Final exam

Exam available: February 1, 2020

Deadline to submit: February 24, 2020, 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 unrelevant 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:	10 points
Question 3:	20 points
Question 4:	30 points
Question 5:	10 points
Question 6:	15 points
Question 7:	10 points
Total:	**100 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 question is whether parents' smoking behaviour has an influence on their children's health.

Smoking was assessed by the four variables *E070M* (mother smoking), *E070V* (father smoking), *E072* (mother smoking in pregnancy), *E074* (mother smoking while breastfeeding).

Children's health was measured by the variables *arztZ01* (number of paediatrician visits), *kw100.e* (KINDL index of physical well-being) and *bmiB* (body mass index). Here, higher values in the KINDL index mean a better health.

As possible influencing factors, *sex* (sex), *age2* (age), *schichtz* (social class), *e0622* and *e0623* (frequency of doing sports in and outside of a club), and *e065z* (total sleep per day) will be included.

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 [10 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 for the analysis (*E070M*, *E070V*, *E072*, *E074*, *arztZ01*, *kw100.e, bmiB*, *sex*, *age2*, *schichtz*, *e0622*, *e0623, e065z)*. [3 points]

Questions 2c) – 2e) are based on this dataframe *kiggs*, but can also be done using the provided dataframe kiggs_finalexam.RData, if there are difficulties during importing.

c) Run the formatting steps in the provided Rmd file data_formatting.Rmd [1 point]
Explain in one sentence what these are doing. [1 point]

d) Save this formatted dataframe on your computer, e.g. on your desktop. [2 points]

e) Give the R command how to load this dataframe into R. [1 point]

Question 3 - Data transformations and data checks [20 points]

a) In order to avoid measuring smoking with 4 variables and to calculate all analyses with these 4 variables, we will combine them in one variable called *burdenS.* This new variable shall contain the total smoke exposure to which the children were exposed and is to be used in all further questions to measure parents' smoking. Carry out the following steps:

   - Check that the variables *E070M*, *E070V*, *E072*, *E074* are all factors. If they are not, transform them into factors. [3 points]
   - Set the value "has not breastfed" of variable *E074* to NA for all children. [1 point]
   - Delete this now empty factor level from the variable. [1 point]
   - Check whether these two steps worked as intended. [2 points]
   - Now calculate the new variable *burdenS* as the sum of the ranks of the four variables *E070M*, *E070V*, *E072*, *E074* for each person (i.e. sum of the numerical factor levels). [5 points]

   What is the meaning of this new variable, does a high value mean that the children were exposed to high levels of smoking, or that they were exposed to low levels of smoking? [1 point]

b) Generate a new dichotomous variable called *sport*, which describes whether a child does a lot or little sport, based on the two variables *e0622* and *e0623*. Think for yourself how to combine and transform these two variables in a sensible way to calculate the new variable. [4 points]

c) Add these variables *burdenS* and *sport* to the dataset kiggs, and save it in its updated form as an RData file (overwrite the previous file). [3 points]

In case of problems with the calculation of the variable *burdenS,* the variable *E072* (mother smoking in pregnancy) can be used alternatively in all further analyses. If there are problems with the calculation of the variable *sport,* the variable *e0622* can be used in question 6. There is no point deduction for this in the following questions.

Question 4 - Descriptive statistics and graphs [30 points]

a) Consider the variables *age2, sex, bmiB, arztZ01, kw100.e, burdenS* and describe them 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, or (iii) median/interquartile distance. [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 6 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. [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, transform factor variables to numeric variables.

Question 5 - Correlation [10 points]

Here, investigate whether parental smoking is associated with children's health.

a) Calculate the Pearson correlation coefficient of the generated smoking variable *burdenS* (or alternatively of the variable *E072*) with the variables *arztZ01*, *kw100.e* and *bmiB*. Perform 2-sided significance tests at the significance level of 0.05 to test whether the correlations are equal to 0. Give the three estimated correlation coefficients, their respective 95% confidence intervals and corresponding p-values [6 points].

b) Interpret the result of the significance tests for the correlation coefficients: Is *burdenS* associated with *arztZ01*? Is *burdenS* associated with *kw100.e*? Is *burdenS* associated with *bmiB*? [3 points]

c) Do you think that Pearson's correlation coefficient is an appropriate statistic here, or do you think that Spearman correlation coefficient would have been more appropriate? [1 point]

Note: If necessary, transform factor variables to numeric variables.

Question 6 - Linear Regression [15 points]

The final step is to examine whether there is an association between child health and smoking exposure, taking into account the possible influencing factors sex, age, social class, sport and sleep.

a) To do this, select either *arztZ01*, *kw100.e* or *bmiB* as outcome for the analysis, in order to measure the health of the children. Which variable did you choose, and why? [1 point]

b) Now calculate a linear regression, with this outcome and the predictors *burdenS*, *sex*, *age2*, *schichtz*, *sport* and *e065z*. [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) [3 points].

c) To answer the question of whether the smoking behavior of parents has an influence on the health of children, adjusting for possible influencing factors, consider the significance test of the regression coefficient of *burdenS* in this regression. Report the regression coefficient of *burdenS*, interpret the coefficient, report its 95% confidence interval, and its p-value of the significance test [4 points].
What is your conclusion: Is there an association or not? In which direction? [1 Point]

d) Has the conclusion changed (whether there is an association) compared to the conclusion from question 5 based on the correlation? [1 point]

e) Since there is evidence that individuals drawn from the same area are correlated with each other, but we are not interested in the effect of the area on the health ... what would be a suitable strategy for accounting for this correlation? [1 point]
Give the hypothetical R code of the model you could use for this analysis, if the area variable were *area*. [2 points]

Question 7 - Logistic Regression [10 points]

Investigate how the results look like if you investigate question 6 using logistic regression.

a) To do this, create a binary variable from *arztZ01* (0 visits vs. more than 0 visits) or *bmiB* (BMI lower than 20 vs. BMI higher or equal to 20) as outcome for the analysis. [2 points]

b) Now calculate a logistic regression with this outcome and the predictors *burdenS*, *sex*, *age2*, *schichtz*, *sport* and *e065z*. [2 points]

c) To answer the question of whether the smoking behavior of parents has an influence on the health of children, adjusting for possible influencing factors, consider the significance test of the regression coefficient of *burdenS* in this regression. Report the regression coefficient of *burdenS*, its 95% confidence interval, and the p-value of the significance test. Interpret the exponentiated regression coefficient, the 95% confidence interval, and the results of the hypothesis test. [6 points]