Preparing statistics homeworks

Johan A. Elkink

October 21, 2016

Introduction

The homeworks submitted for this class attempt to evaluate and develop your performance on a number of dimensions, namely your ability to:

1. perform statistical analyses as taught in the module to date;
2. interpret the software output from these analyses;
3. interpret the results statistically and substantively;
4. present statistical results in an appropriate manner, i.e. as for a publication;
5. develop appropriate research procedures for statistical research.

Let me discuss for each of those what this implies for the preparation of your homework submissions. This will be a "live" document, which I will update based on submitted homeworks and feedback—so please do let me know if you have any suggestions.

I will refer to the do-file, SPSS syntax-file, or R-file as the code, and I will refer to the PDF submission as the write-up. I refer to SPSS, Stata, or R as the software.

Perform statistical analyses

The baseline of course is that you need to perform the types of statistical analysis taught in this module. This is demonstrated by you providing the code file that shows you took all necessary steps and including the output where appropriate (see next section) in the write-up. This does not need further elaboration here, as it is the most obvious aspect of the homework and the one you expect this is all about.

Interpret the software output

Different software provides different types of output for equivalent analyses and often provides a lot of superfluous information. Just being able to execute the right command therefore does not demonstrate that you know what the result means.

- Always include the result from the analyses in the write-up, without copy/pasting the raw output, as you need to demonstrate that you know which part of the output is the relevant information. E.g. say "I find a correlation coefficient of 0.24" as opposed to some table produced by the software.

- *Always* include an interpretation of the finding.
- An exception is any recoding, file opening, or other transformation of the data, which do not lead to statistical results. Here the code suffices.

Interpret the results statistically and substantively

The next important aspect to evaluate—probably the most important part of the course—is the statistical and substantive interpretation of results. You need to demonstrate that you understand statistics—both through interpreting results and by occasionally answering more theoretical questions in the homework—and you need to demonstrate that you know what the implication is of the finding for political science, understanding politics, or understanding development questions.

- When performing regression analysis, always discuss both statistical significance,[1] direction, and magnitude of effects.
- Always translate the finding to substantive meaning in the respective context. E.g. if a question uses data about trust in politicians and voting behaviour, tell me what you conclude about trust and voting behaviour, not only "the test is statistically significant" or "there is a positive correlation between X and Y". Show that you know what it means.
- Where possible or in your view appropriate, add critical comments on the findings. Can we really draw substantive conclusions or are there weaknesses or limitations to the analysis?
- Avoid short answers or interpretations, make sure you discuss everything you think could be important. Always adhere to word counts—if a question has a suggested length for an answer, do not deviate much.

Present statistical results in an appropriate manner

One aspect of statistical analysis is being able to present results. Here students are typically most surprised by feedback as this appears to be nitpicking, but in my view one thing that you should learn in this class is how to translate results from the software into a proper write-up. We often use replication data from published papers for homeworks and labs—do have a look at how those papers present results and you will see that a lot of what is mentioned here makes sense.

- Provide a write-up that looks a bit like a nicely laid out report—with or without question numbers. This will affect the presentation grade which is typically part of the homework grading.
- Write the report on the statistical and substantive interpretation of the results, not the steps taken. So recodes, transformations of variables, etc. do not need to be discussed in the write-up at all, and code should not be included in the write-up.
- Do not copy/paste tables or other results—except for figures—into the write-up. Properly present results as if for a publication.
- All figures and all tables should have captions to describe the contents.

[1] Once we have discussed t-tests.

- Avoid variable names in tables, figures, captions, and the rest of the write-up. Use more descriptive labels instead. "noside1" is a meaningless phrase, it's just the name of a variable in a data set; instead, "number of coalition partners" is much clearer.[2]

- Statistical results should always be rounded to a reasonable number of digits. When measuring something on a ten-point scale and evaluating the average across a number of people, it does not make sense to talk of "4.53254212"—this suggests a level of precision that is simply not reasonable, and it is also harder to read. Instead, use "4.5" or "4.53".

Develop appropriate research procedures

Finally, an important aspect of doing quantitative research is to learn working habits that allow for a good workflow and replicability of your research. This is of minor importance to the above, but absolutely crucial when you actually do quantitative research, so it is important to learn these habits here. This refers primarily to how to manage your code file.

- Include all code necessary for providing the answers to the homework in the code file.
    - Make sure it is exhaustive, all steps are included, including opening the file and any transformations.
    - Make sure it is minimal, no steps should be included that are not relevant to the eventual analysis.

- Make sure the code file is easy to follow.
    - Add a generous amount of comments to the code file.
    - Include explanations of variable names in the comments, e.g. do not say "regress deaths on noside1", which you can see from the command anyway, but say, "regression to explain the number of battle deaths by the size of the coalition". Do not say "open the data file", which again is obvious, but say "open the data downloaded on 25 August 2016 from http://...., which is the replication data of ...".
    - Do not include the conclusions of any analysis in the code file comments—earlier steps in the analysis might change, which would outdate these conclusions, which can lead to serious confusion if you use the same code later.

- Leave enough whitespace in the code, which means blank lines between (sections of) code and spaces between equations and after commas, e.g. `(x - mean(x)) / sd(x)` instead of `(x-mean(x))/sd(x)` and `mean(x, na.rm = TRUE)` instead of `mean(x,na.rm=TRUE)` in R.

- When recoding or transforming a variable, always create a new variable, and use a variable name that is very easy to understand. So not `x = log(gdp)`, but `logGDP = log(gdp)` and not `gender` to denote a dummy variable on gender, but the name of the category that is labeled as '1' in the variable, e.g. `female` if zero is male and one is female.

[2] Some authors of published papers do opt for the use of variable names in the article itself. I think this is not a good style, but if one does that, the write-up needs very good definitions prior to the usage of the name.