Exercises Day 1

PSY8003

Matthias Mittner

spring 2022

Preparation

Make sure your statistical software is ready.

1. Make sure the software is installed
    - Stata: https://i.ntnu.no/wiki/-/wiki/Norsk/Stata
    - R/RStudio: https://cran.r-project.org/ and https://www.rstudio.com/
2. Create a new folder for this course, e.g., `psy-8003_exercises`
3. Copy the file `demo1.R` or `demo1.do` into that folder
4. From within your statistical software, change the "current working directory" to your newly created directory
    - Stata: `File` -> `Change working directory`
    - RStudio: `Session` -> `Set Working Directory`
5. Make sure that you can run the demo file
    - Stata: `Open` -> select the file -> Click "Do"
    - RStudio: Open in "Files" pane -> click "Source"
6. Create a script-file for the solution for this exercise sheet
    - Stata: `File` -> `New` -> `Do file`: create a file called `exercise1.do` in your folder
    - RStudio: `File` -> `New File` -> `R script`: create a file called `exercise1.R` in your folder

Exercise 1: Play the correlation game

This exercise is supposed to build your intuition about the strength of a correlation so that you can quickly and accurately judge a correlation by looking at a scatter plot.

- visit the webpage http://guessthecorrelation.com/
- play until you have (at least) a high-score of 100 points
- preferable, team up one-on-one with someone from your group and use the two-player option to challenge one another; setup two-player-mode:
  - both players must have game website open. one player enters the username of the opponent, that user is then prompted to start a game.
  - closest guess to true correlation wins a coin. first player to 10 coins wins. no coins awarded on draw.

Exercise 2: Simple regression

Open the dataset `loenn.dta` (in Stata, import the file; in RStudio, use "File" -> "Import data" or `haven::read_dta()` to read the file).

This file contains the yearly salary of different persons along with other variables:

- `erfaring`: work experience in years
- `utdann`: education in years
- `kvinne`: is the person female (`=1`) or not (`=0`)
- `fagfor`: is the person organized?
- `alder`: age of the person in years
- `gift`: is the person married?
- `kvierf`: interaction term between `kvinne` and `erfaring` (please ignore for now)

Run a few bivariate regressions with `loenn` (salary) as the outcome variable, using one of the other variables as the predictor. Interpret the result of the intercept and the slope coefficient as well as the global fit measures (standard deviation of residuals, R2). Which of the models you tested gives the best fit?

Exercise 3: Multiple regression

Using the dataset from exercise 2, run a few multiple regression models (i.e., input several of the variables as predictors to predict dependent variable `loenn`). Interpret the coefficients (be careful to include the conditioning on all the variables that you included!).

Can you predict the salary of a male person that is 40 years old, married and organized in a labour union?

Exercise 4: Brain weight and total sleep across species

Is the size of the brain across different species associated with the number of hours they sleep during the night?

- download dataset `total_sleep.dta` from Blackboard and import it into your software
- the dataset contains data about different species
- run correlations between the variables in the datase:
  - brain weight (`BrainWt`)
  - body weight (`Bodywt`)
  - and number of hours sleep per night (`TotalSleep`)
- run two separate linear regression analyses with `TotalSleep` as dependent variable and each of the other two as predictor
- can you tell the effect of the two variables apart?
- calculate the pairwise correlations for the three variables
  - can you spot a potential problem?

Exercise 5: Random data

The dataset `random_data.dta` contains one hundred variables `x1`, …, `x100` that contain 20 completely random data-points (taken from a standard normal distribution). There is no relation at all between any of the variables. One way to think about this data is therefore, that the null-hypothesis is true for a regression model/correlation analysis between any variables in this dataset.

In your software,

- load the dataset
- calculate 20 random regression models between any two variables (e.g., run `regress x4 x20` and repeat for different `x` variables) and note down the p-values for the slope-coefficient
- how often did you get a significant result?

Next,

- calculate all pairwise correlations
- by design, there is no underlying correlation between any of the variables
  - is that fact reflected in the results of the correlation matrix?