UNIVERSITY OF BAYREUTH

Data Analysis in R | Winter Term 22

Take-Home Exam

Due on 1 February 2023 at Noon

Word limit: 5000 words

Part A

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

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

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

3. Briefly describe the role of probability theory in basic data analysis. 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. In doing so, always explain your data and decisions involved in analysing them. Make sure to address **each** of the following tasks - which also help you structure your response.

a) Present your data set to the reader - address its substantive scope and contents; its origin; 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.[1] **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.

f) Regress your DV on your IV 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 one multivariate model. Explain your choice of control variables.

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

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

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

[1] You are not required to cite any literature here - but feel free to do so if you think this could be of help.