# Exercises Day 4

**PSY8003**

Matthias Mittner

spring 2022

## Exercise 1: Logistic regression

Use the `titanic.dta` dataset from the lecture. In the lecture, we have discussed the effect of `Age` and `Sex` on the probability to survive the Titanic disaster ("women and children first!"). In this exercise, use logistic regression to investigate the effect of the price someone paid (`Fare`) as well as the Ticket class (`Pclass`) on the probability of survival. `Pclass` is a categorical variable with levels `1`=first class, `2`=second class and `3`=third class.

- Fit two models where you predict survival using only a single predictor (`Pclass` and `Fare`)
- what are the results?
- fit a model with both predictors at once and compare it to the models with only one predictor
- what does the result indicate?
- finally, add the two variables to a model that contains `Age` and `Sex` and their IA. Do the two variables add to the model-fit? Can you interpret the model coefficients (or their OR)?

## Exercise 2: Logistic regression for classification

In this exercise, we will use the dataset `spam.dta` which contains data from a spam-filter.

The main variable here is `isspam` which indicates if the email is spam or not. The other variables are explanatory variables. They are:

- `crltot` - total length of words that are in capitals
- `dollar` - frequency of the $ symbol, as percentage of all characters
- `bang` - freqency of the ! symbol, as a percentage of all characters,
- `money` - freqency of the word money, as a percentage of all words,

- `n000` - freqency of the text string 000, as percentage of all words,
- `make` - freqency of the word make, as a percentage of all words.

The goal is mainly to predict whether a future email is spam or not based on these explanatory variables.

- Run a logistic regression with `isspam` as the outcome variable. Which of the variables are important?
- Determine the OR for the different variables. What is the difference between the coefficients and the OR? What do the ORs indicate?
- Which of the effect seems to be strongest?
- Predict the probability of an email being spam based on the occurrence of the word "money" and make a curve for the predicted probabilities.
- Use the result of the logistic regression model to classify each of the emails as either "spam" or "no spam". Look at the confusion matrix and judge how well you can make a correct decision based on this model.
- Can you comment on the observed accuracy? Do you trust this value?

## Exercise 3: Poisson regression

Take a closer look at the `affairs.dta` dataset. Focus on the variables

- `affairs` - number of affairs
- `gender` - 1=male, 2=female
- `age`
- `yearsmarried`
- `religiousness`
- `education`

Use Poisson models to investigate how the predictors are related to the number of affairs. Follow your interest! Some thinks you could study:

- is there a gender difference in how many affairs some may have?
- does religiousness play a strong role? Is such an effect the same across genders?
- ...

For your favourite model, do a test for overdispersion. Is overdispersion a problem here? If yes, fit a negative binomial model. Do the coefficients change a lot?