# Komputerowa analiza szeregów czasowych - Sprawozdanie 1

Martyna Świeściak - 243014 Katarzyna Rybarczyk - 243080

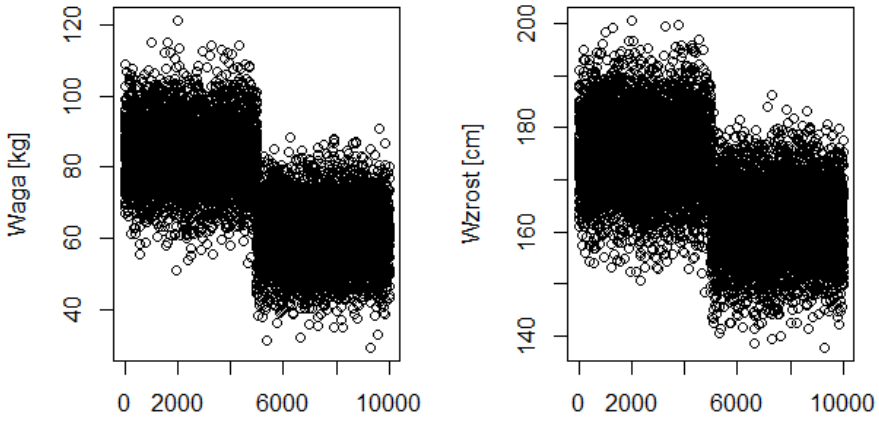
17.12.2019

## 1 Wstęp

Tematem sprawozdania jest zastosowanie modelu regresji liniowej do sprawdzenia zależności na wybranym zbiorze danych. Przedmiotem naszych badań będzie zależność między wagą a wzrostem, o której zakładamy, że jest liniowa.

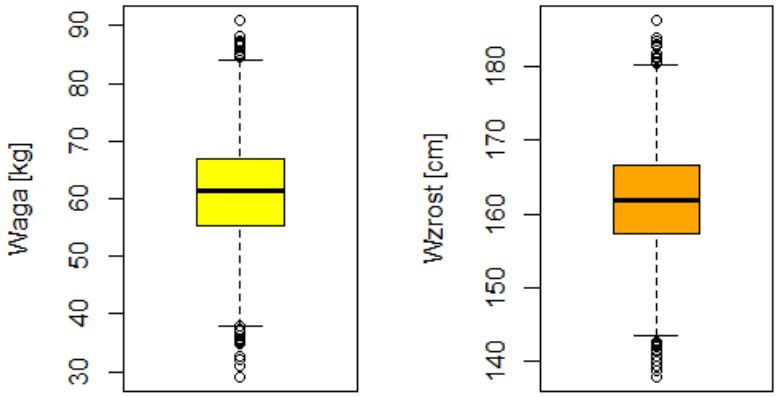
### 1.1 Dane

Dane użyte w zadaniu pochodzą z portalu `www.kaggle.com` i zawierają 10000 obserwacji – 5000 dla kobiet i 5000 dla mężczyzn. Jakość danych jest wysoka – nie istnieją puste obserwacje. Zbiory obserwacji są zauważalnie różne z uwagi na płeć (Rysunek 1), więc skupiłyśmy się wyłącznie na danych dotyczących kobiet. Waga została pobrana w funtach, więc pierwszym krokiem było przeliczenie jej na kilogramy (przyjęte zostało 1lbs = 0,45kg). Analogicznie został przeliczony wzrost z cali na centymetry (1” = 2,54cm).




Rysunek 1: Wykresy badanych zbiorów obserwacji wagi i wzrostu u mężczyzn (lewe strony wykresów) i kobiet (prawe strony wykresów).

W danych występuje dość dużo wartości odstających (Rysunek 2). Zdecydowałyśmy się zostawić je w naszej próbce, ponieważ uważamy, że dobrze obrazują rzeczywistą sytuację społeczeństwa (wiemy, że istnieją osoby bardzo niskie i bardzo wysokie, a także bardzo szczupłe i otyłe). Dodatkowo, w tym wypadku błędy pomiarowe nie miałyby znaczącego wpływu na jakość i rozrzut rozpatrywanych obserwacji.




Rysunek 2: Wykresy typu boxplot dla obserwacji dotyczących wagi oraz wzrostu kobiet

### 1.2 Teoretyczny model regresji liniowej

Główną ideą regresji liniowej jest przewidywanie jaką wartość przyjmie dana zmienna, gdy będziemy znali wartość innej zmiennej – w naszym wypadku chcemy oszacować wzrost kobiety na postawie jej wagi. Skorzystamy w tym celu z klasycznego modelu regresji liniowej o postaci

$$Y_i = \beta_1 x_i + \beta_0 + \xi_i, \tag{1}$$

gdzie:

- $Y_i$ - niezależne zmienne losowe
- $\beta_0, \beta_1$ - współczynniki modelu regresji,
- $x_i$ - deterministyczne obserwacje,
- $\{\xi_i\}_{i=1}^n$ - ciąg niezależnych zmiennych losowych o średniej równej zero i stałej wariancji równej $\sigma^2$, tzn. biały szum.

W przypadku klasycznego modelu zakłada się, że $\xi_i$ są z rozkładu normalnego. Wtedy:

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \xi_i) = \beta_0 + \beta_1 x_i, \tag{2}$$

$$Var(Y_i) = Var(\beta_0 + \beta_1 x_i + \xi_i) = Var(\xi_i) = \sigma^2. \tag{3}$$

## 2 Konstrukcja modelu

### 2.1 Podział danych

Chcąc nie tylko stworzyć model, ale także sprawdzić jego dokładność, zestaw danych został podzielony na dwie próbki równej długości:

- testową,
- użytą do stworzenia modelu.

Podziału dokonałyśmy na danych posortowanych ze względu na wagę – do próbki testowej wzięłyśmy co drugą obserwację. Zabieg ten miał na celu symulację sytuacji, gdzie połowę danych posiadamy do stworzenia modelu, a druga połowa to nowe obserwacje, dla których chcemy sprawdzić dopasowanie. Jednocześnie próbki powinny być dość podobne.

## 2.2 Statystyki liczbowe

Następnie na danych sprawdzone zostały statystyki - miary położenia, rozproszenia, asymetrii i spłaszczenia. Wyliczenia tych statystyk zostały przeprowadzone na całym zbiorze obserwacji, jak również tylko na próbce, która została użyta do stworzenia modelu (m. in. w celu sprawdzenia, czy podział jest sensowny i nie zawiera np. tylko obserwacji odstających).

| | Liczba obserwacji | Średnia | Wariancja | Odchylenie standardowe | Współczynnik zmienności | Min | Max | Skośność | Kurtoza |
|---|---|---|---|---|---|---|---|---|---|
| Wzrost | 5000 | 161,82 | 46,90 | 6,85 | 4,23% | 137,83 | 186,41 | -0,02 | 2,94 |
| Waga | 5000 | 61,14 | 73,28 | 8,56 | 14,00% | 29,12 | 91,01 | -0,01 | 2,93 |

Rysunek 3: Statystyki liczbowe dla całego zbioru obserwacji.

Obserwacje na podstawie statystyk:

- rozkłady obu zmiennych są nieznacznie lewostronnie skośne – jednak bardziej symetryczny jest rozkład wagi (skośność najbliższa 0),
- rozkłady obu zmiennych są delikatnie platokurtyczne (kurtoza $< 3$),
- rozkłady obu zmiennych bliskie są rozkładowi normalnemu,
- pomiary wzrostu cechują się niewielkim rozproszeniem (mały współczynnik zmienności).

Dla porównania, poniżej znajduje się tabelka z wartościami statystyk liczbowych próbki użytej do stworzenia modelu. Statystyki nie różnią się znacząco od oryginalnej próbki, co oznacza, że model nie powinien być zaburzony przez wybrany podział próbki.

| | Liczba obserwacji | Średnia | Wariancja | Odchylenie standardowe | Współczynnik zmienności | Min | Max | Skośność | Kurtoza |
|---|---|---|---|---|---|---|---|---|---|
| Wzrost | 2500 | 161,81 | 48,17 | 6,94 | 4,29% | 140,08 | 186,41 | 0,03 | 2,92 |
| Waga | 2500 | 61,14 | 73,29 | 8,56 | 14,00% | 31,04 | 91,01 | 0,00 | 2,93 |

Rysunek 4: Statystyki liczbowe dla modelowego zbioru danych.

## 2.3 Wyznaczenie estymatorów

W celu stworzenia modelu na podstawie naszych danych, musimy znaleźć estymatory $\hat{\beta}_0$ i $\hat{\beta}_1$ wyznaczone metodą najmniejszych kwadratów (lub metodą największej wiarygodności), wyrażone wzorami

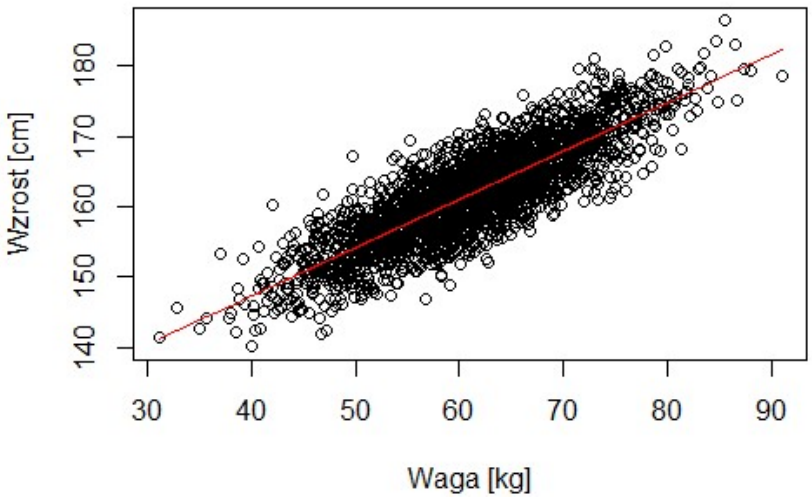
$$\hat{\beta}_0 = \beta_1 + \frac{\sum_{i=1}^{n} \xi_i (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \tag{4}$$

gdzie $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ i $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Użyta została do tego wbudowana w RStudio funkcja *lm*. Otrzymane wyniki to $\hat{\beta}_0 = 119.84$ oraz $\hat{\beta}_1 = 0.69$.

# 3 Analiza

Wywołując funkcję *predict* na stworzonym modelu możemy zobaczyć, jaka prosta regresji została dopasowana do naszych danych:




Rysunek 5: Prosta regresji dopasowana do badanego zbioru danych.

## 3.1 Dopasowanie modelu

Jakość dopasowania sprawdzamy za pomocą dwóch miar:

- $R^2$, które można interpretować jako procent obserwacji położonych na dobranej prostej, tzn. im bliżej 1, tym lepiej. W naszym przypadku $R^2 \approx 0.72$, co daje dość dobry wynik.

$$R^2 = \frac{\sum_{i-1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5}$$

- MSE, czyli błąd średniokwadratowy, to średnia kwadratów różnic między wartością obserwowaną a estymowaną. Dla naszej estymacji mamy MSE $\approx 13.62\text{cm}^2$ czy też łatwiejsze w interpretacji (ponieważ wyrażone w tej samej jednostce, co estymowane wartości) RMSE $= \sqrt{\text{MSE}} \approx 3.7$cm. Przy wartościach z zakresu między 140 a 187cm (Rysunek 4) daje to zadowalająco niski wynik.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{6}$$

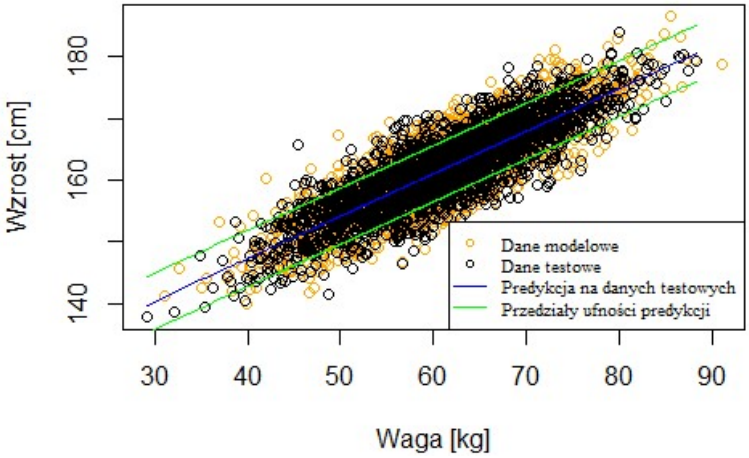
## 3.2 Predykcja

Następnie wykorzystana została próbka testowa do predykcji na podstawie stworzonego modelu. Chcąc sprawdzić, czy model poprawnie przewiduje wartości, musimy sprawdzić, ile obserwacji należy do przedziału ufności na poziomie $\alpha$, tzn. sprawdzamy czy prawdziwa wartość $Y(x_0)$ z prawdopodobieństwem $1-\alpha$ zawiera się w przedziale:

$$\left(\hat{Y}(x_0) - t_{n-2,1-\frac{\alpha}{2}} S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \hat{Y}(x_0) + t_{n-2,1-\frac{\alpha}{2}} S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right), \tag{7}$$

gdzie $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$ oraz $t_{n-2,1-\frac{\alpha}{2}}$ – kwantyl rozkładu t-studenta rzędu $1 - \frac{\alpha}{2}$ z $n-2$ stopniami swobody.

Poniższy wykres prezentuje dane testowe z przewidzianą prostą i przedziałami ufności dla $\alpha = 5\%$. Sprawdzając, ile obserwacji z próbki testowej faktycznie znajduje się wewnątrz przedziału ufności, otrzymujemy wynik 94.72%, co oznacza, że model został dobrany poprawnie, a zatem możemy przewidywać kolejne wartości i z dużym prawdopodobieństwem będą się one rzeczywiście znajdować wewnątrz takiego przedziału.



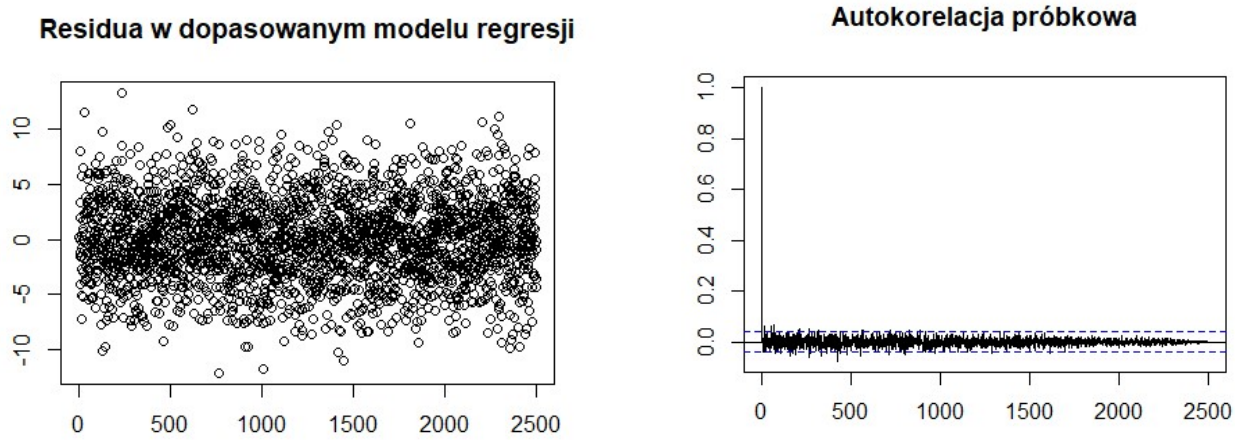

Rysunek 6: Predykcja wykonana dla próbki testowej.

### 3.3 Analiza residuów

Pamiętając o założeniach klasycznego modelu regresji liniowej, nie można zapomnieć o analizie residuów, tzn. realizacji zmiennych losowych $\xi_i$. Residua definiujemy jako $e_i = y_i - \hat{y}_i$, czyli odległości między rzeczywistymi danymi a prostą regresji. W tej sekcji sprawdzimy, czy uzyskane residua:

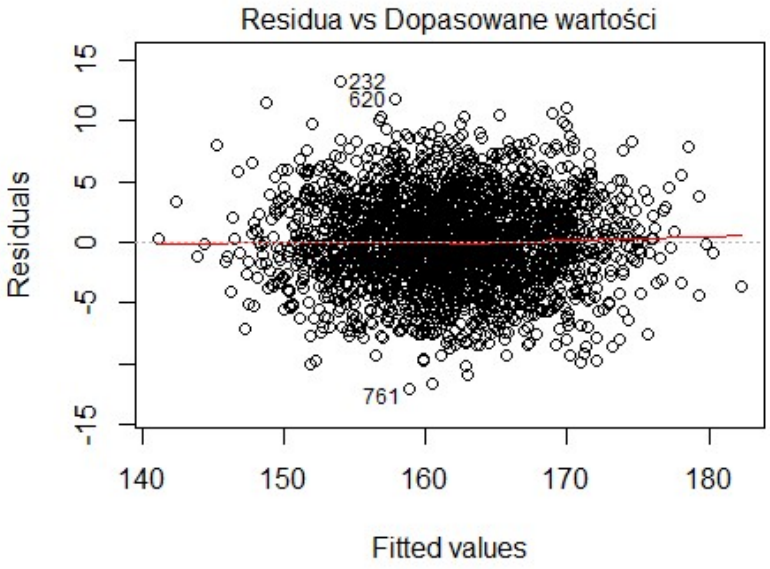
- są niezależne,
- są z rozkładu normalnego o średniej równej zero i stałej wariancji.

W pierwszej kolejności testowana jest niezależność. Szukamy zależności na wykresie próbki oraz sprawdzamy, czy próbkowa funkcja autokorelacji, która sprawdza zależność obserwacji oddalonych o coraz większe kroki, zwróci znaczącą wartość w punktach innych niż 0.



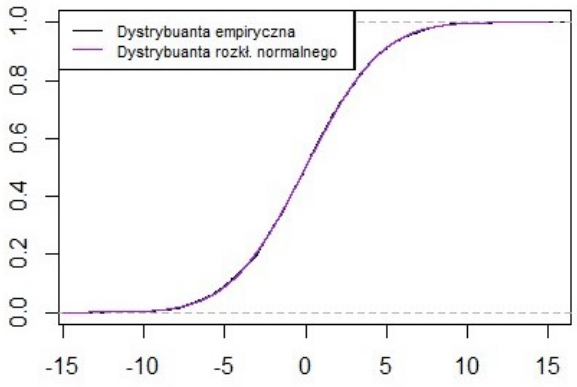

Powyższe wykresy pozwalają nam dojść do wniosku, że residua nie są od siebie zależne.

Następnie testować będziemy, czy $e_i$ pochodzą z rozkładu normalnego o stałej wariancji i średniej równej 0. Średnia próbkowa wynosi $\overline{e} = -1.23 \cdot 10^{-16}$, jest zatem bardzo bliska 0. Odchylenie standardowe wynosi natomiast $S = 3.69$, a niezmienność (brak trendu) w zależności od dopasowanych wartości prezentuje poniższy wykres:

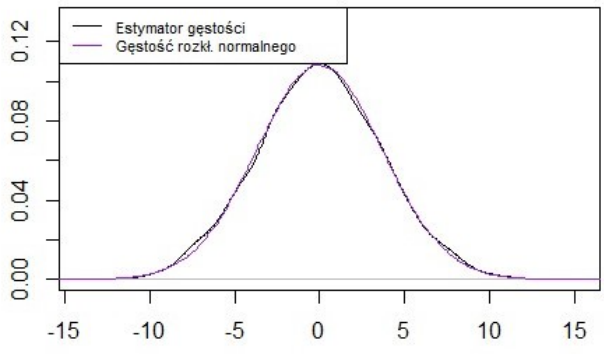



Rysunek 7: Porównanie residuów z dopasowanymi wartościami.

Pozostaje nam zatem sprawdzenie normalności rozkładu residuów. Najprostszym testem jest wizualne porównanie dystrybuanty empirycznej z teoretyczną oraz estymatora gęstości z gęstością teoretyczną rozkładu normalnego. Testujemy zgodność próbki z rozkładem $\mathcal{N}(0, 3.69)$.



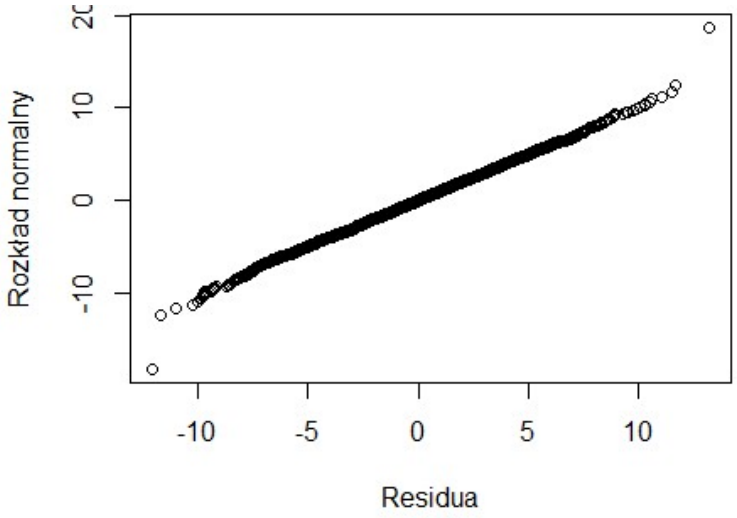

Rysunek 8: Porównanie dystrybuant




Rysunek 9: Porównanie gęstości

Widoczne jest bardzo dobre dopasowanie. Kolejnym testem wizualnym jest wykres kwantylowo-kwantylowy, tzn. wykres rozrzutu utworzony przez wykreślenie dwóch zestawów kwantyli względem siebie. Im bardziej podobne są dwa porównywane rozkłady, tym bliżej prostej $y = x$ leżą punkty.

Poniższy wykres prezentuje dużą zależność między próbkami.




Rysunek 10: Wykres kwantylowo-kwantylowy dla zmiennych losowych z rozkładu normalnego oraz próbki residuów.

Wszystkie testy wizualne potwierdzają nasze przypuszczenia, jednak aby zyskać pewność wykorzystamy bardziej dokładne testy statystyczne. Poniżej znajdują się krótkie, teoretyczne wyjaśnienia oraz uzyskane wyniki.

1. **Test Kołmogorowa-Smirnowa** jest bardzo ważnym, nieparametrycznym testem, badającym zgodność rozkładu empirycznego (próbkowego) z rozkładami teoretycznymi (analitycznymi) poprzez sprawdzenie, czy odległości między dystrybuantą empiryczną a teoretyczną dążą do 0.
   Średnia p-wartość wyliczona na postawie 500 powtórzeń KStest z niezależnymi próbkami rozkładu normalnego i naszymi residuami wynosi około 98.91%.

2. **Test Shapiro-Wilka** jest uznawany za najlepszy test do sprawdzenia normalności rozkładu zmiennej losowej. Największą zaletą tego testu jest duża moc, czyli prawdopodobieństwo odrzucenia hipotezy $H_0$, jeśli jest ona fałszywa.
   P-wartość w tym teście dla naszych residuów wyniosła 88.83%.

3. **Test Jarque-Bera** z uwagi na swoją prostotę i znaną nieskomplikowaną postać rozkładu asymptotycznego jest często wykorzystywany. Konstrukcja statystyki testowej bazuje na wartościach momentów (kurtozy i skośności) wyliczonych z próby empirycznej i porównaniu ich z momentami teoretycznymi rozkładu normalnego.
   W tym wypadku otrzymana p-wartość wyniosła 68.69%.

# 4 Podsumowanie i wnioski

- Regresja liniowa to przydatne narzędzie statystyczne pozwalające nie tylko znaleźć zależność między zmiennymi, ale także przewidywać kolejne wartości.
- Hipotezę postawioną na wstępie – waga i wzrost kobiet są zależne liniowo – możemy uznać za słuszną. Prosta regresji jest dobrze dopasowana do pobranego zestawu danych, o czym świadczą wyniki uzyskane w sekcji 3.1 oraz prawidłowa predykcja na nowych danych w sekcji 3.2.
- Model regresyjny okazał się być istotny statystycznie, tzn. "przydatny" do oszacowania wartości zmiennej zależnej (wzrostu) na podstawie wartości predyktora (wagi).
- Podział danych na próbkę modelową i testową o podobnych parametrach pozwolił na dobranie modelu oraz sprawdzenie jego mocy poprzez predykcję na innych danych.

- Przedziały ufności predykcji stworzone na podstawie poprawnie dobranego modelu pozwalają na określenie z zadanym prawdopodobieństwem, w jakim zakresie znajduje się rzeczywista wartość dla danej obserwacji.
- Analiza residuów wykazała, że z dużym prawdopodobieństwem są one niezależne i pochodzą z rozkładu normalnego o średniej równej zero i stałej wariancji. Potwierdziły to testy wizualne oraz wysokie p-wartości w testach statystycznych. Oznacza to, że poprawnie został dobrany teoretyczny model regresji liniowej.