Exercises Day 2

PSY8003

Matthias Mittner

spring 2022

Exercise 1: Hierarchical regression

Use the datafile `workout.dta` that contains data from people using a gym. This file contains the following variables:

- `whours`: number of hours per months that a person trains in the gym
- `gender`: woman=0, man=1
- `age`: in years
- `educ`: education (levels 1=`secondary/high`, 2=`university`, 3=`more than university`)
- `marital`: single=1, married=0
- `health`: response to "how important is your health to you?" (1=not important to 6=very important)

Run a hierarchical regression analysis where you try to predict how many hours of workout someone does.

- Start by including only `gender` and `age` as predictor variables; is this model significant?
- add `educ` and `marital` in a second block and compare it to the previous model. Is the change significant?
- finally, add the `health` variable in a final step and determine wether the increase in R^2 was significant
- do any of the coefficients change when including additional variables?

Exercise 2: Power analysis

For this exercise, you will most likely want to download the free software G*Power from http://www.gpower.hhu.de/en.html. If you are an R-user, you can also use the `pwr` package https://cran.r-project.org/web/packages/pwr/index.html and, more specifically, the function `pwr.f2.test()`. Stata also has features for power analysis (https://www.stata.com/features/power-and-sample-size/).

Imagine that you are planning a correlational study were you want to investigate the impact of the dose of a certain drug on depressive symptoms. You measure depressive symptoms using "Becks depression inventory" (BDI). You know from previous research that `age`, `gender`, `previous medication` and the five variables of the big-5 personality traits (`openness`, `conscientiousness`, `extraversion`, `agreeableness`, and `neuroticism`) collected with the questionnaire NEO-FFI) affect BDI scores.

a) Reading the relevant literature on the subject, you conclude that the amount of explained variance R^2 in similar studies varied a lot between 0.1 to 0.6. What sample size would you need to collect to find the minimal $R^2 = 0.1$ with power of 0.95 and $\alpha = 0.05$?

b) You feel that the determined number of participants is too high and you want to make a tradeoff: Reduce the number of participants which will reduce the minimum effect-size that you can find with high probability (power). Make a graph of Total sample-size against minimum detectable effect-size f^2 for an $\alpha = 0.05$ and a power of 0.95. From this graph, how many participants would be required for an overall effect of $R^2 = 0.2$?

c) You can maximally invest $N = 150$ subjects. From the plot from the previous exercise, what is the corresponding R^2? You can go back from f^2 to R^2 using

$$R^2 = \frac{f^2}{f^2 + 1}$$

d) Finally, you want to investigate the amount of explained variance of the drug over and above that explained by the other variables. You want to be able to detect a small effect of $f^2 = 0.02$ (small effect according to Cohen's conventions) of drug dose with $\alpha = 0.05$ and power of 0.8. What is the required sample size?

Exercise 3: Regression diagnostics

Using the dataset from exercise 1 (`workout.dta`), investigate the validity of the final regression model that included all the predictors.

- are there any datapoints with very large residuals?
- if yes, are those the same datapoints that have highest leverage or Cook's distance?
- is multi-collinearity a problem in this dataset?

- is the outcome variable `whours` adequatly modeled as a linear function of the predictors?
- is heteroscedasticity a problem?
- are the residuals normally distributed?