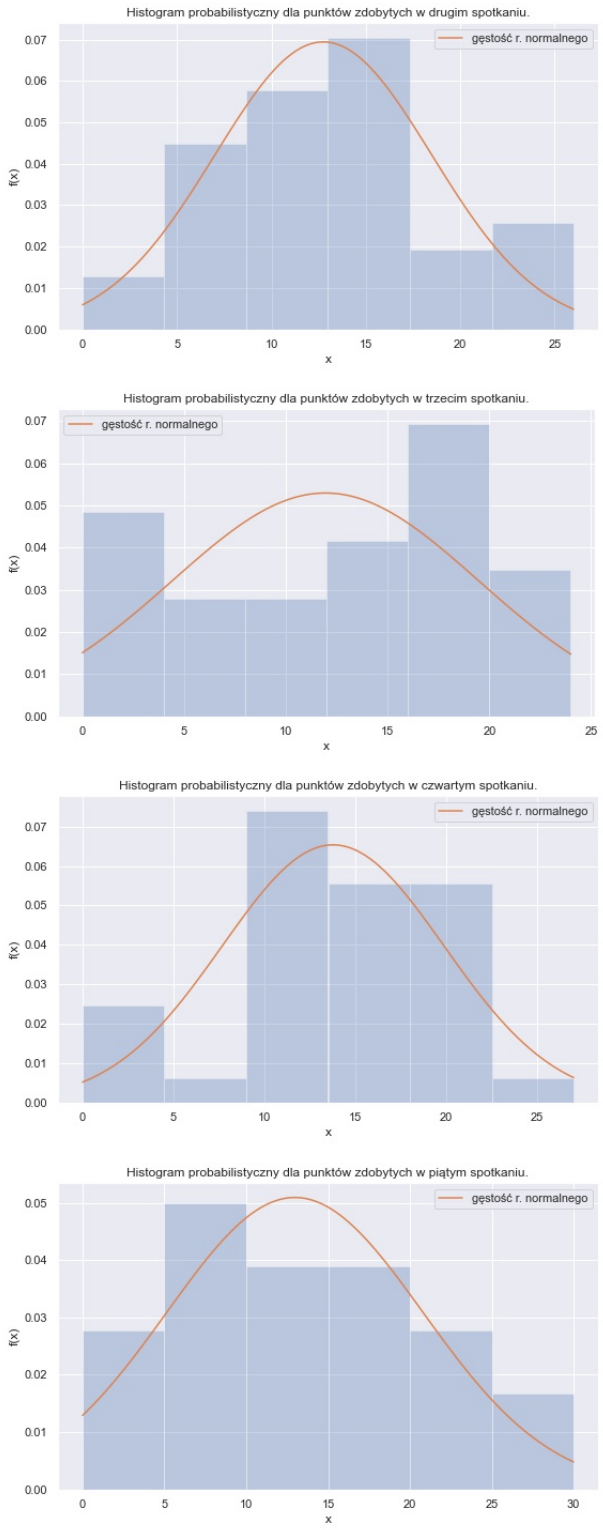
Źródło: opracowanie własne

Przed przystąpieniem do analizy korelacji, dokonane zostało sprawdzenie, czy analizowane dane mają rozkład normalny. Jest to rozkład, którym najczęściej chciałoby się estymować dane. Przyczyną wykonania sprawdzenia jest też fakt, że to czy dane pochodzą z takiego rozkładu, jest jednym z założeń do poprawnej interpretacji współczynnika korelacji Pearsona. Narysowano więc histogramy, za pomocą których zbadano zgodność każdej próby z rozkładem normalnym o średniej i wariancji wyestymowanej na podstawie próby. Oprócz tego wykonany został test Anderson-Darling, który jest wykorzystywany do sprawdzenia, czy dane posiadają rozkład normalny. Hipoteza zerowa tego testu mówi, że tak jest. Natomiast hipoteza alternatywna odrzuca to założenie. Jeśli p-wartość, odpowiadająca zaobserwowanej wartości statystyki testowej, jest mniejsza od przyjętego poziomu istotnościα, to hipoteza zerowa jest odrzucana, bo zaszło zdarzenie nietypowe przy prawdziwości tej hipotezy. A gdy p-wartość jest większa od α, to nie ma podstaw do odrzucenia hipotezy zerowej.



Rysunek 4.4: Wykresy przedstawiające porównanie histogramów probabilistycznych z gęstością rozkładu normalnego. *Źródło: opracowanie własne.*

We wszystkich przedstawionych na rysunkach (4.4) i (4.5) wykresach wartości znajdujące się na osi OX to wartości przyjmowane w każdym z rozważanych zbiorów danych. Natomiast na osi OY znajdują się wartości funkcji gęstości rozkładu. Patrząc na wykresy widać, że histogramy zachowują się w zbliżony sposób do gęstości rozkładu normalnego



Rysunek 4.5: Wykresy przedstawiające porównanie histogramów probabilistycznych z gęstością rozkładu normalnego. *Źródło: opracowanie własne.*

z wyestymowanymi parametrami średniej i wariancji z próby. Można jednak wnioskować, że najgorsze dopasowanie z gęstością tego rozkładu występuje dla zasięgu gracza w ataku, oraz liczby punktów zdobytych przez zawodników w trzecim i czwartym spotkaniu. Należy jednak pamiętać o tym, że analizowane próbki zawierają tylko 36 obserwacji, może być

to powodem trudności w dopasowaniu rozkładu. Analizując więc same wykresy ciężko rozstrzygnąć, czy dane rzeczywiście mają taki rozkład. W tym celu warto zobaczyć wyniki wspomnianego wcześniej testu Anderson-Darling przedstawione w tabeli (4.3). Test został wykonany przy pomocy programu MATLAB. Wykorzystana została funkcja adtest, która zwraca dwie wartości. Pierwsza z nich h przyjmuje tylko dwie wartości 0 lub 1. Jeśli test zwraca 0 jest to równoważne, z tym że powinno się przyjąć hipotezę zerową za prawdziwą. Dla wartości 1, powinno się odrzucić hipotezę zerową. Test zwraca również drugą wartość p, która jest p-wartością wspomnianego testu.

Tabela 4.3: Wyniki testu Anderson-Darling.

	H	W	S	M1	M2	M3	M4	M5
h	0	0	1	0	0	0	0	0
p	0.1059	0.7101	0.0347	0.8752	0.7557	0.0507	0.2022	0.6984

Źródło: opracowanie własne

Analizując wyniki testu widać, że dla większości próbek nie ma podstaw do odrzucenia hipotezy zerowej o normalności rozkładu. Test odrzuca tę hipotezę tylko w przypadku zasięgu ataku zawodnika. W tym przypadku również p-wartość jest mniejsza od poziomu istotności $\alpha = 0.05$, więc odrzucamy hipotezę zerową, że zasięg ataku zawodnika ma rozkład normalny. Dla pozostałych zawartych w tabeli (4.3) parametrów, w których test nie odrzucił poprawności hipotezy zerowej, p-wartości przekraczają poziom α. Można więc założyć, że po zbadaniu korelacji Pearsona interpretacja wyników będzie istotna statystycznie.

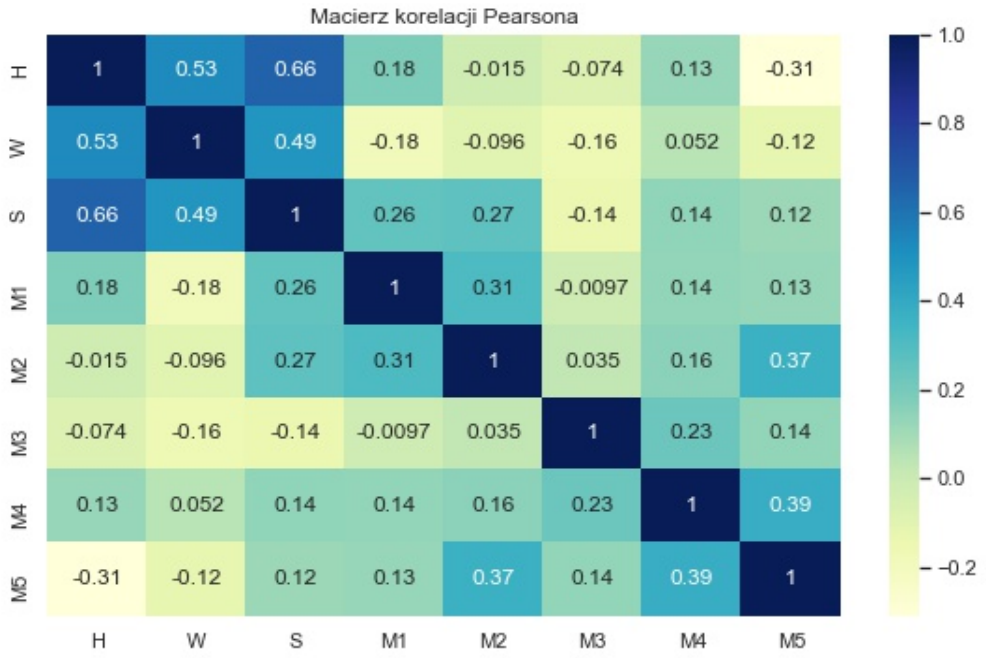
4.2 Analiza korelacji

W tej części przeprowadzona została analiza korelacji między wszystkimi przedstawionymi w poprzednim podrozdziale zbiorami danych. Wyznaczono wartości współczynników korelacji Pearsona, Spearmana i Kendalla. Przeanalizowano macierze korelacji i przeprowadzono test niezależności między zmiennymi.

4.2.1 Korelacja Pearsona

Pierwszy badany współczynnik, to współczynnik korelacji Pearsona, który bada liniową zależność między zmiennymi. Występuje tu założenie o tym, że aby móc interpretować jego wartości zmienne powinny pochodzić z rozkładu normalnego. Zostało to sprawdzone w poprzedniej części, gdzie okazało się, że dla większości zbiorów nie można wykluczyć, że mają rozkład normalny z parametrami średniej i wariancji wyestymowanymi z próby. Można więc od razu przejść do przedstawienia i interpretacji wyników.

Na wykresie (4.6) przedstawiona została macierz korelacji Pearsona. Zarówno na osi OX, jak i OY znajdują się oznaczenia odpowiadające wcześniej wprowadzonym próbkom danych. Wartości znajdujące się na polach, odpowiadają wartości współczynnika korelacji Pearsona pomiędzy odpowiednimi dwoma próbami. Na przekątnej znajdują się wartości równe 1, ponieważ jest to korelacja między tymi samymi zbiorami. Omówiona zostanie



Rysunek 4.6: Macierz z wartościami współczynnika korelacji Pearsona. *Źródło: opracowanie własne.*

tylko dolna macierz trójkątna, ponieważ wartości z górnej są takie same, dla tych samych par zbiorów. Z boku wykresu umieszczono skalę. Im ciemniejszy kolor na wykresie, tym większa wartość współczynnika, a co za tym idzie silniejsza korelacja liniowa między próbkami. Wartości dodatnie świadczą o dodatniej korelacji liniowej, czyli kiedy wartości jednej zmiennej rosną, drugiej także w sposób liniowy. Wartości ujemne natomiast świadczą o ujemnej zależności liniowej, czyli kiedy wartości jednej zmiennej rosną, drugiej maleją. Najciemniejsze poza jedynkami pola, to te odpowiadające korelacji między parametrami fizycznymi zawodników. Słuszna była więc obserwacja odnośnie wykresów (4.1) o możliwej występującej zależności między nimi. Najwyższa wartość współczynnika osiągnięta została dla wzrostu i zasięgu gracza. Jest to logiczne, ponieważ wzrost jest w pewnym sensie składową zasięgu ataku zawodnika. Patrząc na macierz, można zauważyć, że nie ma znaczącego związku pomiędzy parametrami fizycznymi zawodników a zdobywanymi przez nich punktami, może tak być dlatego, że rozważane dane pochodzą z rozgrywek zawodowych sportowców, gdzie parametry fizyczne są określone pewnymi wymogami. Jedyną silniejszą zależność zaobserwowano dla zasięgu ataku gracza i punktów zdobytych w pierwszym i drugim rozgrywanym meczu. Obserwowana jest korelacja między punktami zdobywanymi w kolejnych spotkaniach. Jeśli chodzi o punkty, najwyższe współczynniki osiągają punkty zdobyte w czwartym i piątym spotkaniu, oraz punkty zdobyte w drugim i piątym spotkaniu, a także punkty zdobyte w pierwszym i drugim spotkaniu. Wszystkie wspomniane wartości współczynników są dodatnie, a więc obserwowana jest dodatnia zależność liniowa między próbami.

Aby sprawdzić, czy korelacja rzeczywiście występuje pomiędzy badanymi próbkami, przeprowadzono test istotności. W tabeli (4.4) znajdują się p–wartości odpowiadające współczynnikom przedstawionym w macierzy z rysunku (4.6). W tym przypadku również analizowane będą wartości znajdujące się pod przekątną, na której znajdują się wartości równe 0. Przypominając, hipoteza zerowa tego testu mówi, że próbki są od siebie nieza-

Tabela 4.4: P-wartości dla współczynnika korelacji Pearsona.

	H	W	S	M1	M2	M3	M4	M5
H	0	0.0008	0	0.2834	0.9286	0.6686	0.4567	0.0681
W	**0.0008**	0	0.0026	0.2832	0.5789	0.346	0.7653	0.4962
S	**0**	**0.0026**	0	0.131	0.1151	0.4208	0.407	0.4905
M1	0.2834	0.2832	0.131	0	0.065	0.9551	0.4214	0.4471
M2	0.9286	0.5789	0.1151	**0.065**	0	0.8409	0.3622	0.0264
M3	0.6686	0.346	0.4208	0.9551	0.8409	0	0.1791	0.4152
M4	0.4567	0.7653	0.407	0.4214	0.3622	0.1791	0	0.0198
M5	0.0681	0.4962	0.4905	0.4471	**0.0264**	0.4152	**0.0198**	0

Źródło: opracowanie własne

leżne, natomiast hipoteza alternatywna głosi, że są zależne. Otrzymane p–wartości będą porównywane z poziomem istotności testu $\alpha = 0.05$. Im silniejsza zależność danych, tym mniejsza p–wartość. Więc dla wartości mniejszych od α można mówić o istnieniu zależności liniowej między zmiennymi. Jak widać na przekątnej znajdują się same 0, ponieważ korelacja wynosiła dla nich 1. Widać, że dla parametrów fizycznych zawodników, czyli tam, gdzie zaobserwowano najsilniejszą korelację liniową p–wartości są najmniejsze, mniejsze od $\alpha = 0.05$. Dla punktów zdobytych w czwartym i piątym spotkaniu oraz drugim i piątym p–wartości również mieszczą się poniżej poziomu istotności testu. Występująca między nimi korelacja jest istotna statystycznie. Poziom α zostaje nieznacznie przekroczony dla punktów zdobytych w pierwszym i drugim spotkaniu. Tam, gdzie współczynnik korelacji był niewielki obserwowane są wysokie p–wartości i odwrotnie. Potwierdzony został wniosek, o tym, że punkty zdobywane w kolejnych spotkaniach nie zależą od parametrów fizycznych siatkarzy.

4.2.2 Korelacja Spearmana

Kolejnym sposobem badania korelacji między zmiennymi jest wyznaczenie wartości współczynnika korelacji Spearmana, który jest rangową metodą badania ogólnej zależności monotonicznej, a więc nie musi to być tylko zależność liniowa. Nie posiada żadnych ograniczeń co do rozkładu, a więc można od razu przystąpić do analizy wyników.

Na wykresie (4.7) przedstawiona została macierz korelacji dla współczynnika Spearmana. W tym przypadku również najsilniejsza zależność obserwowana jest pomiędzy parametrami fizycznymi zawodników, tym razem jest to jednak zależność monotoniczna, niekoniecznie musi to być zależność liniowa. Najwyższy współczynnik korelacji został osiągnięty dla wzrostu i zasięgu ataku siatkarza. Nieznacznie mniejszą wartość przyjął współczynnik korelacji dla wzrostu i wagi. Podobnie jak dla korelacji Pearsona, również w tym przypadku parametry te nie są znacząco powiązane z liczbą zdobytych punktów w kolejnych spotkaniach. Dla wzrostu i wagi są to wręcz korelacje ujemne, czyli kiedy wzrost lub waga zawodnika rośnie, liczba zdobywanych punktów maleje. Jest to dość nietypową obserwacją. Patrząc na liczbę zdobywanych punktów, najsilniejsze korelacje występują dla punktów zdobytych w drugim i piątym meczu oraz czwartym i piątym. Wartość współczynnika dla punktów zdobytych w pierwszym i drugim spotkaniu jest

Rysunek 4.7: Macierz z wartościami współczynnika korelacji Spearmana. *Źródło: opracowanie własne.*

mniejsza niż w dla korelacji Pearsona.

Tabela 4.5: P-wartości dla współczynnika korelacji Spearmana.

	H	W	S	M1	M2	M3	M4	M5
H	0	0.0004	0.0002	0.2173	0.9895	0.4154	0.9319	0.1235
W	**0.0004**	0	0.0014	0.3556	0.6363	0.2767	0.9818	0.5125
S	**0.0002**	**0.0014**	0	0.1872	0.261	0.1353	0.5359	0.4681
M1	0.2173	0.3556	0.1872	0	0.1076	0.9391	0.647	0.3504
M2	0.9895	0.6363	0.261	0.1076	0	0.8759	0.7459	0.0389
M3	0.4154	0.2767	0.1353	0.9391	0.8759	0	0.1025	0.4473
M4	0.9319	0.9818	0.5359	0.647	0.7459	0.1025	0	0.0062
M5	0.1235	0.5125	0.4681	0.3504	**0.0389**	0.4473	**0.0062**	0

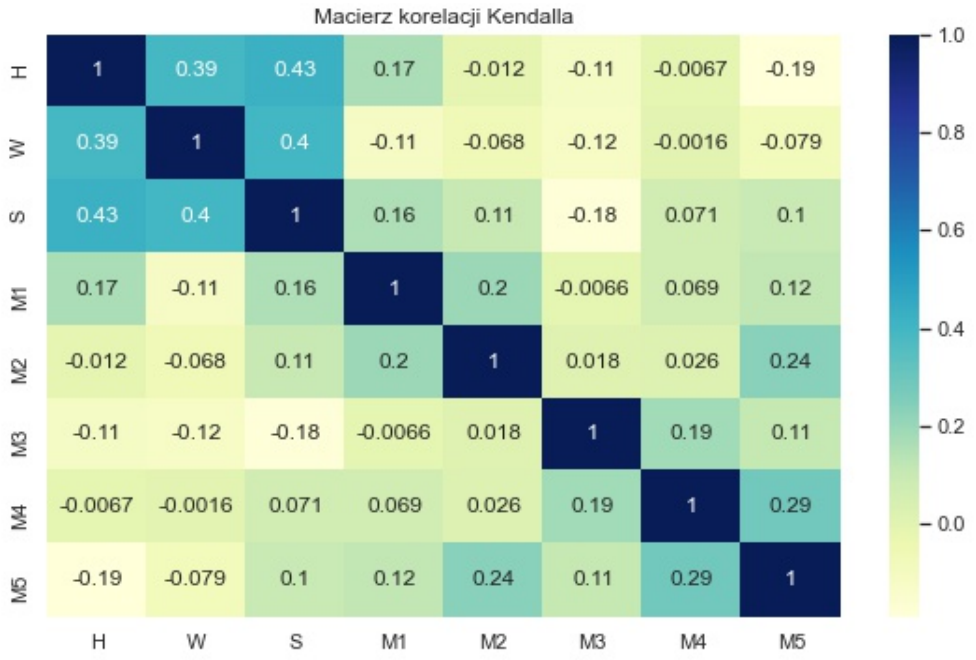
Źródło: opracowanie własne

W tabeli (4.5) przedstawione zostały p–wartości dla przeprowadzonego testu zależności monotonicznej między zmiennymi. Hipoteza zerowa testu mówi o tym, że zmienne są niezależne, z kolei hipoteza alternatywna, że są zależne. Aby zmienne były zależne, oczekuje się, aby p–wartość była mniejsza od poziomu istotności testu $\alpha = 0.05$. W tabeli zaznaczono wartości, dla prób, dla których tak jest. W tych przypadkach odrzucana jest więc hipoteza zerowa o niezależności prób. Są to zbiory, pomiędzy którymi zaobserwowano największą korelację na wykresie z rysunku (4.7). Również tym razem im niższa była wartość korelacji, tym większa p–wartość, świadcząca o tym, że w tych przypadkach nie powinno się odrzucać hipotezy zerowej o niezależności. Zależność monotoniczna obserwowana jest więc pomiędzy parametrami fizycznymi zawodników oraz punktami zdobytymi w drugim i piątym spotkaniu, oraz czwartym i piątym. Dla korelacji Spearmana

p–wartość dla punktów zdobytych w drugim i pierwszym spotkaniu jest znacznie większa niż dla korelacji Pearsona, w tym przypadku jednak nie można już jednoznacznie określić występującej zależności między zmiennymi.

4.2.3 Korelacja Kendalla

Ostatnim etapem badania korelacji występującej w rozważanym zbiorze dotyczącym wyników rozgrywek siatkarskich jest przeanalizowane współczynnika korelacji Kendalla. Jest on kolejną z rangowych metod badania zależności monotonicznej między zmiennymi, która nie musi być liniowa. Współczynnik ten nie posiada założeń odnośnie rozkładu, z którego powinny pochodzić próby.



Rysunek 4.8: Macierz z wartościami współczynnika korelacji Kendalla. *Źródło: opracowanie własne.*

Wykres przedstawiony na rysunku (4.8) to macierz korelacji dla współczynnika Kendalla. Tak jak w poprzednich dwóch przypadkach analizowane będą wartości znajdujące się w dolnej macierzy trójkątnej. Wartości przyjmowane przez ten współczynnik są niższe niż te dla współczynnika Pearsona czy Spearmana. Widać to na pierwszy rzut oka po jaśniejszych kolorach pól macierzy. Najsilniejsza dodatnia zależność monotoniczna obserwowana jest dla parametrów fizycznych siatkarzy, takich jak wzrost, waga i zasięg. Również w tym przypadku nie widać znaczącej zależności między nimi a punktami zdobywanymi w kolejnych meczach. Przyglądając się korelacji między punktami, można zaobserwować zależność między punktami zdobytymi w drugim i piątym spotkaniu oraz czwartym i piątym. Korelacja jest jednak znacznie słabsza niż dla poprzednich współczynników.

W tabeli (4.6) przedstawione zostały p–wartości dla przeprowadzonego testu zależności monotonicznej między zmiennymi. Hipoteza zerowa testu mówi o tym, że zmienne są niezależne, z kolei hipoteza alternatywna, że są zależne. Aby zmienne były zależne, p–wartość powinna być mniejsza od poziomu istotności testu $\alpha = 0.05$. Zaznaczone w tabeli

Tabela 4.6: P-wartości dla współczynnika korelacji Kendalla.

	H	W	S	M1	M2	M3	M4	M5
H	0	0.0012	0.0005	0.1656	0.9234	0.3791	0.9562	0.1179
W	**0.0012**	0	0.001	0.3519	0.5744	0.3374	0.9891	0.5114
S	**0.0005**	**0.001**	0	0.1837	0.3728	0.1348	0.5556	0.3958
M1	0.1656	0.3519	0.1837	0	0.0975	0.9563	0.5654	0.3248
M2	0.9234	0.5744	0.3728	0.0975	0	0.8801	0.8266	0.0473
M3	0.3791	0.3374	0.1348	0.9563	0.8801	0	0.1215	0.3662
M4	0.9562	0.9891	0.5556	0.5654	0.8266	0.1215	0	0.0167
M5	0.1179	0.5114	0.3958	0.3248	**0.0473**	0.3662	**0.0167**	0

Źródło: opracowanie własne

wartości, to te, dla których odrzucana jest więc hipoteza zerowa. Są to próby, dla których zaobserwowano najwyższe współczynniki korelacji przedstawione w macierzy z rysunku (4.8). Im niższa była wartość korelacji, tym większa p–wartość, czyli większe prawdopodobieństwo, że w tych przypadkach nie powinno się odrzucać hipotezy zerowej o niezależności. Zależność monotoniczna jest więc istotna statystycznie pomiędzy parametrami fizycznymi zawodników oraz punktami zdobytymi w drugim i piątym spotkaniu, oraz czwartym i piątym.

4.3 Model regresji liniowej dla dwóch zmiennych

Po przeprowadzeniu analizy korelacji widać, że w analizowanym zbiorze danych występują pewne zależności. Skoro tak jest, możliwe będzie dopasowanie modelu, który pozwoli na przewidywanie wartości jednej zmiennej (zależnej) za pomocą drugiej (niezależnej). Wyniki uzyskane w macierzy korelacji Pearsona (4.6) pokazują, że między danymi występuje zależność liniowa. Dlatego na początku spróbowano dopasować model regresji liniowej, a dokładniej jego najprostszą postać, czyli model z jedną zmienną objaśniającą.

4.3.1 Dopasowanie modelu

Jeśli chodzi o stosowanie modelowania danych w sporcie, przeważnie za cel stawia się przewidywanie wyników wydarzeń sportowych. Z analizy korelacji wyszło, że parametry fizyczne nie mają znaczącego wpływu na liczbę zdobywanych przez zawodników punktów. Z kolei najwyższy współczynnik korelacji Pearsona, a więc najsilniejszą zależność liniową zaobserwowano między punktami zdobytymi w czwartym i piątym spotkaniu. W związku, z czym to te zmienne zostaną wykorzystane do dopasowania modelu regresji liniowej dla dwóch zmiennych.

Aby predykowanie wartości za pomocą dopasowanego modelu miało sens, przyjęto, że:

- Zmienna niezależna (X) – punkty zdobyte przez graczy w czwartym spotkaniu.
- Zmienna zależna (Y) – punkty zdobyte przez graczy w piątym spotkaniu.

Rozważana próba ma postać $(x_1, y_1), \ldots, (x_n, y_n)$, gdzie $x_1, \ldots, x_n$ to wartości przyjmowane przez zmienną X, a gdzie $y_1, \ldots, y_n$ to wartości przyjmowane przez zmienną Y, oraz

n to liczność próby.

Rysunek 4.9: Wykres rozproszenia dla próby. *Źródło: opracowanie własne.*

Na wykresie z rysunku (4.9) przedstawiono wykres rozproszenia dla wspomnianej próby. Na osi OX zaznaczono wartości zmiennej X, czyli punkty zdobyte w czwartym spotkaniu, natomiast na osi OY wartości zmiennej Y, czyli punkty zdobyte w piątym spotkaniu. Kropki odpowiadają kolejnym obserwacjom próby. Patrząc na zachowanie wartości zaznaczonych na wykresie, można zauważyć niewielką dodatnią zależność liniową. Widać, że zawodnicy, którzy zdobywali dużo punktów w czwartym spotkaniu, przeważnie punktowali równie dobrze w piątym.

Do danych zaznaczonych na wykresie dopasowano prostą regresji. Wyestymowano współczynniki modelu z wykorzystaniem metody najmniejszych kwadratów.

Tabela 4.7: Podsumowanie estymacji współczynników.

współczynnik	wartość	błąd std	p–wartość
$\hat{\beta}_0$	6.1397	3.048	0.052
$\hat{\beta}_1$	0.4959	0.203	0.020

Źródło: opracowanie własne

W tabeli (4.7) zamieszczono wyniki estymacji. W kolumnie wartość znajdują się wartości estymatorów. Błąd std. to błąd standardowy określający jak bardzo dany parametr może się zmieniać w różnych badaniach tego samego zjawiska. Widać, że dla współczynnika kierunkowego prostej, czyli $\hat{\beta}_1$ jest on znacznie mniejszy niż dla stałej. Ostatnia kolumna zawiera p-wartości odpowiadające dwóm problemom testowania:

- $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$,
- $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$,

Dla każdego z tych testów hipoteza zerowa jest odrzucana, gdy p-wartość jest mniejsza bądź równa α. Odrzucenie H_0, czyli hipotezy zerowej dla drugiego z tych testów oznacza, że model regresji liniowej ma sens, bo zmienna objaśniająca ma liniowy wpływ na zmienną objaśnianą.

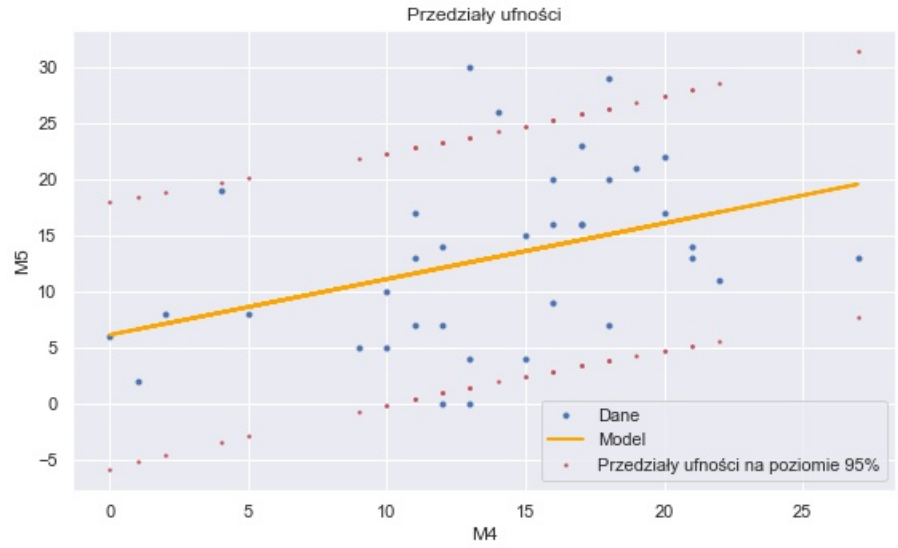
Podstawiając wyznaczone współczynniki do wzoru na prostą regresji, otrzymano:

$$\hat{Y}_i = 0.4959\, x_i + 6.1397 \tag{4.1}$$

Rysunek 4.10: Wykres rozproszenia z dopasowaną prostą regresji. *Źródło: opracowanie własne.*

Rysunek (4.9) przedstawia wykres rozproszenia dla danych z dopasowaną do nich prostą regresji opisaną wzorem (4.1). Widać, że prosta regresji jest funkcją rosnącą, ponieważ otrzymany współczynnik $\hat{\beta}_1$ jest dodatni. Model oddaje więc rosnący charakter danych. Kilka obserwacji znajduję się stosunkowo daleko od wyznaczonej prostej. Możliwe, że są to obserwacje odstające. Jednak ze względu na małą ilość obserwacji w próbie nie będą one usuwane. Kolejnym etapem było wyznaczenie przedziałów ufności predykcji. Predykowane wartości, to te, które zostaną wyznaczone za pomocą modelu (4.1). Przedziały ufności wyznacz się zgodnie ze wzorem (2.1). Wartości rzeczywiste powinny mieścić się w zadanym przedziale z prawdopodobieństwem 0.95.

Wykres z rysunku (4.11) przedstawia wykres rozproszenia dla danych z dopasowaną prostą regresji i wspomnianymi przedziałami ufności na poziomie 95%. Widać, że większość wartości rzeczywistych mieści się w zadanym przedziale ufności. Poza nim znajdują się jedynie pojedyncze obserwacje, które jeszcze przed wyznaczeniem przedziałów wyglądały na odstające.



Rysunek 4.11: Wykres rozproszenia z dopasowaną prostą regresji i zaznaczonymi przedziałami ufności predykcji. *Źródło: opracowanie własne.*

Dopasowując model liniowy ważne, jest, aby dane rzeczywiście miały charakter liniowy. Aby to sprawdzić, można skorzystać ze wskaźników, które oceniają zależność liniową.

Tabela 4.8: Wskaźniki poprawności dopasowania modelu.

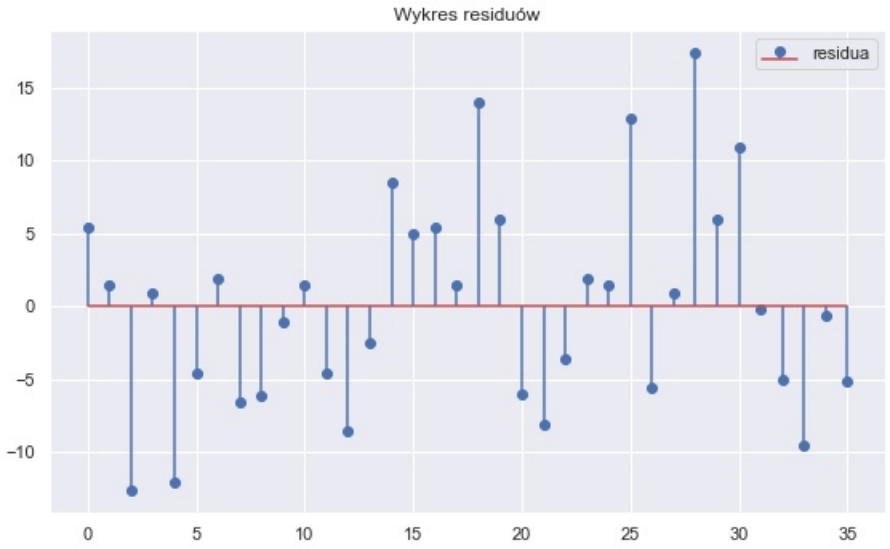
Wskaźnik	Wartość
SSE	320.74
SSR	1822.23
SST	2142.97
r_{XY}	0.39
R^2	0.150
Statystyka F	5.985
p-wartość (F)	0.0198

Źródło: opracowanie własne

W tabeli (4.8) przedstawiono wartości wskaźników oceniających zależność liniową. Zawarty w tabeli został współczynnik korelacji Pearsona i dla tej próby był on najwyższy ze wszystkich analizowanych prób. Współczynnik determinacji R^2 mówiący o tym jak dobrze model dopasowuje się do danych, jest jednak dość niski, wynosi jedynie 0.150. Wartość ta pokazuje procentowo jak duża część zmiennej zależnej, jest wyjaśniana przez użytą w modelu zmienną niezależną. Dla rozważanej próby jest to 15%. W tabeli znajduje się również wartość statystyki F-testu oraz jego p-wartość. Test ten za hipotezę zerową przyjmuje, że wszystkie współczynniki modelu przyjmują wartość 0. Hipoteza alternatywna zaś, że istnieje chociaż jeden niezerowy współczynnik. Standardowo za poziom istotności przyjmuje się $\alpha = 0.05$ i jeżeli p–wartość jest mniejsza od założonego poziomu istotności, można odrzucić hipotezę zerową. Dla analizowanego przypadku $0.0198 < 0.05$, a więc można odrzucić hipotezę zerową i wykorzystywać wyznaczony model do estymacji.

4.3.2 Diagnostyka modelu — analiza residuum

Model regresji liniowej posiada pewne założenie odnośnie residuów. Jeśli model został poprawnie dopasowany residua, czyli wartości resztowe będące różnicą wartości rzeczywistej i predykowanej będą pochodziły z rozkładu normalnego ze średnią równą zero, stałą wariancją i będą niezależne od siebie. W celu diagnostyki poprawności zastosowanego modelu sprawdzone zostaną te założenia. Do sprawdzenia założeń zostały wykorzystane testy statystyczne opisane w rozdziale 2.6, a ich wyniki pochodzą z programu MATLAB. Każdy wykonany niżej test zwraca dwie wartości. Pierwsza z nich h przyjmuje tylko dwie wartości 0 lub 1. Jeśli test zwraca 0 jest to równoważne z tym, że powinno się przyjąć hipotezę zerową za prawdziwą. Dla wartości 1, powinno się odrzucić hipotezę zerową. Test zwraca również drugą wartość p, która jest p-wartością wspomnianego testu.



Rysunek 4.12: Wykres przedstawiający residua. *Źródło: opracowanie własne.*

Wykres z rysunku (4.12) przedstawia wartości residuów dla analizowanej próby w zależności od numeru obserwacji. Widać, że przyjmowane są zarówno wartości dodatnie jak i ujemne. Niektóre błędy są niewielkie, nieprzekraczające 5 punktów, inne natomiast sięgają nawet ponad 15 punktów. Wartości dodatnie i ujemne rozkładają się mniej więcej równomiernie. Na wykresie nie widać żadnej tendencji, więc można przyjąć, że założenie o liniowej zależności między zmiennymi jest spełnione.

Analiza średniej. Średnia residuów powinna przyjąć wartość równą 0. Patrząc na wykres (4.12) widać, że wartości residuów rozkładają się chaotycznie po obu stronach osi OX, nie widać, żadnych tendencji ani trendów więc można wnioskować, że założenie to zostało spełnione.

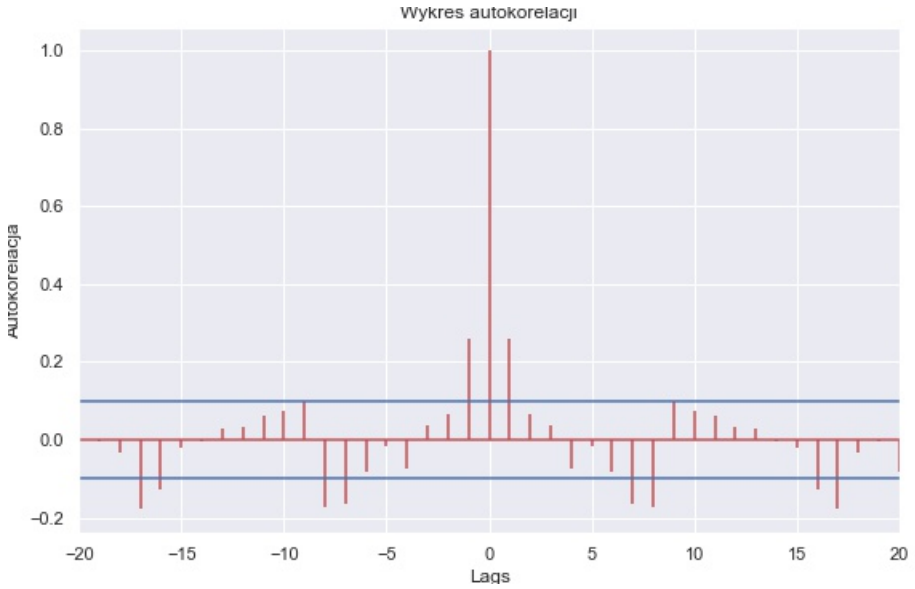
Analiza wariancji. Sprawdzenie założenia o stałej wariancji będzie wykonane przy pomocy Arch testu, który w hipotezie zerowej zakłada, brak heteroskedastyczności, czyli tego, że przynajmniej jedna wartość będzie różniła się wariancją od pozostałych. Jeśli hipoteza zerowa zostanie przyjęta, będzie można wnioskować, że wariancja wartości resztowych jest

stała, w przypadku przyjęcia hipotezy alternatywnej, głoszącej, że w wektorze znajduje się przynajmniej jedna wartość, dla której wariancja odstaje od pozostałych, należy przyjąć brak stałej wariancji. Wyniki Arch testu:

$$h = 0 \qquad p = 0.0879$$

Również w tym przypadku otrzymano wartość $h = 0$ co świadczy o tym, że nie ma podstaw do odrzucenia hipotezy zerowej. Z kolei trzymana p–wartość testu będąca prawdopodobieństwem, z jakim przyjmujemy hipotezę zerową, jest większa od poziomu istotności α, ale nieznacznie. Nie można więc odrzucić stwierdzenia, że wariancja residuów jest stała.

Niezależność. Otrzymane residua zgodnie z teorią powinny być od siebie niezależne. Pierwszym sprawdzeniem może być spojrzenie na wykres (4.12) przedstawiający wykres residuów. Dla wartości niezależnych chmura powinna być równomiernie rozłożona i tak rzeczywiście jest. Jednak przeprowadzając dokładniejszą analizę można narysować wykres funkcji autokorelacji dla residuów. Jeśli wartości będą bliskie 0 dla całej próby, i tylko w zerze otrzymana zostanie wartość 1, można mówić o niezależności. Formalnym sposobem sprawdzenia będzie wykonanie testu Ljung'a-Box'a, który używany jest do oceny zależności między danymi. Za hipotezę zerową test ten przyjmuje, że korelacja między obserwacjami równa jest 0. Hipoteza alternatywna natomiast mówi, że są one zależne.



Rysunek 4.13: Wykres przedstawiający funkcję autokorelacji dla residuów. *Źródło: opracowanie własne.*

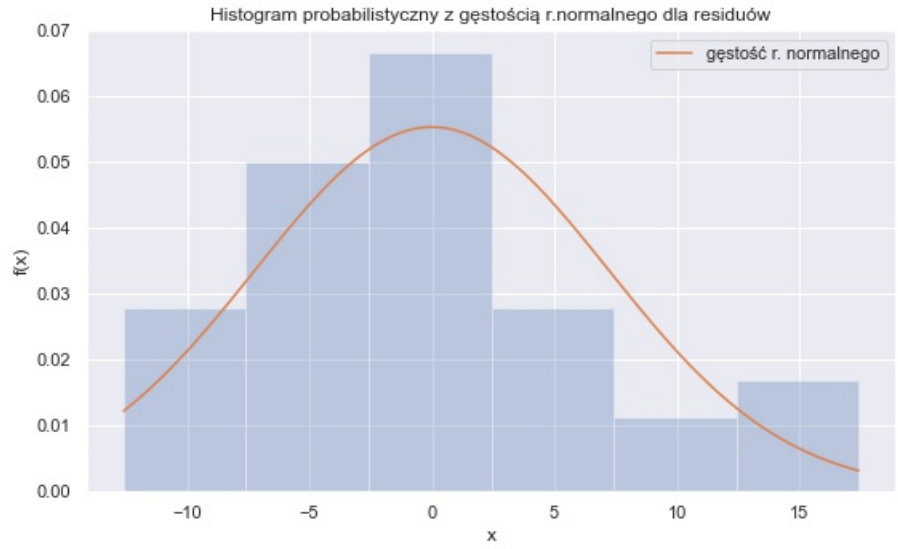
Wartości residuów powinny oscylować w okolicach zera, pomiędzy niebieskimi liniami oznaczającymi przedziały ufności. Na wykresie (4.13) widać, że w pewnych miejscach wartości przekraczają niebieskie przedziały, jest to jednak nieznaczne przekroczenie. W celu formalnego sprawdzenia należy zobaczyć jaki wynik zwraca lbqtest. Wyniki testu:

$$h = 0 \qquad p = 0.9335$$

Otrzymany wynik należy interpretować analogicznie jak w przypadku wariancji. Otrzymane $h = 0$,odpowiada stwierdzeniu, że nie ma podstaw do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej. Z kolei wysoka wartość p daje całkowitą pewność tego,

że nie należy odrzucić hipotezy zerowej. Stąd wniosek, że wartości resztowe z modelu są niezależne, czyli spełniają kolejne założenie modelu.

Rozkład. Ostatnim etapem jest sprawdzenie, czy residua pochodzą z rozkładu normalnego, W tym celu porównano gęstości empirycznej z gęstością teoretyczną rozkładu normalnego ze średnią i wariancją wyestymowaną z próby. Wykonano także test Kołomogorowa-Smirnowa i test Anderson-Darling.



Rysunek 4.14: Wykres przedstawiający histogram probabilistyczny z gęstością r.normalnego. *Źródło: opracowanie własne.*

Patrząc na wykres (4.14) widać, że histogram probabilistyczny zachowuje się całkiem podobnie jak gęstość rozkładu normalnego z parametrami średniej i wariancji z próby. Nie jest on idealnie dopasowany, powodem może być jednak bardzo mała liczba obserwacji w próbie. Nie można więc wykluczyć, że dane mają rozkład normalny. W celu dokładnego sprawdzenia Wykonano K-S test, który za hipotezę zerową przyjmuje, że dane pochodzą z rozkładu normalnego o średniej 0 i wariancji 1. Hipoteza alternatywna odrzuca to założenie. Wyniki testu Kołomogorowa-Smirnowa:

$$h = 1 \qquad\qquad p = 5.0348 \cdot 10^{-6}$$

Wartość $h = 1$ świadczy o tym, że hipotezę zerową należy odrzucić. Bardzo mała p–wartość tylko potwierdza to stwierdzenie. Stąd wniosek, że dane nie pochodzą z rozkladu normalnego o średniej 0 i wariancji 1. Kolejnym testem jest test Anderson-Darling, który zakłada, że dane pochodzą z rozkładu normalnego, ze średnią i wariancją wyznaczoną z próby. Wyniki testu Anderson-Darling:

$$h = 0 \qquad\qquad p = 0.3805$$

Otrzymana wartość $h = 0$ pozwala wnioskować, że nie ma podstaw do odrzucenia hipotezy zerowej. Prawdopodobieństwo, że powinno się odrzucić hipotezę zerową wynosi ok 0.38. Jednak test zwraca, że nie ma podstaw do odrzucenia hipotezy o normalności rozkładu. W związku z tym kolejne założenie modelu zostało spełnione.

Analizując wszystkie otrzymane podczas diagnostyki modelu wyniki, można wnioskować, że model został poprawnie dopasowany i zastosowany do badanych danych.

4.4 Model regresji wielokrotnej

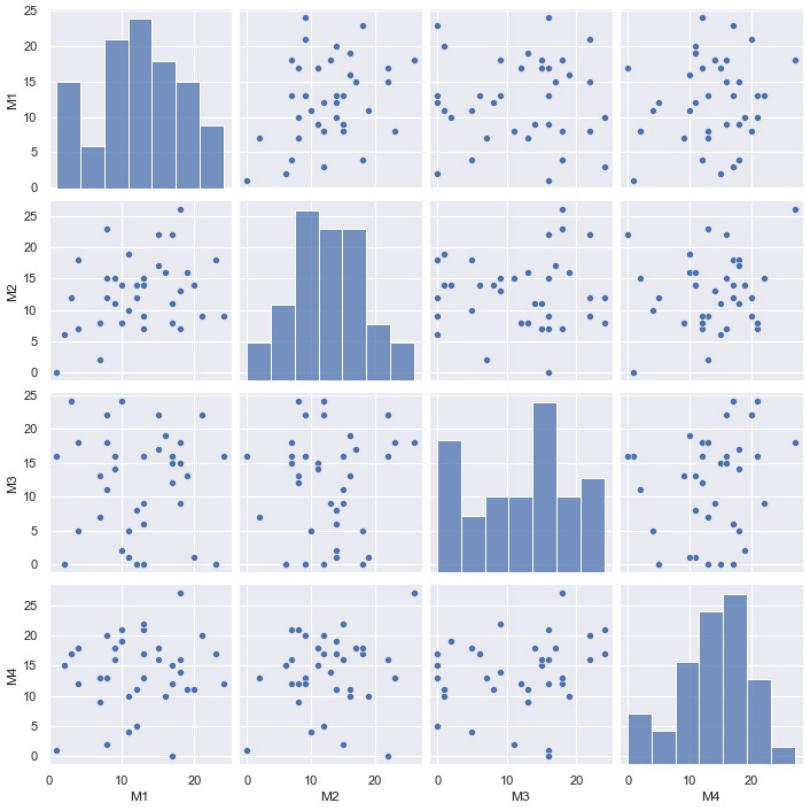
Uzyskane w modelu regresji liniowej dwóch zmiennych R^2 było dość niskie. W celu zwiększenia procentu wyjaśniania zmiennej objaśnianej przez objaśniające można zwiększyć liczbę predykatorów, czyli zmiennych niezależnych. Przed przystąpieniem do dopasowania modelu należy jednak sprawdzić, czy zmienne niezależne, które mają być wykorzystane w modelu, nie są ze sobą skorelowane ani współliniowe. Jeśli by tak było, pojawiłby się problem z odróżnieniem, która zmienna niezależna, w jakim stopniu wpływa na zależną. Za cel predykcji dalej postawiono liczbę punktów zdobytych przez graczy w piątym spotkaniu.

4.4.1 Dopasowanie modelu

Przyjęto więc, że:

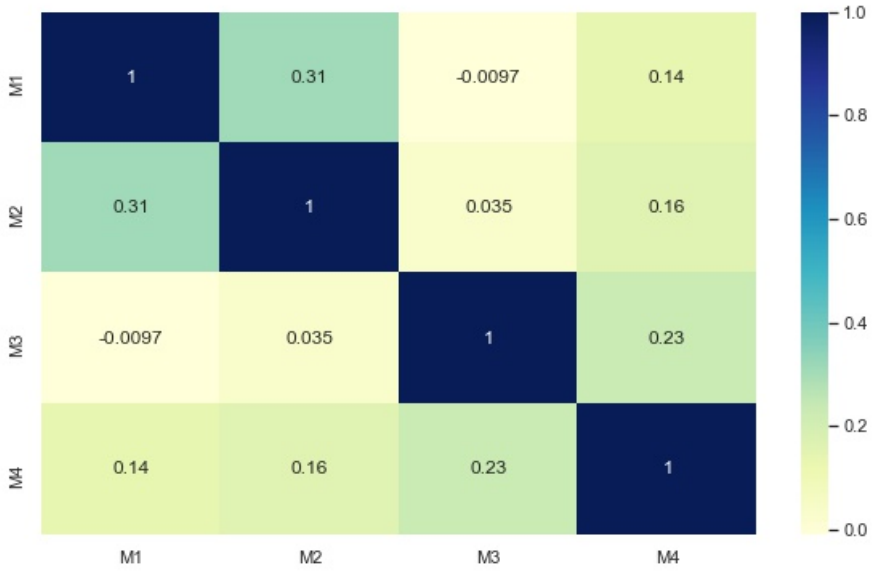
- Zmienna niezależna (X_1) – punkty zdobyte przez graczy w pierwszym spotkaniu,
- Zmienna niezależna (X_2) – punkty zdobyte przez graczy w drugim spotkaniu,
- Zmienna niezależna (X_3) – punkty zdobyte przez graczy w trzecim spotkaniu,
- Zmienna niezależna (X_4) – punkty zdobyte przez graczy w czwartym spotkaniu,
- Zmienna zależna (Y) – punkty zdobyte przez graczy w piątym spotkaniu.

W celu zbadania zależności między zmiennymi, które w modelu mają być niezależne, stworzono macierz wykresów punktowych, aby zobaczyć, jak rozkładają się chmury danych.



Rysunek 4.15: Wykres przedstawiający macierz wykresów punktowych. *Źródło: opracowanie własne.*

Patrząc na wykresy rozproszenia z macierzy wykresów punktowych na rysunku (4.15) na pierwszy rzut oka nie widać żadnych zależności między danymi. Na przekątnej macierzy znajdują się histogramy, które były już analizowane podczas przedstawiania wykorzystanych do analizy danych. Chmury punktów na wykresach rozproszenia we wszystkich przypadkach rozkładają się raczej losowo. Nie obserwuje się żadnych tendencji wzrostowych ani spadkowych. Stąd można wnioskować, że zmienne, które mają być niezależne w modelu, rzeczywiście takie są. Jednak takie nieformalne sprawdzenie nie jest wystarczające. Aby się upewnić, stworzono także macierz korelacji dla zmiennych, które mają być w modelu niezależne.



Rysunek 4.16: Wykres przedstawiający macierz korelacji zmiennych niezależnych. *Źródło: opracowanie własne.*

Wykres z rysunku (4.16) przedstawia wspomnianą wcześniej macierz korelacji. W teorii zakłada się, że zmienne powinny być pozbawione jakiejkolwiek zależności. W praktyce jednak wystarczy, że korelacja pomiędzy poszczególnymi zmiennymi nie będzie przekraczała wartości 0.4. W analizowanym przypadku wszystkie wartości współczynników w macierzy korelacji rzeczywiście są mniejsze od założonej wartości. Jedynie na przekątnej macierzy pojawia się wartość 1, jednak tak jak to było już wspomniane podczas analizy korelacji, wartości na przekątnej to wartości korelacji pomiędzy tą samą zmienną. Można więc stwierdzić, że założenie o niezależności zmiennych zostało spełnione.

Kolejnym krokiem jest sprawdzenie, czy zmienne niezależne w modelu nie są współliniowe. W tym celu wykorzystuje się współczynnik inflacji wariancji (VIF). Wartość współczynnika inflacji wariancji zmiennej niezależnej przedstawia, jak dobrze badana zmienna jest wyjaśniana przez inne zmienne niezależne, zastosowane w modelu [12]. Współczynnik ten przyjmuje wartości większe bądź równe 1.

Interpretacja tego współczynnika jest następująca:

- $VIF = 1$ oznacza brak korelacji między rozważaną zmienną niezależną a innymi zmiennymi niezależnymi,
- $VIF > 10$ wskazuje na wysoką współliniowość między jedną zmienną niezależną a pozostałymi zmiennymi.

Tabela 4.9: Wartości współczynnika inflacji wariancji.

Zmienna	VIF
X_1	5.268057
X_2	5.747671
X_3	3.389732
X_4	5.624077

Źródło: opracowanie własne

W tabeli (4.9) przedstawiono wartości współczynnika inflacji wariancji dla analizowanych zmiennych. Wszystkie wartości są różne od 1 ale także mniejsze od 10, oscylują w okolicy wartości 5. Stąd wniosek, że między zmiennymi nie występuje silna współliniowość i mogą być wykorzystane jako zmienne niezależne w modelu.

Ustalono więc już, że wszystkie zmienne mogą zostać wykorzystane w modelu regresji liniowej wielokrotnej. Podjęto więc próbę dopasowania modelu z wykorzystaniem wszystkich predykatorów.

Tabela 4.10: Podsumowanie estymacji współczynników.

współczynnik	wartość	błąd std.	p–wartość
$\hat{\beta}_0$	1.1854	4.328	0.786
$\hat{\beta}_1$	-0.0179	0.217	0.935
$\hat{\beta}_2$	0.4375	0.224	0.060
$\hat{\beta}_3$	0.0562	0.166	0.737
$\hat{\beta}_4$	0.4180	0.208	0.054

Źródło: opracowanie własne

W tabeli (4.10) znajdują się wyniki estymacji współczynników w modelu. Wartość $\hat{\beta}_0$, tak jak poprzednio jest stałą, czyli wyrazem wolnym. Wartości $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$ odpowiadają wartością współczynników stojących odpowiednio przy zmiennych X_1, X_2, X_3, X_4. W kolumnie wartość znajdują się wartości wyestymowanych współczynników. W kolumnie błąd std. znajdują się wartości błędu standardowego, czyli wartości o ile może się różnić współczynnik w różnych badaniach tego samego zjawiska. Widać, że również tym razem największy błąd osiągnęła stała. Wartości błędów dla pozostałych współczynników są do siebie zbliżone. Ostatnia kolumna zawiera p-wartości, czyli prawdopodobieństwa, że zmienna przy danym współczynniku nie ma wpływu na zmienną objaśnianą. Najmniejsze

p–wartości zostały otrzymane dla punktów zdobytych w drugim i czwartym spotkaniu, a więc to one mają największy wpływ na zmienność zmiennej objaśnianej. Dla punktów zdobytych w pierwszym i trzeci spotkaniu to prawdopodobieństwo jest dość wysokie. Jednak zmienne nie będą usuwane z modelu. Dopiero po sprawdzeniu wyników wskaźników określających poprawność dopasowania modelu, będzie można orzec, czy je zostawić, czy też nie.

Tabela 4.11: Wskaźniki poprawności dopasowania modelu.

Wskaźnik	Wartość
SSE	1672.57
SSR	901.58,
SST	2574.15
R^2	0.251
Statystyka F	2.593
p-wartość (F)	0.0557

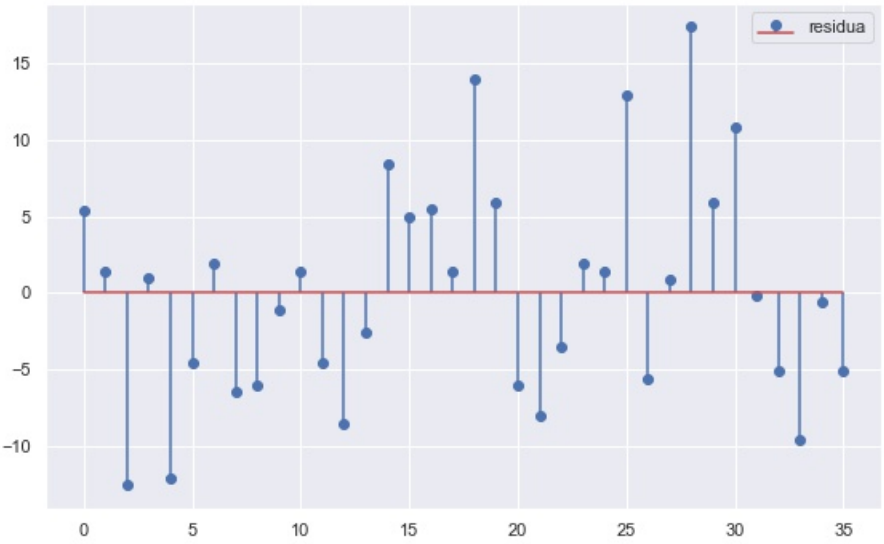
Źródło: opracowanie własne

Tabela (4.11) przedstawia wartości wskaźników, które pomogą ocenić poprawność dopasowanego modelu. W tabeli znajdują się wartości sumy kwadratów błędów, regresyjnej sumy kwadratów, czy całkowitej sumy kwadratów. Współczynnik determinacji mówiący o tym jak dobrze model wyjaśnia zmienną objaśnianą w tym przypadku wynosi $R^2 = 0.251$, a więc model wyjaśnia zmienną w 25.1%. Jednak tak mała wartość tego współczynnika wciąż oznacza, że model nie jest najlepszym rozwiązaniem do predykcji postawionego problemu. Widać jednak, że po dodaniu zmiennych do modelu współczynnik wzrósł w porównaniu do tego co było w przypadku jednej zmiennej niezależnej ($R^2 = 0.150$). W tabeli znajduje się również wartość statystyki F-testu, zakładającego, że wszystkie współczynniki modelu są równe 0, oraz jego p–wartość. Aby móc odrzucić hipotezę zerową oczekuje się, aby p–wartość była mniejsza od $\alpha = 0.05$. W rozważanym przypadku wynosi ona 0.0557, a więc jest tylko nieznacznie większa od zakładanego poziomu istotności. Stąd wniosek, że można zostawić wszytskie zmienne niezależne w modelu i takim modelem próbować predykować wartości zmiennej zależnej, czyli liczbę punktów zdobytych w piątym spotkaniu.

4.4.2 Diagnostyka modelu – analiza residuum

Po dopasowaniu modelu i przeprowadzeniu wstępnej analizy poprawności jego dopasowania należy sprawdzić, czy założenia modelu regresji liniowej również zostały spełnione. Jeśli model został poprawnie dopasowany residua, czyli będą pochodziły z rozkładu normalnego ze średnią równą zero, stałą wariancją i będą niezależne od siebie.

Analiza średniej. Wykres z rysunku (4.17) przedstawia wartości residuów dla analizowanej próby w zależności od numeru obserwacji. Widać, że przyjmowane są zarówno wartości dodatnie jak i ujemne, które rozkładają się mniej więcej równomiernie po obu stronach osi OX. Oscylują one chaotycznie wokół tej osi bez żadnej zauważalnej tendencji. Na wykresie nie widać żadnej tendencji, więc można przyjąć, że założenie o liniowej zależności między zmiennymi jest spełnione. Oraz,że średnia residuów jest bliska 0.



Rysunek 4.17: Wykres przedstawiający residua. *Źródło: opracowanie własne.*

Analiza wariancji. Założenie o stałej wariancji będzie wykonane tak jak poprzednio przy pomocy Arch testu. Test ten w hipotezie zerowej zakłada, że przynajmniej jedna wartość będzie różniła się wariancją od pozostałych. Jeśli hipoteza zerowa zostanie przyjęta, będzie można wnioskować, że wariancja wartości resztowych jest stała. W przypadku jej odrzucenia należy przyjąć brak stałej wariancji. Wyniki Arch testu:

$$h = 0 \qquad p = 0.2488$$

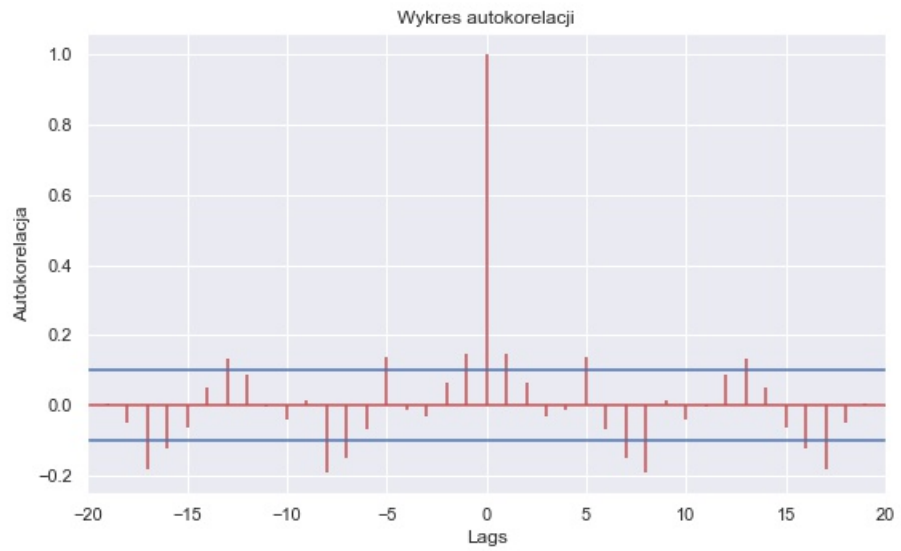
Również w tym przypadku otrzymano wartość $h = 0$ co świadczy o tym, że nie ma podstaw do odrzucenia hipotezy zerowej. Otrzymana p–wartość testu będąca prawdopodobieństwem, z jakim przyjmujemy hipotezę zerową, jest niewielka. Nie można jednak odrzucić stwierdzenia, że wariancja residuów jest stała.

Niezależność. Residua zgodnie z założeniem powinny być od siebie niezależne. W celu sprawdzenia, czy tak jest, można narysować wykres funkcji autokorelacji dla residuów. Jeśli wartości będą bliskie 0 dla całej próby, i tylko w zerze otrzymana zostanie wartość 1, można mówić o niezależności. Można wykonana także test Ljung'a-Box'a, który używany jest do oceny zależności między danymi. Za hipotezę zerową test ten przyjmuje, że korelacja między obserwacjami równa jest 0. Hipoteza alternatywna natomiast mówi, że są one zależne.

Wartości funkcji autokorelacji powinny oscylować w okolicach zera, pomiędzy niebieskimi liniami oznaczającymi przedziały ufności. Na wykresie (4.18) widać, że w pewnych miejscach wartości delikatnie przekraczają niebieskie przedziały. Jednak nie na tyle, aby odrzucić założenie o niezależności residuów. W celu formalnego sprawdzenia należy zobaczyć jaki wynik zwraca lbqtest. Wyniki testu:

$$h = 0 \qquad p = 0.9579$$

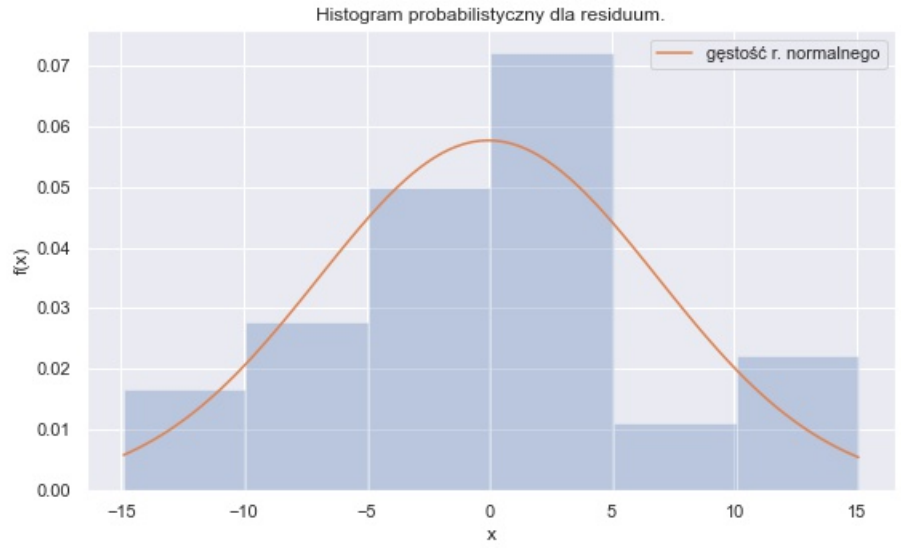
Otrzymany wynik należy interpretować analogicznie jak w poprzednich przypadkach. Otrzymane $h = 0$,odpowiada stwierdzeniu, że nie ma podstaw do odrzucenia hipotezy



Rysunek 4.18: Wykres przedstawiający funkcję autokorelacji dla residuów. *Źródło: opracowanie własne.*

zerowej na rzecz hipotezy alternatywnej. Z kolei wysoka wartość p daje całkowitą pewność tego, że nie należy odrzucić hipotezy zerowej. Stąd wniosek, że wartości resztowe są niezależne, czyli spełniają kolejne założenie modelu.

Rozkład. Na koniec sprawdzono, czy residua pochodzą z rozkładu normalnego, W tym celu porównano histogram probabilistyczny z gęstością teoretyczną rozkładu normalnego ze średnią i wariancją wyestymowaną z próby. Sprawdzenie formalne tak jak dla modelu regresji dwóch zmiennych polegało na wykonaniu testu Kołomogorowa-Smirnowa i testu Anderson-Darling.



Rysunek 4.19: Wykres przedstawiający histogram probabilistyczny z gęstością r.normalnego dla residuów. *Źródło: opracowanie własne.*

Patrząc na wykres (4.19) widać, że histogram probabilistyczny zachowuje się w podobny sposób jak gęstość rozkładu normalnego z parametrami średniej i wariancji z próby. Nie jest on idealnie dopasowany, a powodem braku dobrego dopasowania może być fakt, że próba była bardzo mała. Nie można więc wykluczyć, że dane mają rozkład normalny. W celu dokładniejszego sprawdzenia wykonano kstest, który za hipotezę zerową przyjmuje, że dane pochodzą z rozkładu normalnego o średniej 0 i wariancji 1. Hipoteza alternatywna odrzuca to założenie. Wyniki testu Kołomogorowa-Smirnowa:

$$h = 1 \qquad\qquad p = 1.28 \cdot 10^{-5}$$

Wartość $h = 1$ świadczy o tym, że hipotezę zerową należy odrzucić. Bardzo mała p–wartość tylko potwierdza to stwierdzenie. Stąd wniosek, że dane nie pochodzą z rozkladu normalnego o średniej 0 i wariancji 1. Następnym testem jest test Anderson-Darling, który zakłada bardziej ogólnie, że dane pochodzą z rozkładu normalnego, ze średnią i wariancją wyznaczoną z próby. Wyniki testu Anderson-Darling:

$$h = 0 \qquad\qquad p = 0.7222$$

Otrzymana wartość $h = 0$ pozwala wnioskować, że nie ma podstaw do odrzucenia hipotezy zerowej. Prawdopodobieństwo, że nie powinno się odrzucać hipotezy zerowej, jest również dość wysokie. A więc kolejne założenie modelu zostało spełnione.

Patrząc na wszystkie wyniki otrzymane podczas diagnostyki modelu, można stwierdzić, że model został dopasowany poprawnie i spełnia wszystkie założenia. Pomimo uzyskania niewielkiego wyjaśniania zmiennej objaśnianej przez model, bo tylko 25%, model można stosować.

4.5 Regresja logistyczna

Ostatnim zastosowanym do analizy danych sportowych z siatkówki modelem, jest model regresji logistycznej. Model ten znajduje zastosowanie w sytuacji gdy jedna ze zmiennych przyjmuje tylko dwie wartości, a dokładniej 0 lub 1. W tym celu zmodyfikowano lekko problem, który będzie celem predykcji. W tej części pracy celem modelowania będzie, ocena czy zawodnik dostanie od trenera szansę, aby zagrać w kolejnym spotkaniu, znając liczbę zdobytych przez niego punktów w poprzednim.

Tabela 4.12: Dane wykorzystane w modelu regresji logistycznej.

L.p	M2	P
0	14	1
1	18	1
2	9	0
3	9	1
4	7	0
5	0	1

Źródło: opracowanie własne

W tabeli (4.12) znajduje się fragment danych, które zostaną wykorzystane do dopasowania modelu. W pierwszej kolumnie znajduje się liczba porządkowa gracza, ta sama,

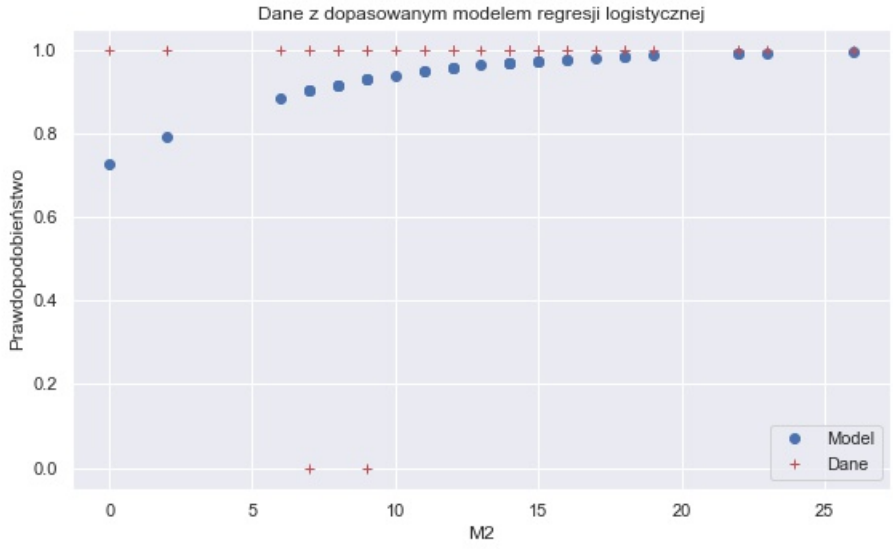
która została przypisana do niego na początku, w tabeli (4.1). W drugiej kolumnie $M2$ znajdują się wartości punktów zdobytych w drugim spotkaniu, i to właśnie one będą zmienną niezależną (X) w modelu. Ostatnia kolumna P, zawiera informację czy zawodnik rozegrał kolejne spotkanie. Wartość 0 przyporządkowywana jest w sytuacji kiedy siatkarz nie dostał szansy rozegrać kolejnego spotkania, a wartość 1, kiedy tę szansę otrzymał. Do przedstawionych w tabeli danych dopasowano model i wyestymowano jego współczynniki.

Tabela 4.13: Podsumowanie estymacji współczynników.

współczynnik	wartość	błąd std	p–wartość
$\hat{\alpha}$	0.9898	1.461	0.498
$\hat{\beta}$	0.1765	0.149	0.235

Źródło: opracowanie własne

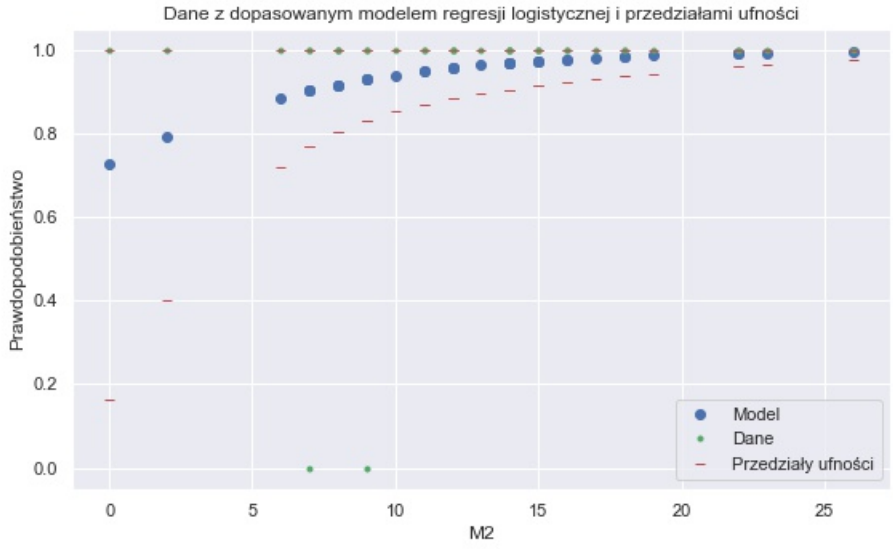
Tabela (4.13) zawiera wyniki estymacji współczynników. Jak w poprzednich przypadkach pierwsza kolumna zawiera wartość współczynnika. Współczynnik α odpowiada stałej znajdującej się w modelu. Współczynnik β odpowiada wartości stojącej przy zmiennej X, będącej liczbą punktów zdobytych w drugim spotkaniu. W kolumnie błąd std. znajdują się wartości błędów standardowych. Widać, że i w tym przypadku dla stałej błąd jest większy. Patrząc na wyniki p-wartości, które mówią, jakie jest prawdopodobieństwo, że dana zmienna nie ma wpływu na zmienność zmiennej zależnej, otrzymano niezbyt satysfakcjonujące wyniki. Ponieważ p-wartości dla obu współczynników znacząco przekraczają poziom istotności 5%.Wyniki te nie wykluczają jednak tego, że zmienne mają wpływ na zmienność zmiennej objaśnianej. Aby to ostatecznie rozstrzygnąć, należy zbadać wskaźniki poprawności dopasowania modelu.



Rysunek 4.20: Wykres przedstawiający dane z dopasowanym modelem regresji logistycznej. *Źródło: opracowanie własne.*

Na rysunku (4.20) przedstawiono dane z tabeli (4.12) oraz dopasowany model regresji logistycznej ze współczynnikami pochodzącymi z tabeli (4.13). Widać, że zgodnie z modelem wraz ze wzrostem liczby zdobytych punktów, wzrasta prawdopodobieństwo

wystąpienia siatkarza w kolejnym spotkaniu. Jest to wniosek, który można było zakładać na początku, przed przystąpieniem do modelowania. Kolejnym etapem było wyznaczenie przedziałów ufności.



Rysunek 4.21: Wykres przedstawiający dane z dopasowanym modelem regresji logistycznej i przedziałami ufności. *Źródło: opracowanie własne.*

Na rysunku (4.21) oprócz informacji zawartych na poprzednim wykresie zaznaczone zostały przedziały ufności. Widać, że większość obserwacji, zaznaczonych zielonymi kropkami mieści się w wyznaczonych przedziałach. Co może świadczyć o poprawności dopasowania modelu.

Ostatnim etapem było zweryfikowanie poprawności dopasowania modelu, z wykorzystaniem odpowiednich wskaźników.

Tabela 4.14: Wskaźniki poprawności dopasowania modelu.

Wskaźnik	Wartość
R^2	0.06
pseudo R^2	0.1033
LLR p-wartość	0.2065

Źródło: opracowanie własne

Tabela (4.14) zawiera wartości poszczególnych wskaźników poprawności dopasowania modelu. Wartość R^2 mówiąca o tym jak dobrze model wyjaśnia zmienną objaśnianą, jest niewielka. Wynosi zaledwie 6%. Kolejnym wskaźnikiem, częściej wykorzystywanym przy modelu regresji logistycznej jest pseudo R^2. Jego wartość jest większa niż jego poprzednika, jednak dalej wynosi nieco ponad 10%. Ostatnim wskaźnikiem jest LLR p-wartość, jest to prawdopodobieństwo, z jakim model popełnia błędy. Zakłada się więc, że ta wartość nie powinna przekraczać poziomu istotności $\alpha = 0.05$. W rozważanym przypadku prawdopodobieństwo to przekracza 20%. Otrzymane wyniki świadczą o tym, że model regresji logistycznej nie jest najlepszym modelem do przewidywania tego rodzaju

zależności i nie powinno się go stosować do predykowania wystąpienia zawodnika w kolejnym spotkaniu znając liczbę punktów zdobytych przez niego w poprzednim meczu.

4.6 Podsumowanie otrzymanych wyników własnych

Podsumowując, do analizy wykorzystano dane dotyczące parametrów fizycznych zawodników oraz liczby punktów zdobytych przez nich podczas rozgrywek fazy grupowej na Igrzyskach Olimpijskich 2020. W pierwszej części wyników własnych przedstawiono dane i zbadano ich charakterystyki oraz rozkład. Po zapoznaniu się z danymi wyznaczono macierze korelacji, w celu zbadania czy występują między danymi jakieś zależności. Dla wszystkich analizowanych współczynników korelacji tj. Pearsona, Spearmana i Kendalla wyszło, że parametry fizyczne zawodników nie mają znaczącego wpływu na to jak punktują zawodnicy. Stało się tak pewnie dlatego, że w analizowanym zbiorze znajdowali się tylko profesjonalni sportowcy, których parametry fizyczne były do siebie zbliżone. Dla wszystkich współczynników zaobserwowano także istnienie zależności pomiędzy parametrami fizycznymi zawodników, czyli wzrostem, wagą i zasięgiem ataku. Okazało się, że istnieje również związek pomiędzy liczbą punktów zdobytych w drugim i piątym spotkaniu oraz pomiędzy punktami w czwartym i piątym spotkaniu. Wszystkie trzy współczynniki wskazywały na istnienie zależności między tymi samymi grupami, różniły się wartościami współczynników i rodzajem badanej zależności.

Kolejnym etapem pracy było dopasowanie modelu regresji liniowej dla dwóch zmiennych. Wybrano liczbę punktów zdobytych przez siatkarzy w czwartym i piątym meczu, ze względu na uzyskanie najwyższego współczynnika korelacji Pearsona, czyli zależności liniowej. Dopasowano do danych odpowiedni model i zbadano poprawność zastosowanego modelu. Mimo że procent wyjaśniania zmiennej zależnej poprzez model był niewielki, bo wynosił ok. 15%, wszystkie wskaźniki poprawności dopasowania modelu, oraz założenia odnośnie residuów zostały spełnione, a więc model można w tym przypadku stosować.

W celu zwiększenia procentu wyjaśniania zmiennej zależnej przez model, w kolejnej części pracy do modelu regresji liniowej dodano więcej predykatorów, czyli zmiennych niezależnych. Tymi zmiennymi były wyniki wszystkich czterech pierwszych spotkań. A celem predykcji wynik uzyskany w piątym meczu. Po tym zabiegu procent ten osiągnął ponad 25%, a więc znacząco wzrósł. Również w tym przypadku wszystkie założenia odnośnie poprawności zastosowania modelu zostały spełnione. Zmienne niezależne nie były współliniowe oraz skorelowane, a residua okazały się być niezależne, mieć rozkład normalny, średnią równą zero i stałą wariancję. Jednak dalej otrzymana poprawność dopasowania modelu nie jest satysfakcjonująca.

Ostatnim etapem pracy było sprawdzenie, czy liczba punktów zdobytych przez zawodnika ma wpływ na to, czy dostanie on szansę rozegrania kolejnego spotkania, przy pomocy modelu regresji logistycznej. Dopasowano model i wyznaczono jego współczynniki oraz przedziały ufności. Jednak w tym przypadku podczas sprawdzenia poprawności dopasowania modelu nie uzyskano satysfakcjonujących wyników. W związku z tym nie jest to model odpowiedni do postawionego problemu. Patrząc na wnioski końcowe, można stwierdzić, że wszystkie postawione we wstępie pracy cele zostały spełnione.

Bibliografia

[1] Agresti, A. *An Introduction to Categorical Data Analysis*, 2 ed. Wiley, 2007.

[2] Agresti, A. *Categorical Data Analysis*, 3 ed. Wiley, 2013.

[3] Gewert, M., Skoczylas, Z. *Analiza matematyczna 1. Definicje, twierdzenia, wzory*, 9 ed. Oficyna Wydawnicza GiS, 2001.

[4] Hilbe, J. M. *Logistic Regression Models*, 1 ed. CRC Press, 2009.

[5] Hilbe, J. M. *Practical Guide to Logistic Regression*, 1 ed. CRC Press, 2015.

[6] Hollander, M., Wolfe, D. A., Chicken, E. *Nonparametric Statistical Methods*, 3 ed. Whiley, 2014.

[7] Kanji, G. K. *100 Statistical Tests*, 3 ed. SAGE Publications Inc, 1993.

[8] Koronacki, J., Mielniczuk, J. *Statystyka dla kierunków technicznych i przyrodniczych*, 1 ed. Wydawnictwa Naukowo–Techniczne, 2018.

[9] Krężołek, D. Miary zależności – analiza statystyczna na przykładzie wybranych walorów rynku metali niezależnych. *Uniwersytet Ekonomiczny w Katowicach* (2014).

[10] Kuan, C. M. Lecture on time series diagnostic tests. *Institute of Economics Academia Sinica* (2003). `https://homepage.ntu.edu.tw/~ckuan/pdf/Lec-DiagTest.pdf`.

[11] Neill, J. Pearson correlation. *Wikipedia – Wolna Encyklopedia* (Online, dostęp 18.10.2021). `https://en.wikipedia.org/wiki/File:Correlation_examples2.svg`.

[12] Robinson, C., Schumacker, R. E. Interaction effects: Centering, variance inflation factor and interpretation issues. *University of Alabama* (2003). `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.620.5853&rep=rep1&type=pdf`.

[13] Rosa, A. Analiza korelacji i regresji. *Uniwersytet Łódzki* (Online, dostęp 20.10.2021). `http://www.demografia.uni.lodz.pl/dlastud/korelacja_i_regresja.pdf`.

[14] Vittinghoff, E., C.Shiboski, S., V.Glidden, D., E.McCulloch, C. *Regression Methods in Biostatistics Linear, Logistic, Survival and Repeated Measures Model*, 1 ed. Springer, 2011.

[15] Weisberg, S. *Applied Linear Regression*, 3 ed. Wiley, 2005.

[16] Wyłomańska, A. Wykład z komputerowej analizy szeregów czasowych. *Politechnika Wrocławska* (2020).

[17] Yan, X., Su, X. G. *Linear Regression Analysis: Theory and Computing*, 1 ed. World Scientific Publishing, 2009.