UNIVERSITY OF BAYREUTH

Data Analysis in R | Winter Term 25/26

Take-Home Exam

Due on 18 March 2026 at Noon

Word limit: 6,000 words

Part A

1. In your own words, explain the advantages and disadvantages of analysing observational data for making causal inferences.

2. Recall the example of the effect of daughters on voting on women's rights in Congress, which we discussed in class.

   2.a) What is the purpose of including control variables in regression models?

   2.b) Would you want to add as many control variables as possible to a regression model estimating the effect? Why or why not?

3. Describe the role of probability theory in basic data analysis. Why do we need it?; and how does it relate to simple regression analysis?

Part B

Now suppose you are writing the empirical section of a research paper using the data set you picked beforehand. Make sure to address **each** of the following tasks - which also help you structure your response - but write a single coherent section.[1] In doing so, always explain your data and decisions involved in conducting your analysis.

[1] You may use subheadings to structure your response, of course.

a) Present your data set to the reader - address its substantive scope and contents; its origin and source;[2] and the structure of the data. *Hint: Recall the commands we used in class to explore new data sets.*

b) As we learned in this class, data are the (operationalised and measured) representation of concepts. Now focus on the substantive side of things - after all, you are interested in the (causal) effect your data may help reveal. Identify a potential relationship of interest in your data and explain why you believe that this relationship might exist and be of interest.[3] **Develop and clearly formulate a hypothesis.**

c) Describe and visually present your dependent variable in at least one figure.

d) Describe and visually present your independent variable in at least one figure - using a different figure than the one you chose to present your DV.

e) Visualise the relationship between your IV and DV in at least one figure and calculate their correlation coefficient - what does this (not) tell you about the relationship?

f) Regress your DV on your IV, thus estimating a bivariate model. What does the estimated coefficient tell you about the relationship between the variables?

g) Now include control variables and run at least two multivariate models. Explain your choice of control variables.

h) Present your models in a single table and interpret your results. Plot the coefficients of at least two of the models.

i) Before reaching a substantive conclusion, make sure to plot predicted values based on a multivariate model. Make sure that this plot shows confidence intervals.

j) Taking into account the regression models and prediction plot, evaluate your hypothesis.

k) Test two regression assumptions of your choice using a multivariate model you estimated before. What do your results tell you about your models?

[2] When available online, also add a link to your data.
[3] You are not required to cite any literature here - but feel free to do so if you think this could be of help.

l) Suppose you have reason to assume that your proposed effect is moderated by a third variable. Choose another variable from your data and explain your choice, briefly expanding on your hypothesis, before re-estimating one of your models with an interaction term. Visually present this moderated relationship.

m) Interpret your findings in light of your previous results and write a brief concluding section.

n) Now, write a brief discussion section in which you consider how you could improve your research design. You are not limited by implementation, but are just thinking of the most suitable strategy here. Sketch out a - realistic and feasible[4] - approach that would be ideal to test your hypothesis. What data would you use? What models do you estimate?

[4] While there is no clear-cut boundary, you may use or envision the collection of data that can realistically be collected. For instance, you may envision a survey - but you cannot read peoples' minds directly.