ERM ST 2019

Individual Assignment

- The sheet consists of two parts (A-B)
- You can find all necessary datasets on Moodle.
- All results are submitted via Moodle using the "Submit results of practice sheet" link. No other forms of submission will be accepted.
- You can change your results as long as you haven't submitted them.
- Please submit your results no later than Sunday, July 21st, 23:55.
- No late submissions will be accepted, so please make sure that you submit your results well before the deadline. And also make sure that you actually submit your results!
- All hints in this document refer to R. However you're also free to use any other statistical package as long as you can be sure to create an equally informative output for submission.
- R Markdown produces dynamic output formats in html, pdf, MS Word:
    - Once you have completed the assignment, create an entire document with all code and comment out "install.packages("XYZ")"!
    - Create output by pressing: Ctrl + Shift + k
    - Make sure to save the output as PDF (directly or indirectly via html/MS Word)
    - I will not open anything else but *.pdf files.
    - Do not try out R Markdown 5 minutes before the deadline. Make sure it works early on. Ask your peers if you face technical difficulties.
- Please use the comment function "#" to indicate which command/result refers to which Part and step. Written answers based on statistical results should be included in the script as comments.
- If plagiarism is found, neither the plagiarizing nor the plagiarized student will receive the bonus. Further disciplinary actions may follow.

PART A

Data: sheet1_partA.csv

A personnel department has to fill an internship vacancy. The underlying data for selecting the best candidate consists of information obtained from the CVs as well as knowledge acquired in an interview. As interviews consume time and money, the company considers cancelling the interview. Prior to this decision, you are asked to examine to what extent the interview contributes to forecasting the performance of a potential intern. You are provided with a sample of 66 former applicants.

Step 1: "Review of the Data Set"

A1: What problem do you see when you look at the values of the variables "SCHOOL" and "BACHELOR" denoting the final school and bachelor grades? Solve it by recoding in missing values.

Step 2: "Associations and Hypothesis"

The performance of an intern is determined through an assessment by his supervisor (PERFORMANCE).

A2: Describe the variables PERFORMANCE and SCHOOL by means of descriptive statistics and a histogram. Indicate the following quantities: median, mean, skewness, kurtosis, standard deviation, 25%-quantile, 50%-quantile and 75%-quantile. Is the variable PERFORMANCE left-skewed? Is the distribution of the variable PERFORMANCE flatter than a normal distribution? How is the appropriate case called?

A3: Formulate a correlation hypothesis (that makes sense given your task) between the dependent variable (assessment of performance by the supervisor) and one independent variable.

A4: Analyze the associations of the variables in a correlation table. Can the null hypothesis stating that there is no association between performance and number of previous internships, be rejected? Provide the correlation coefficient and p-value.

Step 3: “Regression Analysis”

In all models, PERFORMANCE is the dependent variable.

Model 1: Age in years (AGE) and sex (SEX)
Model 2: Model 1 + school leaving grade (SCHOOL) and bachelor grade (BACHELOR)
Model 3: Model 2 + total duration of previous stays abroad in months (ABROAD) and number of previously completed internships at home and abroad (INTERNSHIPS)
Model 4: Model 3 + performance in the interview (INTERVIEW)

Now, you have to include the independent variables of model 1-4 **one after the other** into the examination.

A5: Indicate the explained part of the dependent variable‘s total variance for each model. Indicate also the F-value. Does the model in general possess explanatory value?

Step 4: “Interpretation of the Regression Analysis Results”

A6: Indicate the standardized and non-standardized coefficients of model 4. Which of them has the strongest effect? How do you interpret the (non-standardized) coefficient of the variables AGE and SEX? Can you confirm your (directed) hypothesis on a significance level of 5% or 10%, respectively? Would you cancel the interview based on your empirical analysis?

PART B

Data: sheet1_partB.csv

You would like to examine how the interest Y of students in a subject (e.g. technology and innovation management) is correlated to the number of ECTS credits X obtained in this subject. You are provided with a survey among 100 students in order to clarify the correlation. It has been asked in the survey how many ECTS credits the students have already collected and how much they are interested in the subject.

Step 1: "Associations and Hypothesis"

B1: Formulate a correlation hypothesis (which is as useful as possible)! Analyze the correlation of the variables. Does the result support your hypothesis?

Step 2: "Visualizing the Data"

B2: First, create a scatter plot. Does it confirm your hypothesis?

Step 3: "Regression Analysis"

You would like to examine the association with linear and quadratic models. You select students' interest as dependent variable Y. Your independent variable X is the number of obtained ECTS credits at the beginning.

B3: First, perform a simple linear regression analysis. Interpret your results.

Now, calculate an additional variable X^2 containing the squared values of X.
Perform a multiple regression analysis with X and X^2 afterwards.

B4: Compare and interpret the results of both analyses. By how many percentage points did the ratio of explained and total variance increase? Why do need to be careful with interpretations when the explained variance rises after you have added a variable?