Probability & Statistics

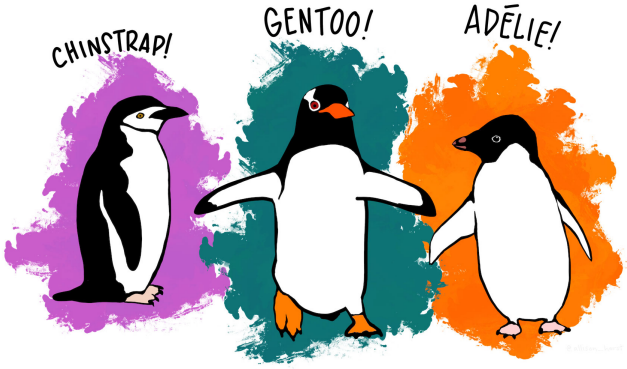
Project 2: Linear Regression

Apr. 2022

Contents

1 Data Description

The dataset `penguins` is accessible by installing the package `palmerpenguins` contains 344 observations and size measurements for 3 penguin species observed on 3 islands in the Palmer Archipelago.

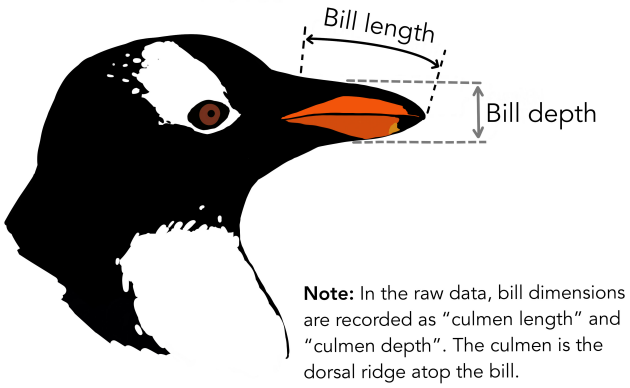


The following variables are included in the dataset:

- `species`: Penguin species (Adélie, Chinstrap and Gentoo),
- `island`: An island in Palmer Archipelago, Antarctica (Biscoe, Dream or Torgersen),

- `bill_length_mm`: Bill length (in mm) (for more information, please refer to the reference image),
- `bill_depth_mm`: Bill depth (in mm) (for more information, please refer to the reference image),
- `flipper_length_mm`: Flipper length (in mm),
- `body_mass_g`: Body mass (in g),
- `sex`: Penguin sex (female, male),
- `year`: Year.

Bill dimensions The culmen is the upper ridge of a bird's bill. In the simplified penguins data, culmen length and depth are renamed as variables `bill_length_mm` and `bill_depth_mm` to be more intuitive. For this penguin data, the culmen (bill) length and depth are measured as shown below



2 Research question

Th goal of this study is to estimate the weight of an individual penguin though a multiple linear regression model which considers features such as the bill length, bill depth and other variables.

3 Instructions

Your report should contain the following two sections:

*** Please note that the significance level should be considered 5% for all parts.**

1. **Statistical analyses:**
   (a) Perform an **explanatory analysis** on the data. Investigate the relation between the predictor variables and the outcome `body_mass_g`, taking into account the different species encoded in `Species`.
   (b) Build the optimal **regression model** (using *forward model building*) which describes the relationship between `body_mass_g` and the other predictors. Consider `bill_length_mm`,

`bill_depth_mm`, and `flipper_length_mm` as candidates for the independent variables. Check quadratic or higher orders effect, and interaction terms if it is necessary.

(c) Check the **assumptions** of your chosen model.

(d) **Add the variables** `species`, `sex` and `island` as predictors to your final model. Is the effect significant? What can you conclude?

(e) Check the **assumptions** of your chosen model. Compare the **assumptions** of the chosen model in step (c) with this one. Discuss your findings on whether the model including categorical variables fits the assumptions better or not?

(f) Check if there are **outliers or influential observations**.

(g) Estimate the body mass for a penguin with following measures:

- **Species**: Adelie
- **Island**: Dream
- **Bill length**: 36
- **Bill depth**: 17.5
- **Flipper length**: 188
- **Sex**: female

Interpret the result. Do you think that the result is reasonable?

2. **Results:**

- Describe the results as if you would write the results section of a paper.
- Create a table containing an overview of the selected regression coefficients in each of the successive models, including their t-value and p-value.
- Create a table for the successive models. Include, for each model, the R^2-value and Residual Sum of Errors (RSE).
- Include a figure inspecting the relation between the predicted and measured weight values.

4 Guidelines

Write the report in an **R Mark Down (.rmd)** file. Structure your report well, with headers of different levels for the sections and research questions. There is a specific syntax in R Markdown to write headers, bullet points, numbered items and tables. Please use this. You can find a complete guide on R Markdown **here**. A cheatsheet summarizes the most important syntax: `https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf`. Use TEX syntax inside R Markdown for mathematical notations. During your work, regularly convert your .rmd file to HTML to avoid unexpected errors in the end.

Feel free to use R functions which haven't been used in the practicals, for example to make nicer graphs. Supplement graphs with informative main titles and axis titles with variable units.

5 Practical information

- **Deadline:** The project can be handed in until **Friday, May 6**. The project is an **obligatory part** of the exam and will count for 10% of the final score. If no (or no decent) project is handed in, the maximal obtainable grade for the entire course is 7/20.
- **Groups:** Work is done in groups of **3 or 4 persons**, as chosen on Ufora. Everyone in a group gets the same score.
- **Submission:** Each group should submit an **RMarkDown file**, and a corresponding **HTML file**. Both files have to be included in a **zip-file** and submitted through Ufora. The names of the uploaded files must have the following structure: `Group_X_project2`, where `X` corresponds to your group number.
- Do not forget to mention the names of the group members in your project.

Good luck!