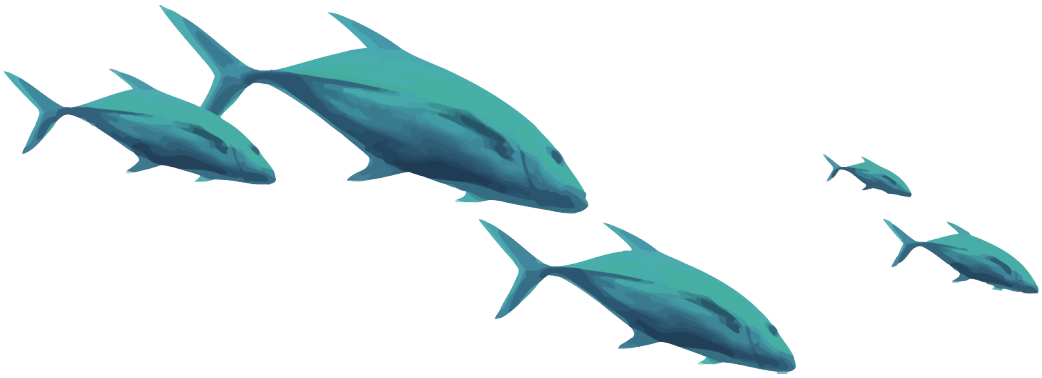
Regresja liniowa

Katarzyna Turbańska, Edyta Łabędzka

Wrocław, 9 lutego 2022



1. Opis danych

Przeanalizujemy dane dotyczące ryb. Będziemy chciały stworzyć model przewidujący wagę ryby w zależności od jej szerokości. Wykorzystamy w tym celu dane z Kaggle – „Fish market" opisujące siedem gatunków ryb sprzedawanych na targach (m.in. leszcz, płotka, okoń, szczupak). Zmienne, na których będzie opierać się nasza analiza, to „Width" oraz „Weight", opis w tabeli 1.

Zmienne	Opis	Typ	Ilość obserwacji
Width	szerokość ryby [cm]	zmienna ciągła	159
Weight	waga ryby [g]	zmienna ciągła	159

Tabela 1.1. Opis zmiennych

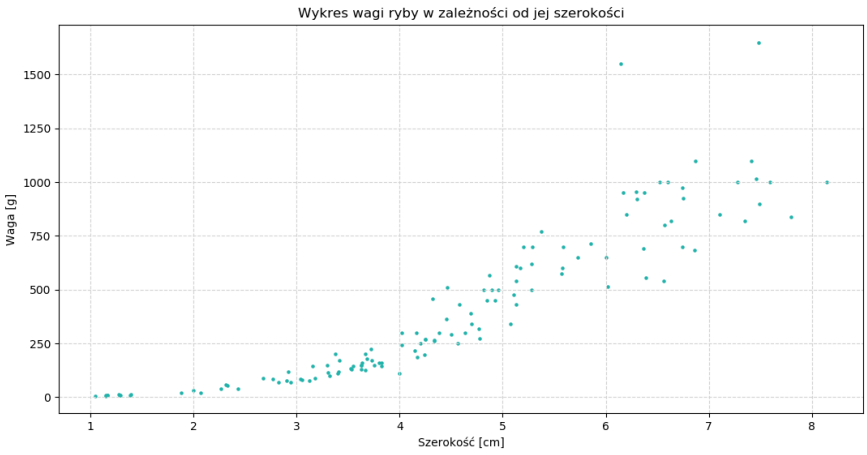
Ponieważ będziemy chciały sprawdzić działanie modelu, dzielimy dane na zbiór treningowy (127 obserwacji) i testowy (32 obserwacje). Poniżej opiszemy zmienne ze zbioru treningowego, ponieważ to z nich zostanie wyznaczona regresja.

2. Statystyki opisowe

W tabeli 2, widzimy opis podstawowych statystyk dla zmiennej objaśnianej – waga i objaśniającej – szerokość. Na rysunku 2.1 widzimy wykres rozproszenia dla tych zmiennych.

Statystyka	szerokość	waga
Ilość obserwacji	127	127
Średnia	4.4598	407.2150
Odchylenie standardowe	1.7265	354.4223
Minimum	1.0476	6.7
Pierwszy kwartyl	3.3857	122.5
Mediana	4.3350	290
Trzeci kwartyl	5.6583	650
Maksimum	8.1420	1650

Tabela 2.1. Wartości statystyk opisowych



Rysunek 2.1. Wykres wagi ryby w zależności od jej szerokości

3. Dobranie prostej regresji do danych

Współczynnik korelacji Pearsona dla „Width" oraz „Weight" wynosi 0.886507, czyli mamy silną zależność liniową pomiędzy tymi zmiennymi. Oprócz tego, patrząc na wykres rozproszenia 2.1, możemy powiedzieć, że dane układają się wzdłuż krzywej.

Zatem mając dwa zbiory danych zaobserwowanych: $x_1, x_2, \ldots, x_n$ oraz $y_1, y_2, \ldots, y_n$, które są liniowo zależne chcemy wyznaczyć prostą regresji. Nasz model ma postać:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

gdzie

- y_i – zmienna objaśniana;
- x_i – zmienna objaśniające;
- β_0 – wyraz wolny;
- β_1 – współczynnik kierunkowy prostej;
- ε_i – błąd losowy.

Współczynniki do regresji liniowej wyznaczamy metodą najmniejszych kwadratów, z której otrzymujemy wzory:

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

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

gdzie $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ oraz $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Estymowane $\hat{y}_i$ mają więc postać:

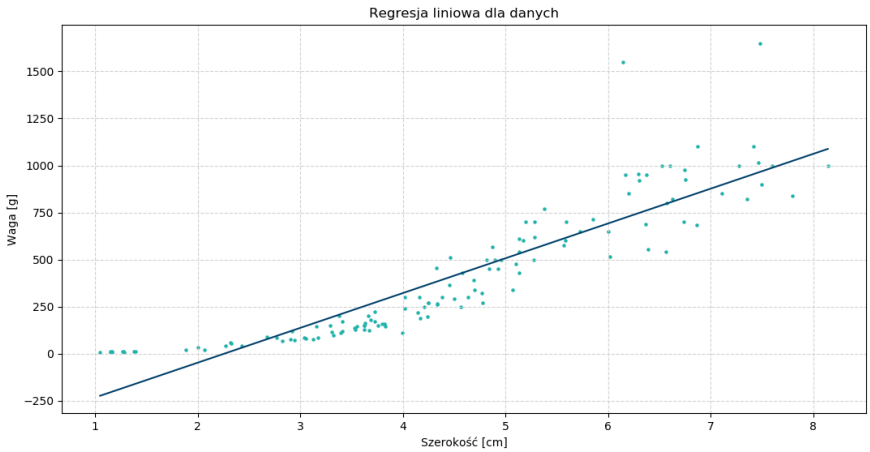
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Korzystając z powyższych wzorów, współczynniki dla naszych danych wyniosły: $\beta_0 = -417.19551833$, $\beta_1 = 184.85424310$. Prosta regresji z wyznaczonymi współczynnikami naniesiona jest na wykresie 3.1.

A więc chcąc sprawdzić jak dobrze dopasowana jest nasza prosta, obliczamy współczynnik determinacji R^2.

Współczynnik determinacji R^2 jest miarą jakości dopasowania modelu do próby. Przyjmuje wartości z przedziału $[0, 1]$, przy czym można powiedzieć, dopasowanie jest tym lepsze, im wartość R^2 jest bliższa 1. Jest on dany wzorem:

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



Rysunek 3.1. Regresja liniowa dla danych

Dla tej prostej regresji współczynnik determinacji wyniósł: $R^2 = 0.81083015$.

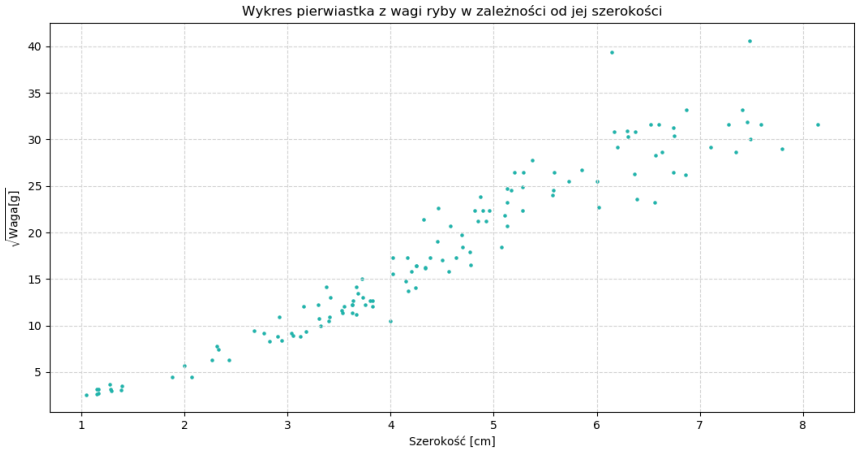
Widzimy jednak że nasze dane nie do końca układają się wzdłuż prostej, a bardziej przypominają funkcję kwadratową. Dlatego przeprowadzimy transformację $y_1, y_2, \ldots, y_n \to \sqrt{y_1}, \sqrt{y_2}, \ldots, \sqrt{y_n}$. Wykres rozproszenia dla przetransformowanych danych znajduje się na rysunku 4.1. Można zauważyć, że teraz dane układają się wzdłuż prostej.

4. Transformacja danych

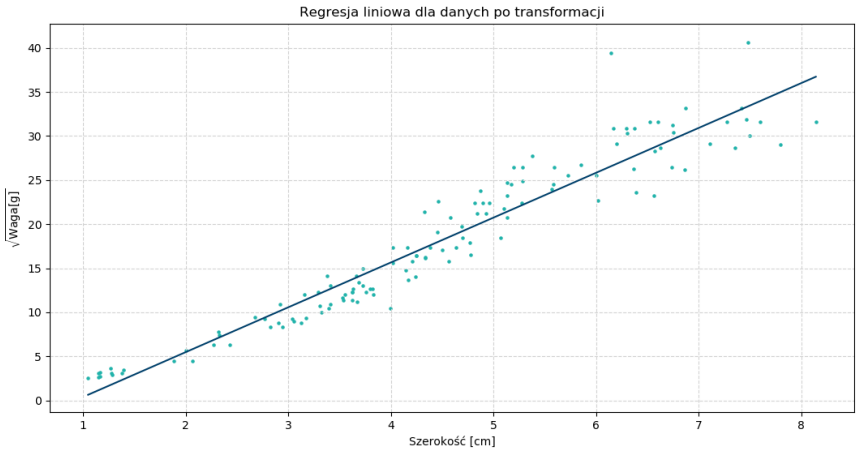
Wyznaczymy prostą regresji dla zmiennych po transformacji, czyli $x_1, x_2, \ldots, x_n$ oraz $\sqrt{y_1}, \sqrt{y_2}, \ldots, \sqrt{y_n}$. Otrzymujemy współczynniki równe: $\beta_0 = -4.68086356$, $\beta_1 = 5.08473851$. Wzór ma postać

$$\hat{y}_i = -4.68086356 + 5.08473851 x_i.$$

Czyli przy rybie szerszej o 1 [cm] jej waga wzrasta o około 0.16311498 [g]. Prosta regresji wygląda teraz następująco 4.2. Współczynnik determinacji zwiększył się do 0.91719614, czyli po transformacji mamy lepsze dopasowanie. Dalsze opracowanie przeprowadzamy już dla regresji z danych po transformacji.



Rysunek 4.1. Wykres pierwiastka z wagi ryby w zależności od jej szerokości



Rysunek 4.2. Prosta regresji dla danych po transformacji

5. Przedziały ufności

Przedział ufności to przedział, w którym z prawdopodobieństwem $1-\alpha$ mieści się poszukiwana przez nas rzeczywista wartość, gdzie α jest zadanym poziomem istotności.

Przedział ufności dla β_0 przy nieznanej σ wyznaczamy następująco

$$P\left(\hat{\beta}_0 - \theta_0 < \beta_0 < \hat{\beta}_0 + \theta_0\right) = 1 - \alpha,$$

gdzie

- $\theta_0 = t_{1-\frac{\alpha}{2},n-2} \cdot S \sqrt{\frac{1}{n} + \frac{(\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}$;
- $S = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i-\hat{y})^2}{n-2}}$ – próbkowe odchylenie standardowe;
- $t_{1-\frac{\alpha}{2},n-2}$ – kwantyl rzędu $1-\frac{\alpha}{2}$ rozkładu Studenta z $n-2$ stopniami swobody.

Przedział ufności dla β_1 przy nieznanej σ wyznaczamy analogicznie:

$$P\left(\hat{\beta}_1 - \theta_1 < \beta_1 < \hat{\beta}_1 + \theta_1\right) = 1 - \alpha,$$

gdzie

$$\theta_1 = t_{1-\frac{\alpha}{2},n-2} \cdot S \frac{1}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Przyjmujemy $\alpha = 0.05$ oraz obliczamy przedziały ufności β_0 i β_1 stosując powyższe wzory. Otrzymujemy:

- przedział ufności dla β_0: $[-5.97355358, -3.38817354]$;
- przedział ufności dla β_1: $[4.81429234, 5.35518468]$.

6. Analiza residuów

Aby dowiedzieć się czy nasz model jest dobrany poprawnie, sprawdzimy czy jego założenia zostały spełnione:
1° $\mathbb{E}\varepsilon_i = 0$;
2° $\mathrm{Var}\varepsilon_i = \sigma^2 = \mathrm{const}$;
3° $\varepsilon_1, \ldots, \varepsilon_n$ są nieskorelowane;
4° $\varepsilon_i \sim N(\mu = 0, \sigma^2)$.
Residua e_i są realizacjami zmiennej ε_i. Dlatego na ich podstawie sprawdzimy założenia o ε_i.

Residuum danej obserwacji to błąd oszacowania. Będziemy je obliczać odejmując od wartości rzeczywistych y_i ich wartości estymowane $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Warunek 1° i 2°

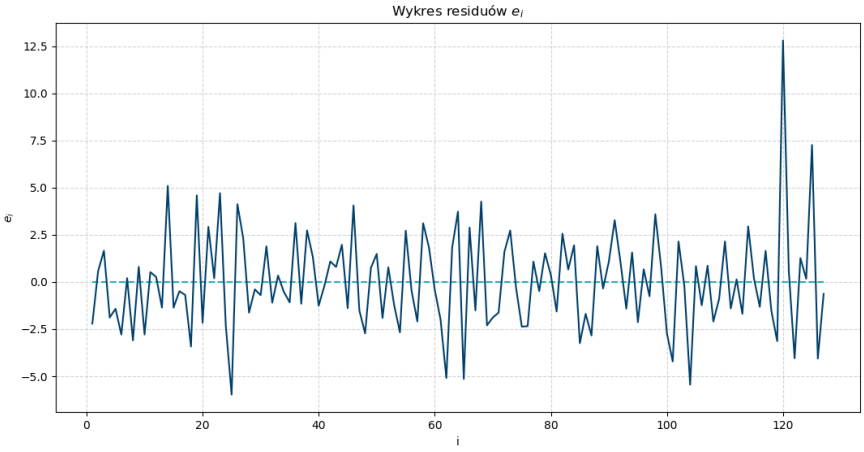
Zaczniemy od wypisania w tabeli 6 podstawowych statystyk dla residuów.

Statystyka	Wartość
Średnia	$8.1055 \cdot 10^{-15}$
Odchylenie standardowe	2.6377
Minimum	-5.9755
Pierwszy kwartyl	-1.6582
Mediana	-0.2922
Trzeci kwartyl	1.5797
Maksimum	12.8103

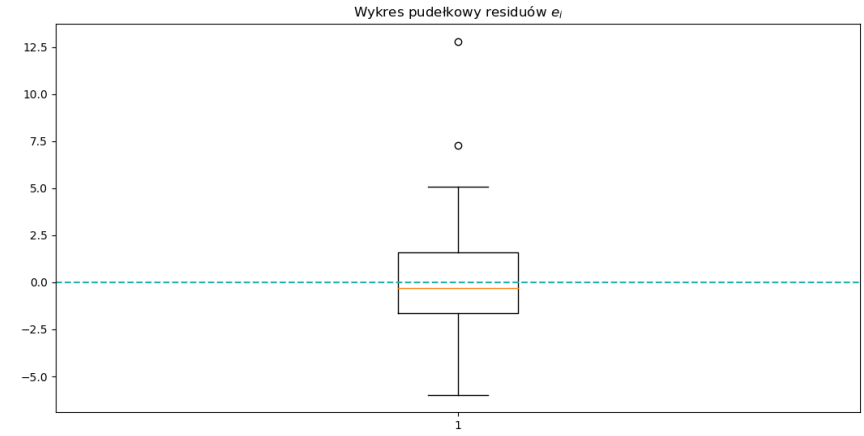
Tabela 6.1. Wartości statystyk opisowych residuów

Jak widzimy w tabeli 6 średnia w przybliżeniu wynosi 0. Na podstawie wykresu residuów 6.1 możemy stwierdzić, że średnia jest bliska 0, ponieważ wartości rozkładają się mniej więcej symetrycznie. Natomiast wariancja (kwadrat z odchylenia standardowego) nie jest ani rosnąca, ani malejąca oraz oprócz kilku obserwacji odstających mieści się w pewnym przedziale. Czyli spełnione są założenia 1 i 2.

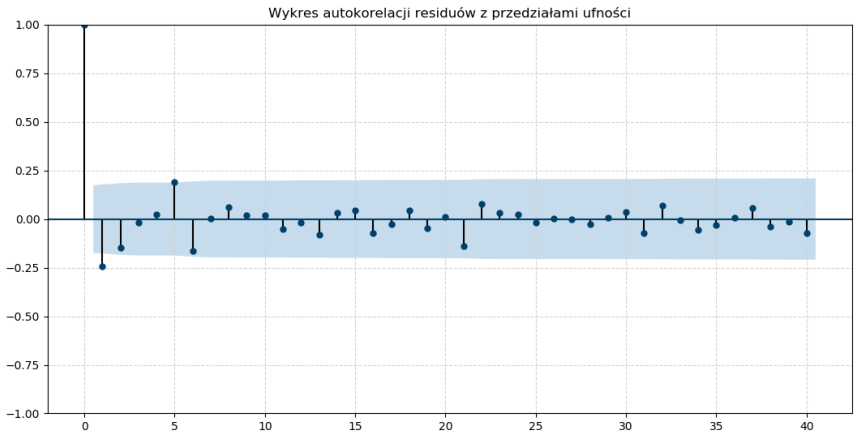
Za wartości nietypowe uznajemy obserwacje, które nie znajdują się w siatce Tukeya $f_{1.5}$. Na wykresie 6.2 widzimy, że mamy 2 wartości nietypowe. Po długości pudełka i wąsów, a także medianie bliskiej 0, możemy stwierdzić, że rozkład może być w przybliżeniu symetryczny.



Rysunek 6.1. Wykres residuów e_i



Rysunek 6.2. Wykres pudełkowy residuów e_i



Rysunek 6.3. Wykres autokorelacji residuów

Warunek 3°

Korelacja dla większości zmiennych jest bliska zero. Jest ona ogólnie słaba i mieści się w przedziale ufności 95%, co możemy zobaczyć na wykresie 6.3. Czyli możemy stwierdzić, że residua są nieskorelowane.

Warunek 4°

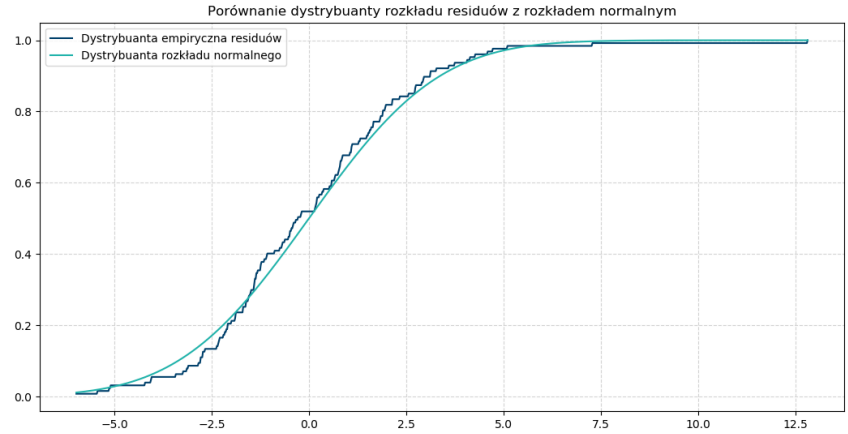
Na wykresach 6.4 i 6.5 przedstawiłyśmy graficzne porównanie dystrybuant oraz gęstości dla rozkładu residuów oraz rozkładu normalnego z parametrami μ i σ, odpowiadającymi średniej i wariancji wyznaczonych residuów. Dystrybuanta jest dobrze dopasowana, natomiast w przypadku gęstości, widzimy, że nie do końca się one pokrywają. Po histogram residuów widać, że ich rozkład jest delikatnie prawostronnie skośny, a nie symetryczny – tak jak założyłyśmy przy boxplocie. Żeby mieć pewność co do rozkładu przeprowadzimy jeszcze testy na normalność rozkładu.

Test Kołmogorowa-Smirnowa dla jednej próby sprawdza, czy rozkład w populacji dla pewnej zmiennej losowej, różni się od założonego rozkładu teoretycznego, gdy znana jest jedynie próba statystyczna. Test ten bazuje na dystrybuancie empirycznej F_n, a jego statystyka dla danej dystrybuanty teoretycznej $F(x)$ zdefiniowana jest jako:

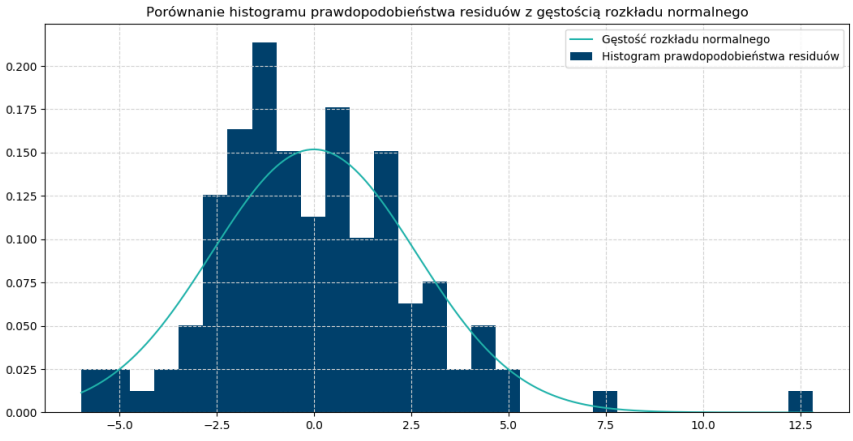
$$D_n = \sup_x |F_n(x) - F(x)|.$$

Wartości testu Kołmogorowa-Smirnowa dla unormowanych residuów: statystyka $D_n = 0.06143673$, p-wartość $= 0.73663690$.

Statystyka D jest bliska 0, czyli szansa na to, że residua pochodzą z rozkładu normalnego jest duża. P-wartość dla testu Kołmogorowa-Smirnowa jest dużo większa niż zadany poziom istotności $\alpha = 0.05$, więc nie mamy podstaw, aby odrzucić hipotezę zerową zakładającą, że unormowane residua pochodzą ze standardowego rozkładu normalnego.



Rysunek 6.4. Porównanie dystrybuanty rozkładu residuów z rozkładem normalnym



Rysunek 6.5. Porównanie gęstości rozkładu residuów z rozkładem normalnym

Test Andersona-Darlinga to test, który sprawdza czy badana próbka pochodzi z zadanego rozkładu, korzystając z ważonej odległości Craméra-von Misesa – w tym przypadku waga wynosi $w(x) = [F(x)\ (1 - F(x))]^{-1}$. Test AD również bazuje na dystrybuancie empirycznej F_n, a statystyka dla danej dystrybuanty teoretycznej $F(x)$ ma postać:

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

Wartości testu Andersona-Darlinga dla residuów: statystyka $A^2 = 0.71156460$, wartość krytyczna dla $\alpha = 0.05$ wynosi 0.764.

W teście Andersona-Darlinga żeby odrzucić hipotezę o rozkładzie normalnym residuów, statystyka musiałaby być większa od wartości krytycznej odpowiadającej przyjętemu poziomowi istotności. Jak widzimy powyżej, w tym przypadku nie mamy podstaw, by odrzucić ta hipotezę.

Test Jarque-Bera sprawdza czy kurtoza i skośność badanej próby są takie same jak dla rozkładu normalnego. Statystykę testową Jarque-Bera definiujemy następująco:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K - 3)^2\right),$$

gdzie n to długość próby, S to skośność, a K to kurtoza.

Wartości testu Jarque-Bera dla residuów: statystyka $JB = 82.21496144$, p-wartość $= 0$.

Statystyka jest bardzo daleko od 0, co sygnalizuje, że residua, nie maja rozkładu normalnego. P-wartość dla testu Jarque-Bera wyniosła 0, czyli powinniśmy odrzucić hipotezę zerową o normalności residuów.

Wnioski Założenia $1° - 3°$ są spełnione, natomiast w przypadku $4°$, residua nie spełniają założenia o rozkładzie normalnym. Na podstawie testu Kołmogorowa-Smirnowa i Andersona-Darlinga nie mogliśmy odrzucić hipotezy zerowej. Natomiast z testu Jarque-Bera wynika, że rozkład residuów nie jest zgodny z rozkładem normalnym. Najprawdopodobniej jest to spowodowane prawostronną skośnością (widoczną na wykresie 6.5) ich rozkładu, przez co test uwzględniający skośność nie przeszedł. W takim wypadku niestety nasz model jest niepoprawny.

7. Predykcje

Na początku podzieliłyśmy dane, dzięki czemu możemy teraz sprawdzić dokładność naszego modelu. Korzystając ze zbioru testowego X, przewidujemy wartości odpowiadających im Y za pomocą naszej regresji. Wyniki możemy zobaczyć na wykresie 7.1. Obliczamy jakość dopasowania dla tych punktów: $R^2 = 0.67127822$. Oprócz tego chcemy wyznaczyć przedziały ufności predykcji dla $y(x_0)$ przy nieznanej σ. Wyznaczamy je następująco:

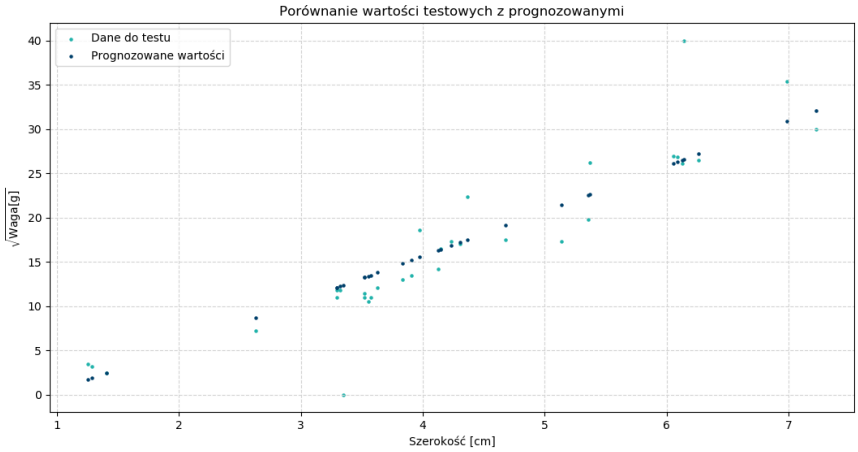
$$P\left(\hat{Y}(x_0) - \theta < Y(x_0) < \hat{Y}(x_0) + \theta\right) = 1 - \alpha,$$

gdzie

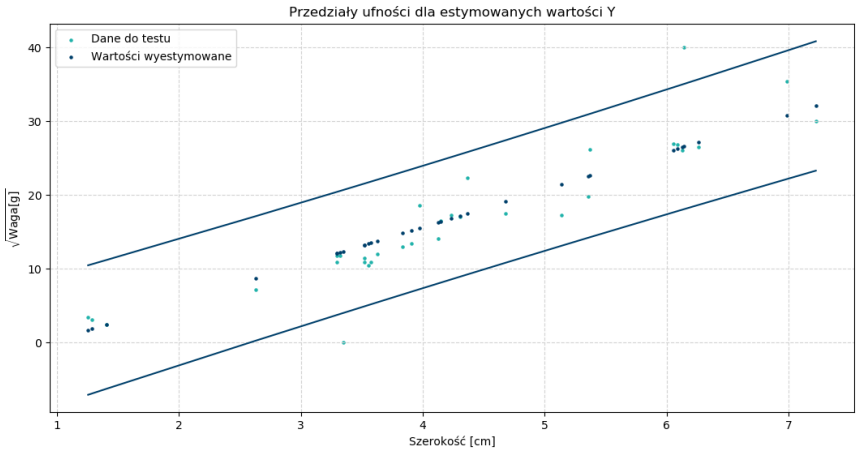
- $\theta = t_{1-\frac{\alpha}{2},n-2} \cdot S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$.

Przedziały wyznaczone z powyższych wzorów dla $\alpha = 0.05$ zostały przedstawione na wykresie 7.2.

Model nie jest zbyt dokładny, co jest dobrze widoczne na wykresie 7.1. Potwierdza to również wartość współczynnika determinacji. Mimo wszystko, tylko dwa punkty nie trafiły do przedziałów ufności, jak możemy zobaczyć na rysunku 7.2. Ogólnie model jest w stanie z akceptowalną dokładnością przewidzieć wartości; nie jest to losowe, skoro $R^2 > 0.5$.



Rysunek 7.1. Porównanie wartości testowych z prognozowanymi



Rysunek 7.2. Przedziały ufności dla estymowanych wartości Y

8. Wnioski

Gdyby nasz model był poprawny, mogłybyśmy stwierdzić, o ile dokładnie zmienia się pierwiastek z wagi ryby w zależności od jej szerokości dla gatunków ryb ze zbioru analizowanych przez nas danych.

Jednak skoro residua nie spełniają założeń, to model może być niepoprawny, a tym samym wnioski wysuwane na jego podstawie błędne. Brak normalności residuów przede wszystkim powoduje błędy w wyznaczaniu przedziałów ufności. Jednakże warto pamiętać, że w przypadku rzeczywistych danych ciężko jest otrzymać rozkład normalny błędów.

W tym przypadku możliwe, że warto byłoby zastosować inny model, mający inne założenia. Można by również rozważyć usunięcie wartości odstających, które często są przyczyną skośności rozkładu, przez co nie jest spełnione założenie o normalności. Nie zrobiłyśmy tego, ponieważ nie jesteśmy w stanie stwierdzić, czy ryba danego gatunku może urosnąć do „odstających" rozmiarów.