# Analiza wieku Słuchotek

## 1. Wstęp

Celem jest zrobienie projektu na danych, których współczynniki korelacji są duże $(0.8, 1)$. Taka trudność pojawia się często w problemach w biologii. Wybraliśmy bazę danych *Abalone*, jest to baza danych opisująca Słuchotki (mięczaki). Problemem będzie znalezienie wieku "mięczaków". W tym celu będziemy estymować liczbę "pierścieni skorupy" od reszty cech.

### 1.1. Opis danych

Nasz zbiór danych opisuje pomiary poszczególnych cech słuchotek. Istotną dla informacji z "punktu widzenia" biologii jest, że uchotki mają kształt okrągły albo owalny, zatem powinniśmy się spodziewać dużej korelacji pomiędzy *Długością* oraz *Średnicą*. Istotne też dla nas będą okresy rozwoju uchotek, ponieważ dojrzałość płciową osiągają dopiero po około 4 latach. **Opis danych**

- *Płeć* - zmienna kategoryczna opisująca płeć osobnika. Możliwe wartości: M - samiec, F - samica, I - niemowlę.
- *Długość* - odległość między dwoma najdalszymi punktami na osi poziomej muszli (mm).
- *Średnica* - szerokość muszli mierzona pod kątem prostym do jej długości (mm).
- *Wysokość* - wysokość muszli z mięsem w środku (mm).
- *Cała_waga* - waga całkowita muszli z mięsem (g).
- *Waga_po_obraniu* - waga mięsa po wyjęciu go z muszli, bez żadnych innych części (g).
- *Waga_trzewi* - waga narządów wewnętrznych, które zostają po wykrwawieniu (g).
- *Waga_powłoki* - waga samej muszli po wysuszeniu (g).
- *Pierścienie* - liczba pierścieni widocznych na muszli.

## 2. Koliniowość ze względu na płeć

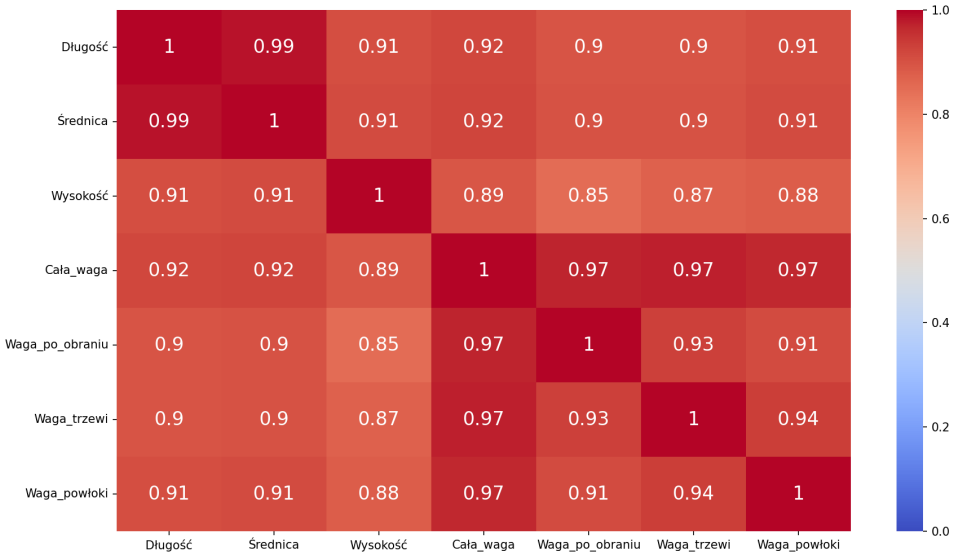
### 2.1. Wstęp

Chcemy sprawdzić, czy zmienna *Płeć* ma istotny wpływ na wartości innych zmiennych objaśniających, ponieważ rozwój osobników może przebiegać odmiennie w zależności od płci (zarówno pod względem wielkości, jak i tempa wzrostu).

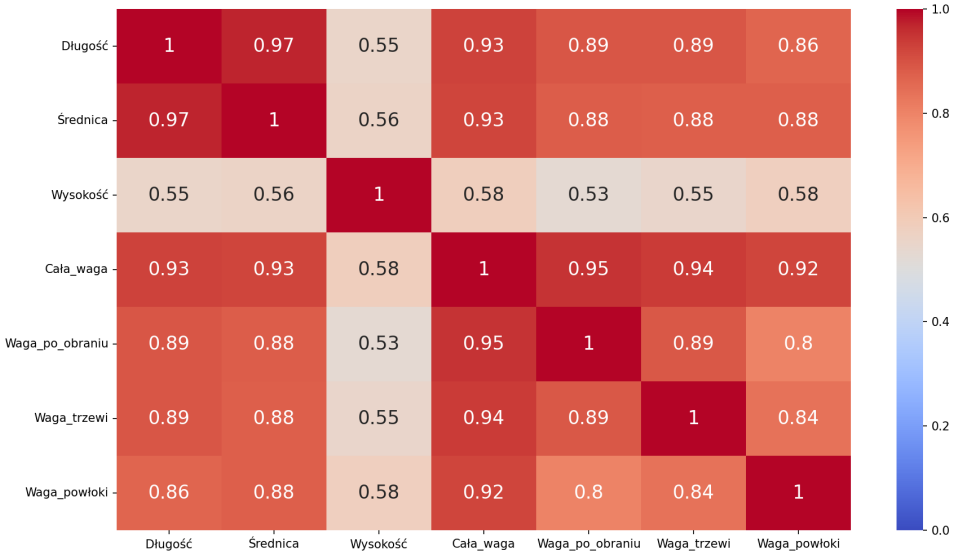
### 2.2. Wnioski

Na podstawie zaprezentowanych macierzy korelacji (Rysunki:1, 2, 3) dla różnych płci, możemy zauważyć, że w przypadku niemowląt obserwujemy bardzo silne korelacje pomiędzy większością zmiennych, a u samic obserwujemy słabsze związki między niektórymi cechami (np. wysokość vs długość). W związku z tym zdecydowaliśmy się stworzyć osobny model dla każdej z płci, *Model_Sex_F*, *Model_Sex_M*, *Model_Sex_I*.

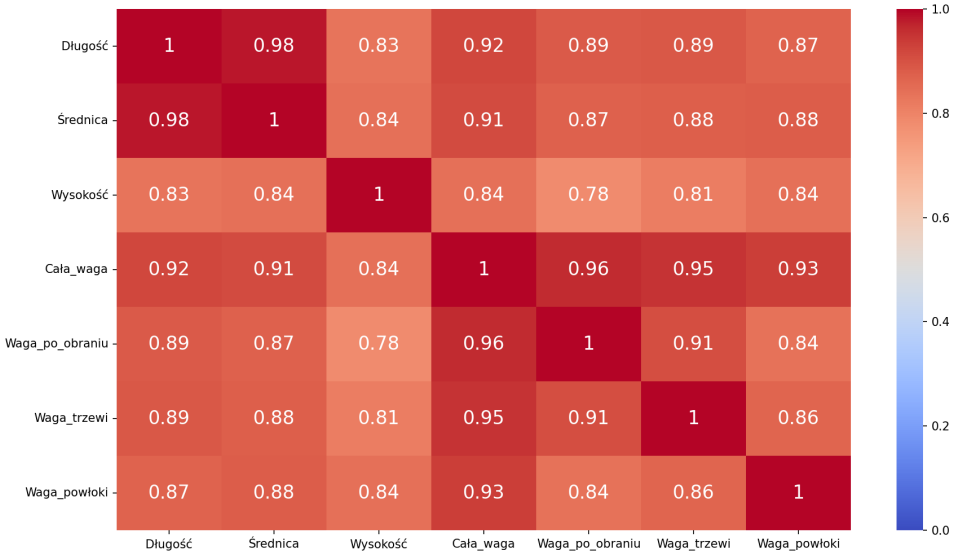
## 2.3. Opracowanie wyników




Rysunek 1: Korelacja - Płeć I




Rysunek 2: Korelacja - Płeć F




Rysunek 3: Korelacja - Płeć M

# 3. Dobór zmiennych objaśniających

## 3.1. Wstęp

Celem tego etapu jest wybór odpowiednich zmiennych objaśniających do każdego z modeli regresji liniowej dla poszczególnej płci, które najlepiej tłumaczą zmienną objaśnianą. Stworzyliśmy dodatkowe zmienne: stosunki mas, wymiarów oraz stosunek masy do objętości, gdzie objętość osobnika przybliżyliśmy przez iloczyn długości, wysokości i szerokości.

## 3.2. Metodologia

Postępowanie przeprowadzamy dla każdego modelu.

1. Sprawdzamy wykresy QQ reszt modelu oraz Wykres reszt vs. przewidywanych wartości oraz przeprowadzamy testy: Normalności (Shapiro-Wilk), homoskedastyczności (Breuscha-Pagana).
2. Przeprowadzamy analize współliniowości (Vif) dla zmiennych objaśniających, odrzucamy te zmienne dla których współczynnik Vif jest większy od 10
3. Przeprowadzamy testy permutacyjne F dla nowo stworzonych zmiennych przy ilości permutacji równej 100,000.
4. Odrzucamy $H_0$ : zmienna nie wnosi istotnej różnicy na poziomie istotności $\alpha = 0.05$
5. Przeprowadzamy ponownie analizę współlinowości (Vif) naszego modelu i odrzucamy najwyższą wartość dopóki jest ona większa od Vif równego 12.
6. Porównujemy $R^2$ naszego modelu ze wszystkimi zmiennymi (bazowe oraz dodatkowe zmienne) oraz $R^2$ modelu ze zredukowaną ilością zmiennych (końcowe proponowane).

## 3.3. Opracowanie wyników

### 3.3.1 Ogólne

Tabela 1: Wartości *p-value* testów normalności i homoskedastyczności

| **Model** | **Test Shapiro–Wilka** | **Test Breuscha–Pagana** |
|---|---|---|
| Model_Sex_F | $2.193e{-23}$ | $1.601e{-19}$ |
| Model_Sex_M | $2.311e{-23}$ | $2.953e{-28}$ |
| Model_Sex_I | $2.887e{-30}$ | $6.007e{-23}$ |

### 3.3.2 Model_Sex_I

Tabela 2: Vify - model podstawowy

| **NAME_INDEX** | **Długość** | **Średnica** | **Wysokość** | **Cała_waga** | **Waga_po_obraniu** | **Waga_trzewi** | **Waga_powłoki** |
|---|---|---|---|---|---|---|---|
| **Model bazowy** | 37.54204416640646 | 40.43 | 7.50 | 102.49 | 23.21 | 20.81 | 25.08 |
| **Model bazowy po redukcji** | | | 5.34 | | 6.57 | | 9.04 |

Tabela 3: F test wyniki

| **Tested feature** | **F statistic** | **P_statistic** | **P value** | **Odrzucenie $H_0$ dla $\alpha = 0.05$** |
|---|---|---|---|---|
| **Długość/Wysokość** | 1.9 | 16732 | 0.17 | False |
| **Masa/objętość** | 4.99 | 2632 | 0.03 | True |
| **Waga_powłoki/Cała_waga** | 2.57 | 10659 | 0.11 | False |
| **Waga_po_obraniu/Cała_waga** | 1.64 | 15941 | 0.16 | False |
| **Waga_trzewi/Cała_waga** | 1.81 | 17660 | 0.18 | False |
| **Waga_powłoki/Waga_po_obraniu** | 4.29 | 3826 | 0.04 | True |
| **Waga_trzewi/Waga_po_obraniu** | 0.52 | 47014 | 0.47 | False |
| **Waga_trzewi/Waga_powłoki** | 0.06 | 80600 | 0.81 | False |

### 3.3.3 Model_Sex_F

Tabela 4: Vify - model podstawowy

| NAME_INDEX | Długość | Średnica | Wysokość | Cała_waga | Waga_po_obraniu | Waga_trzewi | Waga_powłoki |
|---|---|---|---|---|---|---|---|
| **Model bazowy** | 20.64 | 19.57 | 1.55 | 63.39 | 18.13 | 9.8 | 13.02 |
| **Model bazowy po redukcji** | | 7.29 | 1.55 | | 6.02 | 6.49 | 4.92 |

Tabela 5: F test wyniki

| Tested feature | F statistic | P_statistic | P value | Odrzucenie $H_0$ dla $\alpha = 0.05$ |
|---|---|---|---|---|
| **Długość/Wysokość** | 2.2007 | 11516 | 0.11517 | False |
| **Masa/objętość** | 0.35558 | 51252 | 0.51252 | False |
| **Waga_powłoki/Cała_waga** | 0.04639 | 83287 | 0.83287 | False |
| **Waga_po_obraniu/Cała_waga** | 137.94333 | 0 | 1e-05 | True |
| **Waga_trzewi/Cała_waga** | 33.58466 | 0 | 1e-05 | True |
| **Waga_powłoki/Waga_po_obraniu** | 91.54663 | 0 | 1e-05 | True |
| **Waga_trzewi/Waga_po_obraniu** | 26.58653 | 0 | 1e-05 | True |
| **Waga_trzewi/Waga_powłoki** | 4.01512 | 4506 | 0.04507 | True |

### 3.3.4 Model_Sex_M

Tabela 6: Vify - model podstawowy

| NAME_INDEX | Długość | Średnica | Wysokość | Cała_waga | Waga_po_obraniu | Waga_trzewi | Waga_powłoki |
|---|---|---|---|---|---|---|---|
| **Model bazowy** | 31.52 | 30.62 | 4.18 | 84.74 | 23.38 | 13.04 | 16.08 |
| **Model bazowy po redukcji** | | 6.9 | 4.17 | | 6.82 | 7.76 | 5.94 |

Tabela 7: F test wyniki

| Tested feature | F statistic | P_statistic | P value | Odrzucenie $H_0$ dla $\alpha = 0.05$ |
|---|---|---|---|---|
| **Długość/Wysokość** | 0.04972 | 82194 | 0.82194 | False |
| **Masa/objętość** | 32.83583 | 0 | 1e-05 | True |
| **Waga_powłoki/Cała_waga** | 0.00855 | 92656 | 0.92656 | False |
| **Waga_po_obraniu/Cała_waga** | 117.1675 | 0 | 1e-05 | True |
| **Waga_trzewi/Cała_waga** | 19.88455 | 0 | 1e-05 | True |
| **Waga_powłoki/Waga_po_obraniu** | 54.98979 | 0 | 1e-05 | True |
| **Waga_trzewi/Waga_po_obraniu** | 25.9749 | 0 | 1e-05 | True |
| **Waga_trzewi/Waga_powłoki** | 2.54613 | 11085 | 0.11086 | False |

Tabela 8: Porównanie modeli regresji dla poszczególnych płci przed i po redukcji zmiennych

| Model | $R^2$ przed redukcją | $R^2$ po redukcji | Ostateczne zmienne w modelu |
|---|---|---|---|
| Model_Sex_F | 0.4018 | 0.4016 | *Średnica*, *Wysokość*, *Waga_powłoki*, *Waga_po_obraniu/Cała_waga*, *Waga_trzewi/Cała_waga*, *Waga_trzewi/Waga_powłoki* |
| Model_Sex_M | 0.4761 | 0.4747 | *Średnica*, *Wysokość*, *Waga_po_obraniu*, *Masa/objętość*, *Waga_po_obraniu/Cała_waga*, *Waga_trzewi/Cała_waga*, *Waga_powłoki/Waga_po_obraniu* |
| Model_Sex_I | 0.6163 | 0.6013 | *Wysokość*, *Waga_po_obraniu*, *Masa/Objętość*, *Waga_powłoki/Waga_po_obraniu* |

## 3.4. Wnioski i dyskusja

### 3.4.1 Normalność

W żadnym z analizowanych modeli reszty nie mają rozkładu normalnego, co potwierdziły wyniki testu Shapiro–Wilka oraz nie mają spełnionego założenia o homoskedastyczności - test Breuscha–Pagana (Tabela: 1).

### 3.4.2 Analiza współliniowości - model podstawowy

W każdym z modeli po analizie współliniowości modelu podstawowego odrzuciliśmy zmienne *Cała_waga*, *Długość*. Dodatkowo w modelu *Model_Sex_I* odrzuciliśmy *Średnica*, *Waga_trzewi* (Tabele: 2, 4, 6)

### 3.4.3 Permutacyjny test F

Ze względu na niespełnienie założenia o normalności rozkładu reszt (potwierdzone testem Shapiro-Wilka Tabela 1), zastosowanie klasycznego testu F nie było możliwe. W tej sytuacji wybraliśmy permutacyjny test F, który nie wymaga założenia o normalności i pozwala na wiarygodną weryfikację istotności statystycznej nowo wprowadzanych zmiennych objaśniających. W poniższych wnioskach korzystaliśmy z wyników z kodu oraz tabel: 3, 5, 7.

**Model_Sex_I** W modelu dla niedojrzałych osobników po przeprowadzeniu testu permutacyjnego na poziomie istotności $\alpha = 0{,}05$ odrzucamy zmienne: *Długość/Wysokość*, *Waga_trzewi/Waga_powłoki*, *Waga_trzewi/Cała_waga*, *Waga_po_obraniu/Cała_waga*, *Waga_trzewi/Waga_po_obraniu* oraz *Waga_powłoki/Cała_waga*. Następnie ze względu na wysoki współczynnik *vif*, usunięto zmienną *Waga_powłoki*.

**Model_Sex_F** W przypadku modelu zbudowanego dla samic po przeprowadzeniu testu permutacyjnego na poziomie istotności $\alpha = 0{,}05$ odrzucamy zmienne: *Długość/Wysokość*, *Masa/Objętość* oraz *Waga_powłoki/Cała_waga*. Dodatkowo ze względu na wysoki współczynnik *vif*, z modelu usunięto zmienne *Waga_trzewi/Waga_po_obraniu*, *Waga_trzewi*, *Waga_po_obraniu* oraz *Waga_powłoki/Waga_po_obraniu*.

**Model_Sex_M** Dla modelu dla samców po przeprowadzeniu testu permutacyjnego na poziomie istotności $\alpha = 0{,}05$ odrzucamy zmienne: *Długość/Wysokość*, *Waga_trzewi/Waga_powłoki* oraz *Waga_powłoki/Cała_waga*. Następnie ze względu na wysoki współczynnik *vif*, z modelu usunięto zmienne *Waga_trzewi*, *Waga_powłoki* oraz *Waga_trzewi/Waga_po_obraniu*.

### 3.4.4 Poprawność wykonania redukcji

Aby zweryfikować poprawność zastosowanej metodologii redukcji zmiennych, porównaliśmy współczynniki determinacji $R^2$ dla modeli zawierających wszystkie zmienne oraz modeli ze zredukowaną liczbą zmiennych (Tabela 8). Zaobserwowane różnice w wartościach $R^2$ są nieznaczne, co potwierdza, że proces redukcji nie wpłynął istotnie na jakość modeli. Możemy zatem uznać, że zastosowana metodologia selekcji zmiennych jest poprawna.