# Modelowanie Matematyczne - Lab 4

Adam Przemysław Chojecki

04 czerwca 2024

## Punktacja

- Zadanie 1: 3 pkt
- Zadanie 2: 1 pkt
- Czystość kodu i zrozumiałe komentarze: 1 pkt

## Zadanie 1: Regresja logistyczna śmierci na Titanicu

W tym zadaniu będziemy modelować przeżycie pasażerów ze statku Titanic. Zbadamy, jaki był wpływ ich klasy na przeżycie. Użyjemy regresji logistycznej oraz testu Walda do oceny istotności zmiennych.

1. Wczytaj zbiór danych `Titanic` używając funkcji `data(Titanic)`.
2. Przekształć zbiór danych `Titanic` do ramki danych (`data.frame`), używając funkcji `as.data.frame(Titanic)`.
3. Wyświetl pierwsze kilka wierszy danych, aby zapoznać się ze strukturą zbioru, używając funkcji `head()`.
4. Chcemy zamodelować zależność zmiennej `Survived` od zmiennych `Class`, `Sex` oraz `Age`.
   Jak interpretować zmienną `Freq`?
   Poczytaj w `?glm` czym jest parametr `weights`.
5. Dopasuj model regresji logistycznej, gdzie na kolumnie `Class` wykonany będzie one-hot encoding (czyli domyślnie):
   `glm(... ~ ..., data = ..., family = binomial, weights = ...)`
6. Przeanalizuj model. W jakiej klasie miało się największą szansę przeżycia?
7. Zwróć uwagę, że interpretacja zmiennej `Class` ma w sobie relację porządku (`1st` > `2nd` > `3rd`), której nie ma explicite wpisanej w dane.
   Zastanów się, gdzie w tym porządku miałoby sens wpisać `Crew`.

8. Zbuduj własny encoding dla kolumny `Class`. W tym celu dodaj do danych 3 kolumny. Jeśli obserwacja jest pierwsza w porządku, to wszystkie 3 kolumny mają `0`. Jeśli jest druga w porządku, to pierwsza z kolumn ma wartość `1`, a pozostałe `0`. Jeśli obserwacja jest trzecia w prządku, to pierwsze dwie kolumny mają `1`, a ostatnia ma `0`. Jeśli zmienna jest ostatnia w porządku, to wszystkie 3 kolumny mają wartość `1`.

9. Ponownie dopasuj model logistyczny, tym razem ze swoim encodingiem (czyli zamiast kolumny `Class` użyjemy naszych nowych trzech kolumn).

10. Zinterpretuj dopasowany model. W jakiej teraz klasie model przewiduje największą szansę przeżycia?

11. Przeanalizuj wyniki testu Walda dla zmiennej `(Intercept)`. Czy jest ona istotna w modelu? Jak to interpretować?

12. Jeśli jest nieistotna, to dopasuj model bez interceptu:
`glm(... ~ ... - 1, data = ..., family = binomial, weights = ...)`
Zwróć uwagę, że wszyskie zmienne muszą być zmiennymi numerycznymi `0-1`. Nie mogą one być `TRUE-FALSE`, ani `1-2`.

## Zadanie 2: Klasyfikacja gatunków irysów

W tym zadaniu wykorzystamy regresję logistyczną do klasyfikacji gatunków irysów na podstawie ich płatków.

1. Przeanalizuj strukturę danych i zidentyfikuj zmienne, które zostaną wykorzystane jako zmienne objaśniające i zmienna odpowiedzi.

2. Zauważ, że kolumna `Species` ma 3 rodzaje. Połącz klasy `versicolor` oraz `virginica`.

3. Dopasuj model regresji logistycznej do klasyfikacji gatunków irysów
`glm(... ~ ., data = iris, family = binomial)`
Czy model dopasował się poprawnie?

4. Wywołaj `summary()` i zwróć uwagę na wyniki testów Walda. Czy wynik jest podejrzany?

5. Identyfikuj, co było problemem w dopasowaniu modelu.

6. Zbuduj inny model w którym będziemy przewidywać `versicolor` vs `virginica`.