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 information

- Final project: 22 April 2022, 23h59

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. Students are allowed to build upon their projects from the class on "Introduction to Statistics and Programming with R I" with the same or new team members. Please consider the feedback I have given you to improve your project. Those who will submit a completely new project and would like to have my feedback, please send your draft version at the latest two weeks before the submission deadline.

If you are struggling with finding a research question and can not settle on a project, we provided some simple research questions and ideas in the file "Research Questions and Data Example.pdf". You can find it on Moodle in the final project section.

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 between 2000-4500 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, using the replication files of the authors 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 For new projects

3.1 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 on Moodle (Folder "Required literature + all supporting files Multiple files") or 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 (often, 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.

3.1.1 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. However, 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.

3.2 Step 1: Write a summary of your idea

To streamline your work it is helpful to start with one page short summary of your idea/proposa. You (or your group) can 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 can 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.

3.3 Step 2: Produce first draft version with initial draft analyses

The next step will be to create an Rmarkdown file, which

- loads the data you have selected

- runs any pre-processing 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.

You can send your Rmd file and the resulting knitted PDF (plus the data to be able to run the Rmd file) to Mariyana via Slack in a private message. Arrange a meeting in person or over zoom to discuss your draft. 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.

3.4 Step 3: Write up final project (see the guidelines for the final write up below - section 4)

4 For those who build up upon their project from Stats I course

4.1 Short guidelines what to include in your final project

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 the gap in the literature and your contribution,
- discuss your hypothesis and
- briefly describe the underlying logic (the causal mechanisms) 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 effects with 95% confidence intervals) and
- output for the main regression of interest (e.g. regression tables with standard errors and p-values).
- This section should also
  - summarize your results,
  - assess the extent to which you find support for your hypothesis
  - discuss the uncertainty (confidence intervals and p values) of your estimated effects,
  - 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 **are expected** comment on the statistical significance of the estimated coefficients (those of interest to answering your RQ and testing your stated hypotheses). **You should also discuss whether or not you believe the coefficient to represent a causal effect (refresh the material on causal effects and internal validity by revisiting Chapter 2 from the QSS book and the accompanying materials from class and assignments).**

4.2 Detailed guidelines with grading scores

Take the word count for each section as a suggestion. Only the total word count is a strict requirement. The total word count should be at least 2000 words and at most 4500 words. Figures, tables, content and reference list are excluded from the final word count.

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 of the research question you want to address.
- Shortly discuss the motivation [1] for your research question and 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. *Tip: Do not invest too much time on the literature review, since the empirical part is more important for this assignment. However, this part is still essential to the project, therefore you need to have a literature review and theory, which give you 18! points.*

- Avoid just summarizing related papers one after another, instead connect the papers so that they describe the big picture and lead to your research question. [2] 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.
- Identify the problems and gaps in the literature, which you want to answer with your paper. State what is your contribution!

[1] If you use one of the research questions we 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.).

[2] For more practical tips how to write a good literature review (e.g. also for future projects) see the relevant chapter from Powner (2015).

The theoretical section should introduce the reader to your theoretical expectation and concrete hypotheses. [3] One hypothesis is sufficient for this task, but if you want, you can have several hypotheses. 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 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. [4]
  - 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:**
  - 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. Include formatted tables and not plain R code output.
  - You can also use graphs, histograms, barplots, boxplots etc. – whatever is appropriate for your variables and your purposes.
  - Show descriptive patterns in the relationship of interest. For example, you can create a scatterplot of the outcome variable against the predictor variables.
  - Always discuss each table and graph you include in your paper!
- **In addition, discuss your control variables:**
  - What other independent variables (potential confounding factors, also called controls) will you include in your regression analysis? Why?
  - Provide a 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**

[3] Note that we provided one hypothesis for the research questions we proposed on Moodle.

[4] If you analyse one of the questions we 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.

- Using a regression table. Conventionally, regression tables report the effects (coefficients) with their standard errors and the p values. P-values are usually presented in the form of stars - one star * for p values below 0.1, two stars ** for p values below 0.05, or three stars *** for p-values below 0.01. Alternatively, **if you want**, instead of standard errors, you can report the exact p values for each coefficient below or next to the coefficient, **or** present 95% confidence intervals for each coefficient in your regression tables. This is not necessary, but is another way to report your results in a meaningful way. In general, it is sufficient to include the estimated regression coefficients, include in brackets their standard errors and the associated p-values using the stars notation in your regression table.
    - You can produce beautifully formatted regression tables with standard errors and p-values using the function `stargazer()`. The `stargazer()` function also allows to include exact p values and confidence intervals below each coefficient. See the instructions how to use `stargazer()` on Moodle. Do not hesitate to ask questions on Slack or arrange a meeting with us if you encounter any problems here!
    - See published articles to get an idea what is the convention to include in regression tables.

- **Discuss your regression results**
    - Assess the extent to which you find support for your hypotheses (or hypothesis)
    - Discuss the impact of your explanatory variable
        - How strong is the effect?
        - Mention the effect sizes (coefficients or sample average treatment effects (SATE)) in the text.
        - How do you interpret the found effects?
        - How much does your outcome variable change when you change your explanatory variable?
    - You are expected to discuss the significance level of your effects, the p-values and/or the confidence intervals of your coefficients. Explain what do these (the standard errors, p-values and confidence intervals) mean substantively.
        - Can you for example reject the null hypothesis? If yes, at what statistical level can you reject the null hypothesis?
        - What is the null hypothesis?
        - Explain what does it substantively mean to reject the null hypothesis.
        - You are not expected to discuss all variables from your model, just the ones of interest for your research question and hypotheses.
    - 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?

- **Present your substantive effects in a meaningful way**
    - Use graphs to show the reader your findings and substantive effects.
    - Plot the estimated effect sizes with 95% confidence intervals. **Alternatively**, you can plot the predicted values with 95% confidence intervals of your outcome variable for different values of your explanatory variables.
        - You can calculate the substantive effect of your explanatory variables directly from the regression output (refresh material how to interpret the coefficient estimates) or by using the function `predict()` in R. You can also plot the predicted values, which you can get using the `predict()` function. Note that the `predict()` function allows to create predicted values with 95% confidence intervals. Also here you can find examples from chapter 4 of the QSS book.

- **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 will accept 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). These limitations can be presented in positive terms as they open new avenues for future research.
- State how your analysis could be improved (e.g. better 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 these and emphasize the importance of including them.

7) Reference List (5 points):

The reference list should:

- Include all works you cited in your final project
- Use the APSA style for your citations and references
- The corresponding APSA citation style language file (`american-political-science-association.csl`) is included together with the rmd template for your final project.
- We also provided a file called `references.bib`, where you can add the works you want to cite in your final report.

Do not hesitate to ask us on Slack or arrange a meeting if you encounter any problems here. Please note that manually inserted citations in the text and manually created reference lists (e.g. if you do not use the references.bib file) will be penalized by lower points.

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 any issue!

Lets cite this work Acemoglu (2015). Another to put it in braackets (see Acemoglu 2015, 45).

References

Acemoglu, Daron. 2015. "Why Nations Fail?" *The Pakistan Development Review* 54(4): 301–12.

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.