THE GRADUATE INSTITUTE GENEVA

INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT
GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Statistics for International Relations Research I

Final Exam

Rémi Viné

Due: Friday the 23rd of December 2022, at midnight (Geneva time)

The Graduate Institute — Geneva

Instructions

- You have **nine days to complete the test**.
  - The total number of points assigned is 40, one fourth of the full class.
  - Late submissions are accepted but penalties will be applied: submitting 1 day late corresponds to a lower grade by 5 points (out of 40), submitting 2 days late corresponds to a lower grade by 10 points. Beyond 2 days, the grade is reduced by 20 points (so non-passing grades).
- 4 points are assigned to the replicability of your code.
- You are expected to upload your work on Moodle by Tuesday the 20th, midnight, Geneva time.
- Your .pdf document and .R code should address questions in the order in which they appear on the exam.
  - Your .pdf file (or any other well-suited format you prefer) should be a standalone document providing all required answers.
  - Provide details of your computations, process you followed to reach a conclusion. State clearly the underlying hypotheses you use. You will not receive full credit if you only write the final result.
  - Make sure your answer is contextualized, using appropriate levels of measurements.
  - Your .R file should be directly replicable, from the data shared with the exam.
- You are expected to implement the materials covered in classes and lab sessions, and possibly use other R commands that can be handy.
- Citations must be accurate and correct, that is necessary to avoid plagiarism.
- This is an individual exercise, do not work with other students.

What you are provided with

The exam consists in an analysis based on 3 different data sources: (i) World Development Indicators as of 2019 (from the World Bank), (ii) World Urbanization Prospects for cities' population and cities' share of the population in the country with 2018 estimates (from the Population Division of the United Nations), (iii) Urban Social Disorder data (from the Peace Research Institute of Oslo - Thomson et al. (2022)). Along with databases, the preliminary codebook of the latter database is provided. The first database provides indicators at the country level while the remaining databases are at the city level.

The explained variable (there should be only one "original" variable) should be taken from the Urban Social Disorder data. It should be a variable from either the "cities" or the "events" datasets. You might want to transform it or to construct one from the data (*e.g.* the number of deaths in a city over the period of time considered - as long as you explain how you built it and why it suits your work).

What you need to do

Your report should contain the following elements:

1. You should have a clear **research question** in line with what we want you to analyze.

2. You should have a brief **theory** about how various factors may relate to what you are trying to explain. This includes laying out the theoretical relationship between your main predictor(s) and the response variable, as well as talking about how the control variables may be related to the response variable.

   Compelling theories should out forward a suggested causal mechanism. It is not interesting to posit that x and y increase together—tell us how a change in x should set in motion processes that will lead to a change in y. However, be cautious whenever you interpret your results as to whether causal link is statistically shown or not (and remind the core challenges to derive causal conclusions in the context of the data).

3. Based on your theory, you should formulate **research hypotheses** whose observable implications you will be testing. The hypotheses should be on the relationship you want to test between the main explanatory variable(s) and your response variable. The observable implications of the hypotheses should be clearly laid out. The hypotheses need not include only two variables: you can state your expectations about how the relationship between two variables may be moderated by a third, for example (think of possible interactions).

4. Since you are given different dataset, you are expected to **prepare the data** (to clean it) according to your research question and working hypotheses. Explain your choices. You are expected to use the three datasources.

5. You should briefly discuss the **methods** you will use to test your hypotheses. This is artificial, since we require you to analyze the data using multiple regression. But you should still write a paragraph or two about this.

6. You should briefly introduce the **data** you will use, and present **descriptive statistics** about your variables and observations. Descriptive statistics can be used to assess the face validity of your theory and detect patterns (and often, mistakes) in the data. Use any numerical or graphical method as you deem appropriate.

7. You should then use **linear regression analysis** to examine the observable implications of your hypotheses. We expect you to produce at least 3-4 multiple regression models and compare them, before finally settling on one that does a justifiably good job at explaining the variation in your response variable. Do not present too many models, stick to your working hypotheses. One of your models may be a simple, bivariate regression.

   Stronger reports will contain appropriate transformations and interactions (typically, creating dichotomous variables from what you are given), and will also discuss substantive significance.

   Even stronger reports would provide an understanding of the different observational units and the role of this on interpretation.

   Keep in mind that you are not asked to conduct a *p-hacking* exercise. If your working hypotheses are not delivering statistically significant results, so be it! Also remember that you are asked to produce a quantitative analysis that you fully understand: you can use elaborate tools but you must justify why this is necessary in the context of your working hypotheses.

8. Once you choose your model, you should run **diagnostics** to check whether the statistical assumptions underlying the analysis are severely violated. When you identify problems, you should discuss how they could be fixed, but you are not expected to fix them.

9. Finally, you should end with a **conclusion** which summarizes your goals, methods, and main findings, and gives suggestions for future research.

Grading scheme

Introduction & theory (4 pts)

The report proposes a compelling research question relevant to the response variable. It identifies appropriate explanatory and control variables, discussing how they may theoretically relate to the response variable. It presents theoretically-motivated research hypotheses linking the explanatory variable(s) to the response variable. You may motivate your work

citing actual research papers (this is not mandatory). It presents a clear structure for the analysis to come.

Cleaning & descriptive statistics (10 pts)

The report presents the data used to test the hypothesis/es clearly, providing appropriate graphical and numerical summaries for the variables of interest. It also discusses how the data was prepared for the analysis. Remember that the measurement units are important.

Model building (12 pts)

The report uses multiple regression to test the research hypotheses. It methodically fits and compares at least three different models including more than one predictor. It selects the "best" model, provides its prediction equation, and interprets the results appropriately. It makes reasoned decisions on statistical significance, both for the partial coefficients as well as the model as a whole, and discusses substantive significance.

Diagnostics & conclusion (8 pts)

The report recognises the assumptions underlying multiple regression, and employs appropriate plots or tests to assess whether they are violated in the selected model. It uses the diagnostics to show that the model chosen accurately estimates the relationships between our variables. It provides reasoned conclusions on the research question and expected relationships in light of the analysis, discussing how alternative explanations could be responsible for the observed relationships. It suggests ways of improving the analysis.

Statistical literacy & presentation (2 pts)

The report demonstrates sound statistical knowledge by presenting information in a concise and direct manner. Correct notation is used throughout. Figures and/or tables are not raw software outputs or codes, but neat, clearly labeled, and easy-to-read plots and tables.

Readability & reproducibility of the script (4 pts)

The R script reads the dataset provided and runs without any error or need for any additional input to reproduce all the findings presented in the report. The script contains appropriate comments to allow the reader to follow and know which parts of the code produced which elements of the report.

References

Thomson, H., K. Bahgat, H. Urdal, and H. Buhaug (2022). Urban social disorder 3.0: A global, city-level event dataset of political mobilization and disorder. *Journal of Peace Research*, 00223433221082991.