Final Project Information *

Contents

*Substantive parts of this assignment description have been adopted from teaching material developed by Matthew Blackwell, including the links to data sources and the proposal example.

1 Important deadlines and information

- Proposal: Saturday 19 November 2022, 23h59
- Draft Analyses: Saturday 17 December 2022, 23h59
- Final project: 8 January 2022, 23h59 (exact date will be announced in class)

Submit your Rmd file along with the knitted pdf and html files, and the data you used for your analysis. We should be able to run the Rmd file and get all results presented in the paper, so make sure it works.

Students are allowed to team up and write the final project in groups of up to 3 people. Write on Slack (channel final_project) to inform us who is in your team. Do so at the latest before the submission of your proposal. To discourage free riding, each team member can decide at any time to submit a solo project.

2 Final Project Information:

For the final project you are expected to conduct data analysis in the pursuit of answering a research question. The final project should be about 2000-4000 words (tables, figures and references excluded) and should include the following sections:

- title page
- abstract (100-250 words)
- short introduction (about 250 - 500 words)
- concise literature discussion with theoretical expectations (hypotheses) (about 250 - 500 words)
- research design and data section: description of the empirical approach to test the proposed hypotheses, data description, including descriptive statistics and plots (about 500 - 1000 words)
- data analyses and discussion of the results (about 800 - 1500 words)
- concluding discussion (about 200 - 500 words)
- reference list (only works that you cite throughout your final project)

Typically a final project will be between 8 to 15 pages (tables and figures take considerable space). Only the total word count is a strict requirement.

The core of the research project is the data analyses, not the theory and literature parts. You are allowed to use theory and hypotheses, which have been proposed and tested before. Your contribution will be to test these hypotheses with different data or in a different way, run the analyses on your own and interpret (discuss) the findings. Note, that mere replications of existing analyses will not be accepted. Take this task as a preparation for your next empirical research project, e.g. your master thesis.

Below, we outline the steps involved in this project.

3 Start with finding a data source

The biggest part of the final project is finding a data source.

- You are free to use any of the data from the QSS book (found in the `qss` package you have installed).

You can also use the following datasets (see folder on Moodle):

- American National Election Survey, 2016
- Fox News roll-=out
- Afrobarometer

Additionally, here is a list with resources and repositories with data sets :

- Very extensive list of links to political science data sets: https://github.com/erikgahner/PolData

Other useful links

- Harvard Dataverse - Social Science:
- Data.gov - Data sets released by the US government
- Data published by FiveThirtyEight
- Harvard Program on Survey Research Data Collections:
- Roper Center Public Opinion in the US:
- Pew Research Center Data Sets:
- Harvard OpenData Group Directory

You can also search for the replication data of your favorite article. Usually the data used for the analysis is publicly available, the link is often mentioned in the article (usually at the beginning or end, or some footnote) or on the personal webpage of the author. If you can not find the replication file, you can write to the author and ask for it. This is a common practice, do not hesitate to ask for data used in a published article. You can use this data set for your own new analyses (but not mere replication analyses).

If you find a data set that you think is interesting, but you have problems loading the data into R, please contact the course staff. R can load almost anything, so we can likely help you.

3.1 Some advice for choosing data sources

- More often than not, preparing the data for analysis takes longer than the actual analysis itself. Try to find a data set where you do not need to extensively recode / clean up the data before you run your analyses. Usually replication data files form published articles are already cleaned up, but often these contain only the variable used for the analyses in the published article. You might be better off using raw data (e.g. from surveys) as these contain many variables, which might be useful for your project, but then you need to recode and clean up the data. Data from experiments is usually simple to analyze, since the analysis commonly involves simple comparisons of group means.
- Look for a 'codebook' or some other document that explains what the variables mean and how they are coded.
- If you want to analyze the relationship between X and Y, make sure that these two variables are included in the data set. If you want to look at effects for subgroups, make sure there is a variable that you can use for subsetting.
- If the data set is greater than about 50MB, R commands and analyses tend to take longer.

4 Step 1: Write your short proposal (due Sat 19 Nov 2022, 23h59)

You (or your group) should write a one-long-paragraph note to describe what data set you will use and what your tentative research question is. Your research question should ask how one dependent variable is related to one or more independent variables. That is, your research question should be able to be answered by a regression analysis. In this paragraph, you should do the following:

1. State your research question.
2. Formulate a hypothesis related to the research question. This hypothesis should be rooted in some sort of theory. In other words, you need to present a plausible story why the hypothesis might be true, or why do you expect some relationship between X and Y. Explain the underlying logic or causal mechanisms behind this hypothesis - why do you expect this relationship. As social scientists, we are not interested in idiosyncratic or personality driven explanations; we want to understand systematic patterns and relationships!

3. Describe your explanatory variable(s) of interest and how it is measured. Importantly, we need to observe variation in this variable in order to study it!
4. Describe your outcome variable of interest and how it is measured.
5. What observed pattern in the data would provide support for your hypothesis? More importantly, what observed pattern would disprove your hypothesis?

For instance, this would be a comprehensive paragraph that addresses each of these points in detail:

"Does unified government enhance legislative productivity? In this study, I plan to examine the extent to which periods of unified government produce more landmark laws. I hypothesize that legislative productivity increases during periods of unified government in which one party controls both Houses of Congress and the presidency relative to periods of divided government. During periods of unified government, I expect that it is more likely for major bills to pass both Houses and gain the president's signature. During periods of divided government, it is more difficult to reach a consensus around legislation that can pass each House and gain the president's approval. My sample is comprised of each of the 79th (1945-1946) through 103rd (1993-1994) Congresses. My unit of analysis is a Congress (e.g., the 88th Congress). The explanatory variable of interest is whether there is unified government (both Houses and the presidency are controlled by the same party) or divided government. The variable is coded =1 for unified government and =0 for divided government. My outcome variable is the count of landmark pieces of legislation passed in a given Congress. For instance, if the variable were coded =11, it would mean that 11 pieces of landmark legislation were signed into law in that Congress. This variable is measured from David Mayhew's data set on landmark legislation and relies on Mayhew's expert knowledge to classify legislation as "landmark" (see Mayhew (2005)). If I observe greater landmark legislative productivity under unified government relative to divided government, this would provide support for my hypothesis. If, on the other hand, I observe less productivity or the same level of productivity under unified government, this would provide evidence against my hypothesis. When I run my regression of the count of landmark legislation on the unified government indicator variable, a positive, significant coefficient would indicate support for my hypothesis."

Your paragraph might be less detailed and may not refer to the political science literature, but it should address the above items.

Note that you can change your research question, data or the whole project at any time. This proposal is to get you started working on your project and keep you on track.

Upload the PDF of your proposal to Moodle. You can find a template Rmd file for the proposal (called final-project-proposal.Rmd) on Moodle.

5 Step 2: Produce draft analyses (due Sat 17 Dec 2022, 23h59)

The next step will be to create an Rmarkdown file, which - loads the data you have selected - runs any preprocessing that you need to conduct to prepare your data for analysis (e.g. renaming, recoding of variables, creating new variables, merging files, tackling NAs etc.) - produces summary statistics (e.g. tables and plots) of the relevant dependent and independent variables - conducts the main regression of interest for the project.

Upload the Rmd file and the resulting knitted PDF Moodle by Saturday, December 11th at 23h59. Note that these analyses do not have to be final. You may change them or submit something completely different in your final report. We just want to make sure you are making some progress.

6 Step 3: Write up final report (due 8 January 2022, 23h59)

The exact submission date will be announced in class.

6.1 Short guidelines

The final project will include the following sections:

(1) **Abstract**, where you summarize the project within 100-250 words;

(2) **Introduction** where you introduce the research question and motivation;

(3) **Literature review and theory**, where you briefly discuss relevant literature and their findings, identify he gap in the literature and your contribution, discuss your hypothesis and briefly describe the underlying logic behind the hypothesis;

(4) **Research design and data section** that briefly describes your general approach to test your hypotheses, describes the data source, how the key dependent and independent variables are measured (e.g., a survey, expert coding, composite index etc.), and also produces tables and plots that summarize the dependent variable and main independent variables;

(5) **Results and discussion section** that contains a scatterplot of the main relationship of interest, figures with the substantive effect (e.g. predicted values or marginal effects) and output for the main regression of interest (e.g. regression tables). This section should also summarizes your results, assess the extent to which you find support for your hypothesis, and describe limitations of your analysis and threats to inference;

(6) **Concluding section** that provides a short summary of your paper, discusses the limitations of your approach and makes concrete suggestions for future research.

For the data section, you should note if your research design is cross-sectional (most projects will be of this type) or one of the other designs we discussed (randomized experiment, before-and-after, differences-in-differences).

For the results section, you should interpret (in plain English** the main coefficient of interest in your regression. You should also comment on the statistical significance of the estimated coefficient and whether or not you believe the coefficient to represent a causal effect.

6.2 Detailed guidelines with grading scores

Take the word count for each section as a suggestion. Only the total word count is a strict requirement.

1) Abstract (100 - 250 words): the abstract should provide information about the research question, theory and general approach in a concise manner. Take examples from the articles related to your topic, most of them will have an abstract.

2) Introduction (about 250-500 words, 5 points): The introduction should provide a discussion research question you want to address. Provide a short motivation for your research question, mention the gap in the literature you want to address (this is your contribution). The introduction should also discuss the general theoretical expectation and give a short account of your approach to test it (the research design) and your results. Basically, the introduction should provide a short version of your term paper (about 1/2 - 1 page). Take examples from published articles, most start with an introduction and provide the most important information about the article - research question, motivation, the gap in the literature, theoretical expectations, research design and a short summary of the most important results. Try to do something similar. Be concise!

3) Literature Review and Theory (about 250 - 500 words, 18 points): The literature review should provide a short and concise discussion of the literature related to your research question. Avoid just summarizing related papers one after another, instead connect the papers so that they lead to your research question. Identify the problems and gaps in the literature, which you want to answer with your paper. State

what is your contribution! [If you use one of the research questions I provided, you can state that your contribution is to test this research questions say for the 2017 general elections in Austria. Think about some reason why this might be interesting, relevant to do (maybe nobody has done it so far for this election? Maybe this election is different from other elections). *BUT do not invest too much time here, since the empirical part is more important for this assignment.*] Avoid describing the individual articles one by one, instead describe the big picture. For more practical tips how to write a good literature review (e.g. also for future projects) see the relevant chapter from Powner (2015). Check published articles to see how they deal with their literature review. Good articles situate their research question in the existing literature. Try to do something similar.

The theoretical section should introduce the reader to your theoretical expectation and concrete hypotheses. [Note that I provided one hypothesis for the research questions I proposed on Moodle. One hypothesis is sufficient for this task, but if you want, you can have several hypotheses. Please avoid having more than 2-3 hypotheses, as this also increases the number of models you need to run and interpret]. Discuss the underlying logic behind your hypothesis - the causal mechanism - or basically why you expect something to happen, or why you expect a variable X to have a positive or a negative effect on variable Y. Simple logical explanations are sufficient here, no need to develop any theories, rely on theories developed in the literature. You can of course, adjust these theories slightly, add something new, maybe propose a conditional hypotheses (interaction in the model). Pick something simple. Check published articles to see how they write their theory section. Good articles have a theoretical section and discuss the theoretical expectations, hypotheses, and their underlying logic. Try to do something similar.

4 & 5) Research Design and Data Description Section & Analyses and Results Section. (about 1300 - 2500 words, 65 points) These two sections are the most important part of this task. Plan and devote sufficient time:

- Discuss how you want to test your theory and hypothesis/hypotheses
- What data you will use?
- What are your outcome (dependent) and predictor (independent) variables and how are these measured? State your outcome and predictor variables, discuss how you will measure these. If you analyse one of the questions I provided, you will use AUTNES or other similar survey data. In such a case it would be sufficient to describe the variable, state the survey question(s) you use to construct your variables and describe how you constructed them in R if you do any manipulations. In essence, provide enough information so that the reader can understand how you measure your variables, what data you use and how you created them (mention if you scale your variables, or kick out missing values etc.). Do not forget to cite the source of your data.
- Provide some descriptive statistics of your variables, e.g. include a table which shows the min, max, mean, median, standard deviation and number of missing cases for each variable which is used in your regression model. Discuss these. You can also use graphs, histograms, barplots, boxplots etc. – whatever is appropriate for your variables and your purposes. Show descriptive patterns, e.g. plot the outcome and the predictor variables. Discuss each table and graph you include in your paper!
- Describe your model, what other control variables (potential confounding factors) will you include in your regression analysis? Why? Provide a very succinct discussion how you measure these control variables. No need to state the exact question wording for the control variables, it would be sufficient to summarize it with your own words.
- Present your regression results – conventionally regression results are presented in a table, where you report the effect (coefficients) with their standard errors and the p values. See published articles to get an idea what is the convention. You can mention the effect size and the confidence intervals in the text, or plot these if you would like to.
- Create a scatterplot of the main relationship of interest. Include graphs which show the substantive effect of your explanatory variable (e.g. by using the function predict() in R and then plotting the predicted values), include confidence intervals for your predictions and effects.
- Discuss your regression results: assesses the extent to which you find support for your hypothesis. Discuss the impact of your explanatory variable, how strong is the effect? How do you interpret this effect? How much does your outcome variable change when you change your explanatory variable?

How sure are we that this effect is indeed different from zero? In essence is your effect statistically significant and at which level? What is the fit of your model? Check the R^2 and the adjusted R^2. What does it tell you about your model? How much of the variance in your outcome variable is explained by your model, and how much is left unexplained?

- Describe the limitations of your analysis and threats to inference. Discuss whether you can interpret the results in terms of causal effects or mere association (**always think about possible confounding variables**). Are there any confounding variables, which you need to control for, but you do not have data for? Which are these? Discuss why are they important. It is useful to see how published articles present their results to get a feeling how you should do it. We are fine with any format, as long as you cover the above bullet points.

6) Concluding Discussion (200 - 500 words, 7 points): The final section should provide a short discussion of your paper with suggestions for future research.

- The conclusion should identify the limitations of your approach (either in the theory you propose or in the empirical analyses), which open new avenues for future research.
- State how your analysis could be improved (e.g., improved data that would be useful to collect).
- Basically, state what your paper could not do and propose how future research could go about it. If for example you know that there are important confounding factors which you could not include, mention this and emphasize the importance of including these.

7) Reference List (5 points): The reference list should include all works you cited in your term paper. Use the APSA style for your citations and references! The corresponding apsa citation style language file (american-political-science-association.csl) will be included together with the rmd file. We will also provide a file called references.bib, where you can add the works you want to cite in your final report.

8) Rmarkdown File (.Rmd) and Data: Submit your Rmd file along with the knitted pdf file and the data you used for your analysis. We should be able to run the Rmd file and get all results presented in the paper, so make sure it works.

Have fun with this project and do not hesitate to reach out for anything!

References

Mayhew, David R. 2005. *Divided We Govern: Party Control, Lawmaking and Investigations, 1946-2002.* Yale University Press. https://campuspress.yale.edu/davidmayhew/datasets-divided-we-govern/.

Powner, Leanne C. 2015. *Empirical Research and Writing: A Political Science Student's Practical Guide.* SAGE Publications.