INAF U6614: Data Analysis for Policy Research Using R

PROJECT GUIDELINES

This document includes guidance for your project submissions organized into three sections: (1) details for your submitting your reports; (2) suggested paper outlines; and (3) grading guidelines.

1. Submitting project reports

Final projects must be submitted via Courseworks by April 20th (Tuesday), 11:59pm ET. If you are working in a team of two, only one student should submit on behalf of the team. No late submissions will be accepted. Your final report submission makes up 30% of your overall course grade, in addition to 20% from your project presentations.

You are required to submit the following:

1. A knitted .pdf file, along with the corresponding .rmd file used to generate it.
2. All *input* (raw) datasets compressed in a .zip file.
3. All *cleaned* data frames compressed in another .zip file. These will be the only datasets that you load directly into the .rmd file, not input data.
4. R scripts including all code that was *actually used* for your project work (not superfluous code) in another .zip file. Make sure to use comments liberally throughout your code. We recommend one (or more) R script for your data cleaning code, and another R script for your analysis and visualizations.

Here are some guidelines to ensure a clear and reproducible submission:

- You (and the teaching team) should be able to execute the code in your R script(s) that inputs your *raw* datasets and generates the *cleaned* data frames used in your .rmd file.
- All data cleaning and preprocessing should be done in R scripts – NOT in your .rmd file, so that your .rmd code chunks are streamlined and manageable. Your .rmd file should contain *minimal code* necessary to output tables, graphs, and statistics within your report.
- Knitting your .rmd file should produce the exact pdf you submit via CW, and only load the *cleaned* data frames that you actually use to generate your properly formatted .pdf report.
- In your .rmd file, use in-line code references to statistics (don't hard-code).

Report length and formatting:

- The body of your report should contain a maximum of 8 single-spaced pages (excluding charts, tables, references and appendices); that's about 4,000 words.
- All tables and charts should be clearly titled, labelled and formatted, with notes to explain technical details as needed.
- We've provided a sample .rmd file with preferred formatting options. At a minimum, make sure you start with the YAML header included in **report_template.rmd**.

- You may wish to include a Data Appendix in your .rmd file (that doesn't count towards the page/word count), in order to document additional details on your data sources and how your analysis variables were constructed. Here is an example of a short data appendix for an online policy report, and another more detailed example describing survey data.

2. Paper Outline

Here is a suggested outline that you may and aligns with many academic research papers. You will find this outline pre-typed in the report_template.rmd.

Section 1: Introduction
- Clearly state your research questions. Motivate your study and describe the policy context.

Section 2: Background
- More policy background details (if necessary, otherwise fold into Section 1).

Section 3: Data Description
- Describe data sources, representative population, and any other context the reader should know about the data as it pertains to the analysis you do. More details can be included in the Data Appendix.

Section 4: Descriptive Findings
- Describe the distribution of the key variables you're analyzing. Focus mostly on your treatment or policy variable(s) of interest and how they vary across relevant groups and/or time, as well as key covariates that need to be accounted for.
- For example, you should include difference-in-means tables or plots, and/or time series plots of key variables (by subgroups, when appropriate).

Section 5: Empirical Strategy
- Carefully describe the econometric methods and main regression specifications you estimate using clear and unambiguous notation.

Section 6: Findings
- Can be split into multiple sections if appropriate.

Section 7: Conclusion
- Summarize your key findings and policy implications of these findings.
- Discuss the limitations of your analysis and next steps.

Section 8: References

Section 9: Appendices

- Use appendices for more detailed data description, and supplementary tables or charts that provide supporting information that is useful but not central for the story you are telling with your data.

If you prefer, an alternative outline that aligns with many policy or advocacy reports might replace some of the above sections with separate sections organized around key findings. You're free to make your own decisions regarding paper organization, this outline is just a good starting point.

3. Grading Guidance

This submission counts toward 30% of your overall course grade, further broken down as follows:

<u>Research design / introduction (10% in total):</u>

Research questions are clearly formulated and motivated (5%)

- Clearly explain the policy context pertaining to your analysis, what you hope to learn, and why it is policy relevant.
- Provide sufficient background information for readers to understand the problem.

High-level description of your empirical strategy (5%)

- What variation will you be exploiting to inform your research questions, i.e. what comparisons will you be making?
    - EXAMPLE: In this paper, we'll be exploiting variation in neighborhood characteristics between subway station areas to understand how race, poverty, and crime explain differences in fare evasion intensity. This allows us to compare enforcement intensity between stations with comparable poverty rates and crime but different racial composition.

<u>Data, descriptive findings, and empirical strategy (30% in total)</u>

Data description (5%)

- Describe the data sources, any key sampling issues and other data issues, and the representative population.
- If applicable, describe important steps about how you arrived at your analysis datasets from the raw datasets (e.g. sample selection or aggregation) – don't describe routine data cleaning or refer to R functions in your write-up, stick to accessible, policy-relevant terminology. As an example, if you need to aggregate microdata to observations for spatial units like states or provinces, you could briefly explain why/how you did that.

Descriptive findings (15%)

- Clearly define and summarize key variables; key variables include your outcomes, treatment or explanatory variables of interest, and control variables that address major potential sources of omitted variable bias.
- *Do not try to be exhaustive in presenting summary statistics; only show what you think the readers should know in order to follow along in an informative manner.*
    - Use tables and/or plots to present these statistics, the idea is to describe the variation in X and Y that is central to your research design; *difference-in-means tables* and *time series plots* for different groups are two useful approaches.
- Frame your findings in policy-relevant terms, not just statistical terms (i.e. don't just aimlessly report numerous statistics, focus on describing key differences between important groups, or how your treatment variable varies over time, for example).
- Ensure technical accuracy, and clear and precise use of statistical terms.
- Relegate technical details that are not necessary to emphasize to report appendices.

Detailed empirical strategy (10%)

- Describe how your preferred regression specification(s) help inform your research questions in policy-relevant language.
    - DON'T SAY: "We ran an OLS regression to explore the relationships between arrests, race and poverty"
    - DO SAY: "Next we estimate how fare evasion arrest intensity is explained by poverty rates, and how this relationship differs between predominantly Black and non-Black subway station areas. To do this, we estimate the following regression equation separately for predominantly Black and non-Black station areas."
- Present your preferred regression specification using correct mathematical notation, clearly defined variables, and technically accurate and informative use of statistical terminology.
- Discuss threats to internal validity that you are able to address in your preferred specification.

<u>Presentation and interpretation of regression estimates (40% in total):</u>

Presentation and visualization of findings (15%)

- Clear and informative tables and/or charts to summarize your regression estimates
- Make sure the readers can easily follow your tables and charts, and understand the key takeaways.
    - For charts and tables: use clear axis labels, legends and notes, if applicable; make sure the units are clear; and be clear about any relevant statistical details (e.g. SE formulas). Ask yourself: can the reader follow along, can they distinguish between plot markers and lines in your knit document, etc.
    - Regression result specifications and results should be indicated clearly (in the text, chart notes, footnotes, and/or appendices).

Interpretation of findings (15%)

- Ensure technical accuracy, and clear and precise use of statistical terms.
- Also interpret your statistical findings in policy-relevant language that connect to your research question(s).

Conclusion summarizes the key takeaways (10%)

- Succinctly summarize the story of what we've learned from your analysis.
- Also discuss the limitations with some degree of detail.
    - Avoid vague generalities like "we still have an OVB problem" or "we need individual data" – more specifically, what is the OVB problem, and what would more/different data help you to account for?
- Outline possible next steps and the policy implications of your research.

<u>Submission requirements and coding (20% in total):</u>

Files are submitted according to Section 1 of this guideline (5%)

- For example, your .rmd file should not contain code for data cleaning.
- Make sure your report does not exceed the length requirements.

Report is properly written and formatted (5%)

- Ensure your report is free of spelling, punctuation, or grammatical errors.
- Format your report for readability (e.g. sections should be clearly labeled, figures correctly placed, and all text clear and legible).
- Cite sources as needed (in-text and in the References section).

R code is clear, correct, well-organized and efficient (10%)

- Use intuitive and readable variable names, with sufficient comment descriptions.
- Code efficiently: if you find yourself repeating the same functions over and over, you may want to reconsider your approach.
- Only include necessary code, and generously use comments, indentation and white space to ensure readability.
- Reproducibility: we should be able to run your code to reproduce the same results.