Final exam

Exam available:	February 11, 2021
Deadline to submit:	March 8, 2021 at 11:59 pm
Submission:	Upload to Moodle, in case of problems by email to: stefan.konigorski@hpi.de
To be submitted:	2 files: (i) a Word/pdf/html document containing only the requested analyses and results (i.e. results, tables and graphs) and their requested description/interpretation, and (ii) a file with the R code for calculating the results (R Markdown, with comments which R code belongs to which question). Clearly write to which question the output and the R code belong. Any extensive unnecessary and irrelevant computations can yield point deductions. Results can be given with 2 or 3 decimal places. To assess statistical significance in hypothesis testing, the significance level α=0.05 should be used.

Points:	Question 1:	5 points
	Question 2:	5 points
	Question 3:	30 points
	Question 4:	35 points
	Question 5:	15 points
	Total:	**90 points**

Background to the questions:

In the questions of this exam, different data analysis steps of an epidemiological study will be performed and R Markdown will be used for documentation and reporting of results. The main aim is to investigate blood levels of *HbA1c*, which is a blood biomarker informative of blood sugar levels and therefore an indicator of diabetes.

Variables that will be investigated for their association with HbA1c are *bmiB* (body mass index), *sex* (sex), *age2* (age), and the amount and frequency of eating chips/crackers (*fq44*, *fq44a*).

Question 1 - R Markdown [5 points]

As described on page 1, two files should be submitted: a Word/pdf/html document with explained results, and an Rmd file with the R code for the calculation of the results.

Create an R Markdown file containing all relevant R code (in R chunks) that was used to calculate the results. Also include text in this R Markdown script to answer all questions so that all the requested results of the analyses (i.e. results, tables and graphs) are included and described/interpreted. Then knit the R Markdown script to a Word/pdf/html document and submit these two files. [5 points]

Alternatively (if you have problems with knitting), a manually generated Word/pdf/html file with the explained results, and an Rmd file with the R code can be submitted. This means that no points can be obtained for question 1, but all other questions are unaffected.

Question 2 - Import, extract and save data [5 points]

a) Download the SPSS data file KiGGS03_06.sav from moodle and import it into R. [2 points]

b) Create a new dataframe in R named *kiggs*, which contains all variables (and only these) for the analysis (*HbA1c*, *bmiB*, *sex*, *age2*, *fq44*, *fq44a)*. [2 points]

c) Run the formatting steps in the provided Rmd file data_formatting.Rmd. Save this formatted dataframe on your computer as a RData file, e.g. on your desktop. [1 points]

Question 3 - Descriptive statistics and graphs [30 points]

a) Describe the variables *HbA1c*, *bmiB*, *sex*, *age2*, *fq44*, *fq44a* with regard to the following criteria:

- What is the scale (measurement level) of each of these 6 variables (nominal, ordinal, metric)? [3 points]
- For each variable, decide which descriptive statistic is best suited to describe it, and explain why. Available for selection: (i) Frequencies, (ii) mean/standard deviation. [3 points]
- Calculate the descriptive statistics that you have chosen and display them in a table [12 points]. You can also describe them in continuous text, but only reach maximally 9 of the 12 possible points.
- Also indicate how many missing values each variable has, and how many observations have complete data for all 6 variables (i.e. no missing values for any of the variables). [4 points]

b) For each variable, select whether a barplot or a histogram is more suitable for displaying their distribution. [2 points]

Then create these 6 diagrams using functions in the ggplot package. [6 points]. You can also create the diagrams using functions in base R, but then only maximally get 3 of the 6 points.

Note: If necessary, for this question and all further questions, transform factor variables to numeric variables.

Question 4 - Linear Regression [35 points]

Now, examine whether there is an association between the BMI and snacking habits of children with blood sugar levels, accounting for differences in the age and sex of children.

a) For this, calculate a linear regression, with *HbA1c* as outcome and the predictors *bmiB*, *fq44*, *fq44a, sex*, *age2*. [2 points]

   Check for each predictor, how you take it into the regression model (factor, ordinal or metric) and justify for each variable why you did it that way (e.g. because the variable has the measurement level xyz) [5 points].

b) To answer the question of whether the BMI and whether snacking habits of children have an influence on the blood sugar levels of children, adjusting for possible influencing factors, consider the significance tests of the regression coefficient(s) of *bmiB, fq44* and *fq44a* in this regression. Report the respective regression coefficients, interpret <u>each</u> coefficient (without considering whether the p-values are <0.05 or not), report their 95% confidence intervals, and report the p-values of the significance tests [10 points].

   What is your conclusion based on the p-values: Is there an association between BMI and *HbA1c*? Is there an association between *fq44* and *HbA1c*? Is there an association between *fq44a* and *HbA1c*? [3 Point]

c) Of the 7 assumptions of linear regression discussed in class, choose 5 assumptions. Investigate whether they are satisfied in the regression model, and answer for each assumption whether they are satisfied and why yes/no. [10 point]

d) Perform a 2-way ANOVA with *HbA1c* as outcome and the factors *fq44*, *fq44a* and interpret the results: Is there an association between *fq44* and *HbA1c*? Is there an association between *fq44a* and *HbA1c*? [4 points] Explain the difference from this analysis to the regression analysis above. [1 point]

Question 5 - Logistic Regression [15 points]

Now investigate the same question, but with *HbA1c* as a binary variable using logistic regression.

a) To do this, create a binary variable from *HbA1c* (low level vs. high level) as outcome for the analysis. [1 point]

b) Calculate a logistic regression, with binary *HbA1c* as outcome and the predictors *bmiB*, *sex*, *age2*, *fq44*, *fq44a*. [1 point]

c) To answer the question of whether the BMI and snacking habits of children have an influence on the blood sugar levels of children, adjusting for possible influencing factors, consider the significance test of the regression coefficient of *bmiB, fq44* and *fq44a* in this regression. Report p-values of the significance tests [3 points]. What is your conclusion based on the p-values: Is there an association between BMI and *HbA1c*? Is there an association between *fq44* and *HbA1c*? Is there an association between *fq44a* and *HbA1c*? [3 points]

d) For interpretation, consider odds ratios. Report the respective exponentiated regression coefficients and interpret each OR (without considering whether the p-values are <0.05 or not)[6 points].

e) Would you rather report results from the linear regression or from the logistic regression? Which model do you think makes more sense to you, and why? [1 Point]