Aleksandra Janczewska

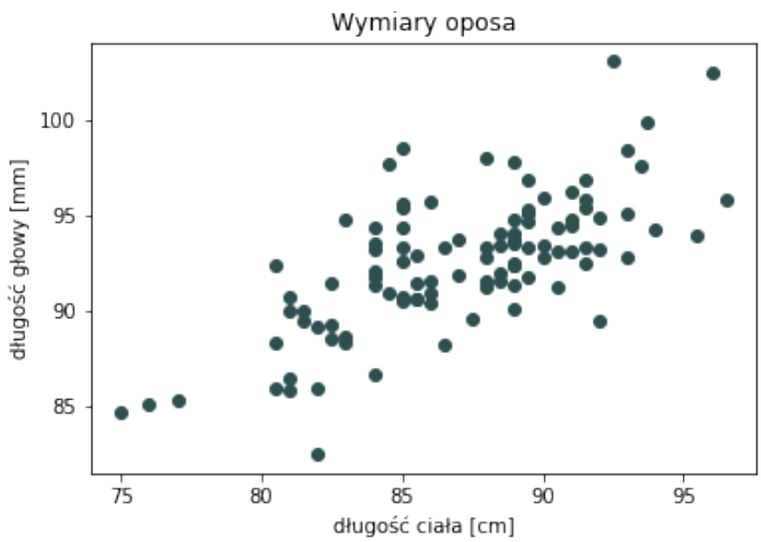
Analiza danych z wykorzystaniem regresji liniowej

1. Opis danych.

Dane, z których będę korzystać w moim sprawozdaniu pochodzą ze strony https://www.kaggle.com i składają się ze 104 obserwacji dotyczących oposów górskich z obszaru od Wikorii Południowej do środkowego Queensland w Australii.

Dane zawierają 14 kolumn dotyczących pomiarów morfometrycznych i charakterystyk badanych oposów. W sprawozdaniu wykorzystam dwie z tych kolumn. Jako pierwszy zbiór potraktuję kolumnę „totlngth". Zawiera ona pomiary długości ciała oposa podane w centymetrach. Jako drugi zbiór wykorzystam kolumnę „hdlngth". Jest to długość głowy oposa podana w milimetrach.

1.1 Wykres danych.



Rysunek 1: wykres danych

2. Statystyki opisowe danych.

Wartości w tabeli 1 są zaokrąglone do dwóch miejsc po przecinku.

nazwa zmiennej	średnia	wariancja	minimum	maksimum	mediana
długość ciała [cm]	87,09	18,40	75,00	96,5	88,00
długość głowy [mm]	92,60	12,65	82,5	103,10	92,80

Tablica 1: Podstawowe statystyki opisowe

3. Regresja liniowa.[1]

Jest to metoda oparta o liniowe kombinacje zmiennych i parametrów dopasowujących model do danych. Model teoretyczny regresji liniowej wyraża się wzorem:

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

gdzie:

* x_i to zmienna objaśniająca (z danych),
* y_i to zmienna objaśniana (z danych),
* β_1,β_0 to stałe,
* ε_i to residuum.

Wzór na prostą regresji dla danych:

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

gdzie:

* x to zmienna deterministyczna,
* $\hat{y}_i$ to estymowana wartość y_i,
* $\hat{\beta}_1$ to estymator parametru β_1 z modelu teoretycznego,
* $\hat{\beta}_0$ to estymator parametru β_0 z modelu teoretycznego.

Estymatory $\hat{\beta}_1$ oraz $\hat{\beta}_0$ wyznaczone są za pomocą następujących wzorów:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \approx 0.573,$$

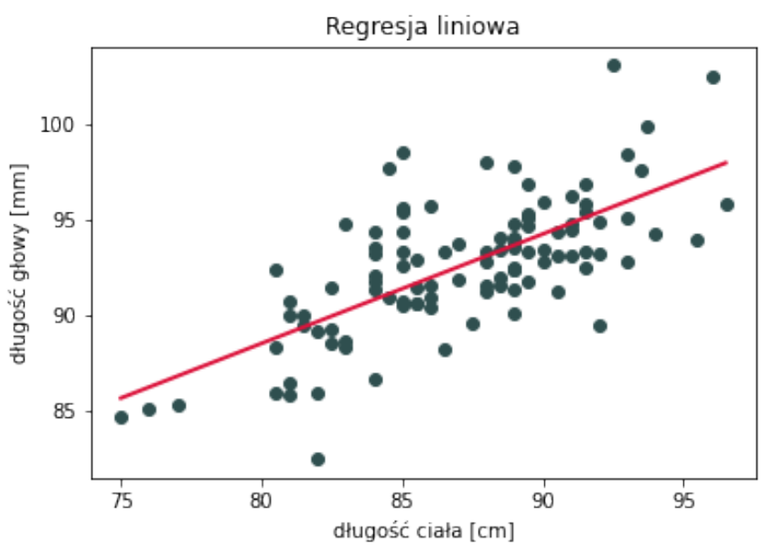
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \cdot \overline{x} \approx 42.709.$$

gdzie:

* n to liczba obserwacji,
* x_i to zmienna objaśniająca (z danych),
* y_i to zmienna objaśniana (z danych),
* $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
* $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Wykres 2 przedstawia prostą regresji liniowej dla danych:

$$\hat{y}_i = 0.573 \cdot x_i + 42.709.$$



Rysunek 2: regresja liniowa

3.1 Współczynnik determinacji.[2]

Współczynnik determinacji r^2 - kwadrat współczynnika korelacji próbkowej pomiędzy zmienną objaśniającą, a objaśnianą. Opisuje on jakość dopasowania modelu do danych. Obecnie, współczynnik ten wykorzystuje się głównie w celach pomocniczych. Przyjmuje wartości z przedziału $[0, 1]$. Wyraża się wzorem:

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

gdzie:

* n to liczba obserwacji,
* y_i to zmienna objaśniana (z danych),
* $\hat{y}_i$ to estymowana wartość y_i,
* $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

Współczynnik wyliczony dla danych wynosi: $r^2 \approx 0.478$. Oznacza to, że regresja wyjaśnia tylko mniej niż 50% zmienności zmiennej objaśnianej, a zatem model regresji liniowej może być niedokładnie dopasowany do danych.

4. Przedziały ufności dla $\hat{\beta}_0$ oraz $\hat{\beta}_1$.

Zakładamy, że wartość σ jest nieznana oraz $\varepsilon \sim N(0, \sigma^2)$. Wyznaczamy statystykę T.

$$T = \frac{\hat{\beta}_0 - \beta_0}{s \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}} \sim t_{n-2}$$

gdzie:

* t_{n-2} to rozkład t-studenta z $n-2$ stopniami swobody,
* B_0 to teoretyczna wartość estymowanego parametru $\hat{\beta}_0$,
* $s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$.

Dla poziomu ufności $1-\alpha$ statystyka T mieści się pomiędzy kwantylami rozkładu t-studenta ($-t_{1-\frac{\alpha}{2},n-2}$ oraz $t_{1-\frac{\alpha}{2},n-2}$) z prawdopodobieństwem $1-\alpha$:

$$P(-t_{1-\frac{\alpha}{2},n-2} \leq T \leq t_{1-\frac{\alpha}{2},n-2}) = 1 - \alpha.$$

Podstawiamy statystykę T i po przekształceniach mamy:

$$P\left(\hat{\beta}_0 - t_{1-\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}} \leq \beta_0 \leq \hat{\beta}_0 + t_{1-\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}\right) = 1 - \alpha.$$

W taki sposób wyznaczamy przedział ufności dla parametru $\hat{\beta}_0$ na poziomie istotności α.

$$\left[\hat{\beta}_0 - t_{1-\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}\right]$$

Dla parametru $\hat{\beta}_0 \approx 42.709$ i $\alpha = 0.05$ wygląda on następująco: $[32.450, 52.970]$.

Analogicznie będziemy wyznaczać przedział ufności dla $\hat{\beta}_1$. Najpierw wyznaczamy statystykę T.

$$T = \frac{\hat{\beta}_1 - \beta_1}{\frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}}} \sim t_{n-2}$$

gdzie:

* t_{n-2} to rozkład t-studenta z $n-2$ stopniami swobody,
* B_1 to teoretyczna wartość estymowanego parametru $\hat{\beta}_1$,
* $s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$.

Dla poziomu ufności $1-\alpha$ statystyka T mieści się pomiędzy kwantylami rozkładu t-studenta ($-t_{1-\frac{\alpha}{2},n-2}$ oraz $t_{1-\frac{\alpha}{2},n-2}$) z prawdopodobieństwem $1-\alpha$:

$$P(-t_{1-\frac{\alpha}{2},n-2} \leq T \leq t_{1-\frac{\alpha}{2},n-2}) = 1-\alpha.$$

Podstawiamy statystykę T i po przekształceniach mamy:

$$P\left(\hat{\beta}_1 - t_{1-\frac{\alpha}{2},n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\frac{\alpha}{2},n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\right) = 1-\alpha.$$

Wyznaczamy przedział ufności dla parametru $\hat{\beta}_1$ na poziomie istotności α.

$$\left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2},n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2},n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\right]$$

Przedział wyznaczony dla $\hat{\beta}_1 \approx 0.573$ na poziomie istotności $\alpha = 0.05$ prezentuje się następująco: $[0.455, 0.691]$.

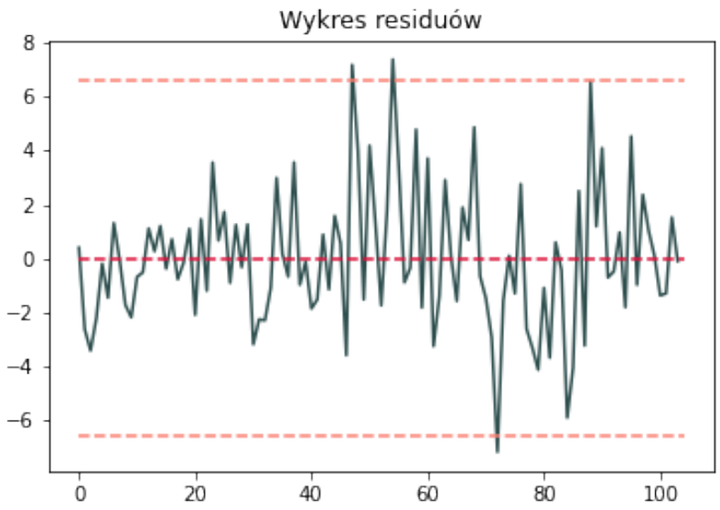
5. Analiza residuów.

W poprawnie dobranym modelu muszą być spełnione następujące założenia dotyczące residuów (ozn. ε_i) $\forall i = 1, ..., n$:

1° $E(\varepsilon_i) = 0$,
2° $var(\varepsilon_i) = \sigma^2 = const$,
3° ε_i są niezależne, $cov(\varepsilon_i, \varepsilon_j) = 0$,
4° $\varepsilon_i \sim N(0, \sigma^2)$

5.1 Analiza średniej i wariancji (warunek 1° i 2°).

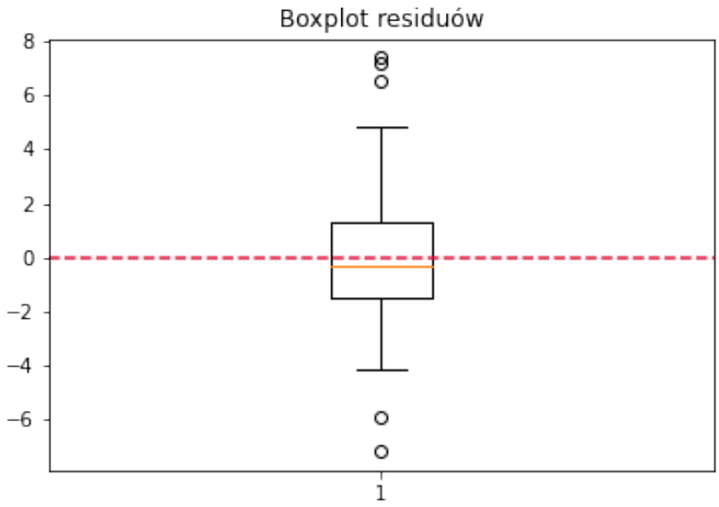
W celu analizy średniej i wariancji residuów przedstawię ich wykres oraz boxplot.



Rysunek 3: wykres residuów

Na wykresie 3 znajdują się wyplotowane residua i jest zaznaczona ich średnia, która oscyluje wokół wartości 0 ($E(\varepsilon_i) \approx -4.51 \cdot 10^{-15} \approx 0$). Liniami przerywanymi jest zaznaczona również wariancja, która ma wartość $var(\varepsilon_i) \approx 6.61$. Z wykresu można odczytać, że 3 obserwacje są większe, niż wariancja, co oznacza, że nie jest ona stała. Dla wszystkich obserwacji z danych, residua spełniają założenie 1°, ale nie spełniają założenia 2°.

Na wykresie 4 widać, że średnia jest lekko poniżej wartości 0. Oscyluje ona wokół tej wartości. Ponadto wykres ten pokazuje, że w zbiorze danych jest 5 obserwacji odstających.



Rysunek 4: boxplot residuów

5.2 Analiza niezależności (warunek 3°).

W celu analizy niezależności residuów, sprawdzę ich korelację za pomocą funkcji autokorelacji. Jeżeli zmienne są nieskorelowane to są również niezależne.

Korelacja to kowariancja unormowana przez iloczyn odchyleń standardowych. Wyraża się ona wzorem:

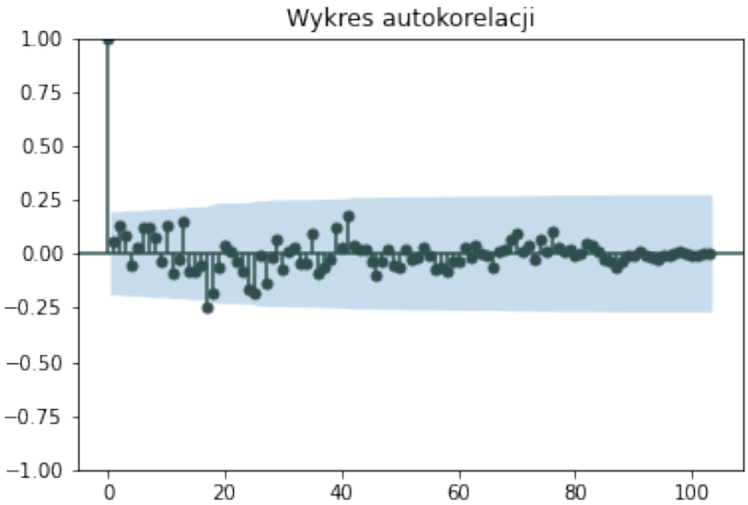
$$corr(\varepsilon_i, \varepsilon_j) = \frac{cov(\varepsilon_i, \varepsilon_j)}{\sigma_{\epsilon_i} \cdot \sigma_{\epsilon_j}}$$

gdzie:

$\sigma_{\varepsilon_i} = \sqrt{var(\varepsilon_i)}$ to odchylenie standardowe zmiennej ε_i,
$\sigma_{\varepsilon_j} = \sqrt{var(\varepsilon_j)}$ to odchylenie standardowe zmiennej ε_j.

Kowariancja jest to liczba określająca odchylenie elementów od sytuacji idealnej, w której występuje zależność liniowa. Wzór na kowarinację:

$$cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) - E(\varepsilon_i) \cdot E(\varepsilon_j).$$



Rysunek 5: wykres autokorelacji

Rysunek 5 prezentuje funkcję autokorelacji dla residuów. Słupki, które należą do obszaru zaznaczonego na wykresie, pokazują dane nieskorelowane. Jak widać, tylko jeden słupek nie mieści się w tym przedziale, możemy więc stwierdzić, że z wyjątkiem tej jednej obserwacji, dane są nieskorelowane, a więc są one niezależne.

5.3 Analiza rozkładu residuów (warunek 4°).

Sprawdzę teraz, czy residua mają rozkład $N(0, \sigma^2)$. W tym celu wykonam testy statystyczne oraz sprawdzę pokrycie danych z rozkładem teoretycznym na wykresach.

5.3.1 Testy statystyczne.

Test Kołmogorowa-Smirnowa[3] jest to test nieparametryczny używany do porównywania rozkładów jednowymiarowych cech statystycznych.

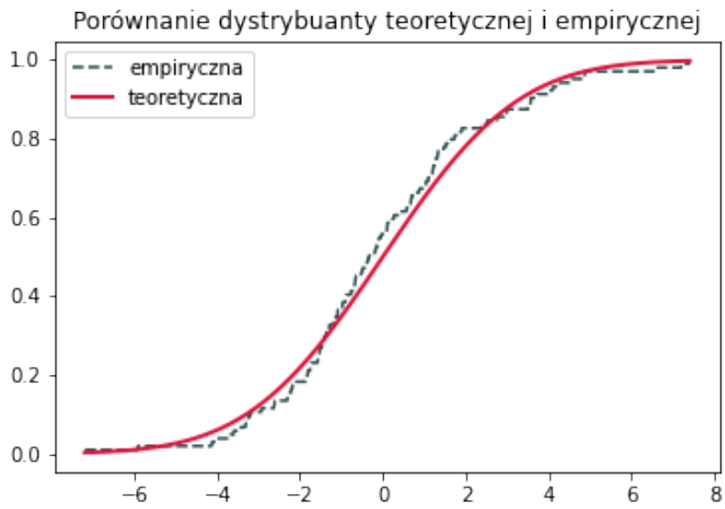
Będę używać testu dla jednej próby, który sprawdza, czy rozkład w populacji dla pewnej zmiennej losowej, różni się od założonego rozkładu teoretycznego, gdy znana jest jedynie pewna skończona liczba obserwacji tej zmiennej. Jeżeli p-value jest większa od poziomu istotności α, wtedy próba ma rozkład normalny. Ustalam poziom istotności $\alpha = 0.05$. Wykonuję test i otrzymuję: $p - value = 0.6192122550752286$. Oznacza to, że $p - value > 0.05$, a co za tym idzie według testu Kołmogorowa-Smirnowa residua mają rozkład normalny.

Kolejnym testem jaki wykonam będzie **test Jarque–Bera**[4]. Jest to test zgodności dopasowania sprawdzający, czy dane próbkowe mają skośność i kurtozę zgodne z rozkładem normalnym.

Jeżeli p-value jest większa od α, wtedy próba ma rozkład normalny. Ustalam poziom istotności $\alpha = 0.05$. Wykonuję test i otrzymuję: $p-value = 0.0774400366217578$. Oznacza to, że $p-value > 0.05$, a więc według testu Jarque–Bera residua mają rozkład normalny.

Na podstawie testów mogę stwierdzić, że residua mają rozkład normalny.

5.3.2 Wykresy.

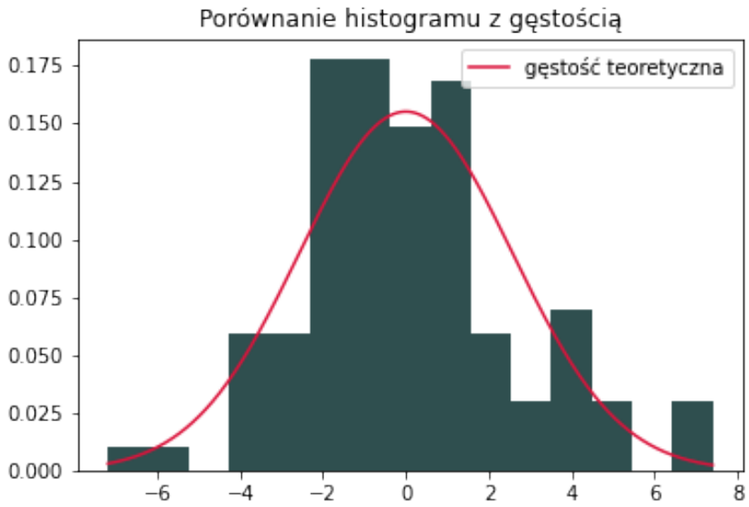


Rysunek 6: dystrybuanty

Na wykresie 6 pokazane jest porównanie dystrybuanty teoretycznej z rozkładu normalnego z parametrami średniej $\mu = E(\varepsilon_i) \approx 0$ oraz wariancji $var(\varepsilon_i) \approx 6.61$ ($N(\mu = 0, \sigma^2 = 6.61)$) z dystrybuantą empiryczną z próby residuów. Dystrybuanty te pokrywają się w większości punktów. Są jednak miejsca, w których na siebie nie nachodzą.

Na rysunku 7 prezentowany jest histogram residuów z nałożoną na niego gęstości teoretyczną z rozkładu normalnego ($N(\mu = 0, \sigma^2 = 6.61)$). Jak można zauważyć gęstość i histogram w pełni się nie pokrywają, jednak nie są też zupełnie różne.

Na podstawie wykresów nie można jednoznacznie stwierdzić, że residua mają rozkład normalny.



Rysunek 7: histogram i gęstość

6. Wnioski.

Podsumowuję teraz jakość modelu regresji liniowej dla danych oraz analizę residuów. Współczynnik determinacji, który wynosi $r^2 \approx 0.478$ świadczy o tym, że model nie jest idealnie dopasowany do danych. Jeżeli spojrzymy na wykres danych, zobaczymy, że są one dość rozproszone, co moim zdaniem może być powodem tego niedokładnego dopasowania.

Jeśli chodzi o residua, nie spełniają one wszystkich założeń, które powinny spełniać w poprawnie dobranym modelu. Spełniają warunek dotyczący średniej $E(\varepsilon_i) = 0$, nie spełniają jednak warunku związanego z wariancją $var(\varepsilon_i) = const$. Residua nie spełniają również w pełni warunku dotyczącego niezależności, ponieważ na wykresie autokorelacji znajduje się obserwacja, która nie jest niezależna od innych. Jeżeli chodzi o ostatni warunek dotyczący sprawdzenia normalności residuów, to testy statystyczne potwierdzają, że mają one rozkład normalny. Porównanie dystrybuant oraz histogramu z gęstością teoretyczną pokazuje, że nie pokrywają się one idealnie. Ostatecznie jednak można stwierdzić, że residua mają w przybliżeniu rozkład normalny. Nie spełniają jednak pozostałych założeń.

Biorąc pod uwagę analizę residuów oraz współczynnik determinacji stwierdzam, że model regresji liniowej nie jest najlepiej dobranym modelem dla tych danych. Co świadczy o tym, że długość głowy oposa nie jest dokładnie liniowo zależna od długości jego ciała. Ponadto stwierdzam, że model można by lepiej pasować do danych, gdyby usunąć z nich obserwacje odstające. Jednak są to jednak jedynie moje spekulacje.

Bibliografia

[1] https://pl.wikipedia.org/wiki/Regresja_liniowa
[2] https://mfiles.pl/pl/index.php/Wsp%C3%B3%C5%82czynnik_determinacji
[3] https://pl.wikipedia.org/wiki/Test_Ko%C5%82mogorowa-Smirnowa
[4] https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test