Part I: Simple linear regression

For ‘Part 1’ of the assignment you will need to download and import ***Data1***.

**1. Considering simple linear regression:**

a. Why is a case/data point/individual/observation that has high *leverage* not necessarily *influential*?

**2. Data1 contains (mostly) financial information of a sample of publicly listed manufacturing firms.**

a. What is the correlation between the following two variables: R&D expenditure and Firm size (measured as total assets)? Interpret the result.

b. Show for one of the firms (just randomly pick one of the firms within the sample) to what extent this firm over- or underinvests in research & development in comparison to the average level of investment in research & development.

c. Use the data from the entire sample and develop a simple linear regression model with respect to the extent that R&D expenditure is predicted by Earnings before interest and taxes (EBIT). Discuss and interpret the (relevant) results.

d. Check whether the independent variable - with respect to the model estimated in question 1.2.c - [*EBIT*] follows a normal distribution. What is one thing you can do to ‘improve’ the distribution of a variable such that it more closely approximates a normal distribution? If neccessary, re-run the statistical model and highlight the most relevant differences.

*Since we ran simple regression models (in Q1.2.c-d), we now have both observed and predicted values for R&D expenditure.*

e. Inspect whether the firm you selected for question 1.2.b actually over- or underinvests in research & development, given their EBIT [hint: think about residuals].

Part II: AN(C)OVA

For ‘Part II’ of the assignment you will need to download and import ***Data2***.

**1. There is a global shortage of microprocessors/computer chips. In order to improve its competitive position, you got hired as a business analyst by a large publicly listed (high-tech) manufacturing firm in the semiconductor industry. They have been struggling for quite some time to improve their production rate (amount of chips produced per day). Due to the ever increasing digital transformation of our society, the firm cannot keep up with the demand from their customers. Being able to improve the productivity now would secure many years of prosperity. The chief operating officer (COO) came up with the plan to buy and install three different types of machines (“Model-3”, “Model-X”, and the “Model-S”) in their production plants (equally distributed among the 90 production plants worldwide).**

a. You have been asked to help the firm figure out whether the most expensive machine (Model-S) significantly produces more chips than the other two machines AND whether there is a significant difference between the two ‘cheaper’ machines (Model-3 vs. Model-X). The COO would like to implement some changes (working with only one machine type) in order to be able to further streamline the production process in the future. However, she does not know which machine would be ‘the best’ pick. Analyse the data (from the data2 file) and provide a recommendation for the COO.

*A quality engineer approaches you and informs you about his concern/observation that the humidity may not only impact the productivity of the employees (damages productivity by making workers feel tired and lethargic) but also the production rate of the machines.*

b. Verify what the effect of the different types of machines on the amount of chips produced per day is when controlling for the humidity. Check whether the assumption of “independence of covariate and treatment” have been met. Discuss the relevant analyses/outcomes.

Part III: Multiple regression

In this third and final part of the assignment you will need to use the ***Data1*** file again.

**1. Considering multiple regression with an interaction: Building on the simple regression form part 1; you want to provide ‘a more complete representation of reality’ with respect to the amount of resources organizations allocate to R&D. In order to do so you decide to add variables to your statistical model.**

a. Use the data from the entire sample and develop a multiple linear regression model with respect to the extent that R&D expenditure is predicted by EBIT, controlled for Firm size. Discuss the results.

b. The dataset also contains information about the CEO - whether the CEO has a research orientation (1) or not (0). This variable is coded “1” if the person prior to becoming CEO holds a PhD degree in science or engineering OR the CEO has academic experience OR the CEO’s dominant functional experience is in R&D OR the CEO holds any patents. If none of these criteria are met, this variable is coded as “0”. Build on the previous model (from question 3.1.a) and develop a multiple linear regression model to check whether the extent to which EBIT influences spending on R&D is dependent upon whether the CEO has a research orientation or not. Discuss the results.

c. Several researchers indicate that financial slack, which refers to the amount of financial excess resources available to an organization during a given period, might impact the extent to which EBIT influences spending on R&D - i.e., they believe that when firms lack financial slack, they become more reluctant to allocate resources to R&D. For the financial slack measure you should use firms’ ‘current ratio’: measured as current assets divided by current liabilities. Build on the previous model (from question 3.1.a) and develop a multiple linear regression model to check whether the extent to which EBIT influences spending on R&D is dependent upon the level of financial slack [note that you do not have to include the CEO variable]. Discuss the results.