Problem Set 1: Most Common Problems

POLS 8810

Spring 2025

Note that the Problems are are not numbered in order of severity. Problem 2 is clearly the most important as this was the fundamental purpose of this assignment. Given this, for all but the most minor violates of Problem 2, the best possible grade you could receive on this Problem Set was a C^+. Further note that a failure to resolve the issues present in Problem 2 (and 2A) when these variables (or others) are used in analyses in future Problem Sets will render any estimates meaningless for inferential purposes and will thus result in a grade of F on those assignments.

Problem 1: Discussing Variables and Their Coding

See, also, the handout I provided on "How to Write the Data/Methods/Results Sections of a Research Paper."

The way in which many of you discussed the variables is, essentially backwards. I do **not** want to read about how the variables are named and structured in the original ANES data. At least, these should not be the primary focus of your discussion of the variable(s). I want to read about how **your** variables are coded. So you would **never** write something like: "Variable V000519 was used to measure the party identification of the respondent." Rather, you would say something like: "My first variable, *Party Identification*, measures the party with whom the respondent self-identifies." There is no need to discuss the original variable in the ANES dataset here except to provide a notation of the source of your data, but this is better and more appropriately placed in a footnote. So after the example sentence above, one could add a footnote that says something like: "My *Party Identification* variable is derived from the variable V000519 in the original ANES data."

Similarly, we to do discuss the coding of the original variable, at least not for its own sake. Rather we want to discuss how the variabel we are using is coded **after** any manipulations. Assume that we decided to simplify the party id measure to three categories—Democrat, Republican, and Independent—we would then say something like:

"*Party Identification* is a categorical variable coded '1' for respondents who self-identify as Democrat, '2' for respondents who self-identify as Republican, and '3' for respondents who self-identify as Independent." We would then go on to discuss any observations dropped from the original data, any categories in the original coding that were combined, etc. In other words, we only discuss the original source data names and coding in so far as it is necessary to make your data replicable and, even then, the focus should **always** be on your variable with the discussion of the original data as a mere means to an end.

Problem 2: Your Data is Useless for Future Analysis

The purpose of this assignment was "to prepare some data for future use in an analysis. This means that you will want to thoroughly examine the data and the codebook to ensure that the variables are coded in a way that makes sense for use in a regression analysis and, if not, to recode them as needed." Consistant with this, "[t]he primary grade will be based on an evaluation of the student's understanding of how to appropriately prepare... the data."

Many of these write-ups fail to do the most basic, primary purpose of the assignment. At the end of your work for this problem set, your data should be ready to go in terms of suitability for some sort of analysis. If I flagged your paper for this problem, you did not meet this most basic requirement.

The specifics of this issue took many forms, but the most severe of them were cases where you simply renamed a variable from the ANES but did nothing to it. None of the variables in the ANES were ready to use. Ask yourself some basic questions. Does it make sense to include "don't know" responses or non-responses? Almost never as these are not meaningful. This is effectively just missing data and must be dealt with as such. Does it make sense to include categories with a tiny number of responses? Again, almost never, at least in that form. Maybe smaller categories need dropped, maybe they need combined, maybe they need something else, but rarely are they left alone.

Data preparation, or getting your data ready for analysis, is always more than simply renaming things. It requires looking at the data and thinking about what you need to do to it to make it useable. This is not always quick or easy, but it is vitally important and thus the most important aspect of this assignment.

2A (A Special Case): Nominal Variables with Multiple ($k > 2$) Categories

What do we do with nominal variables with multiple categories like race? In real life applications, the answer to this question is theory driven. Here, we have no theory to drive us, yet we do have data which can often tell us what **not** to do, even if it cannot provide us with a single answer to how to do things correctly. Most of you took some race variable from the ANES that had multiple categories. Some of you left those categories unchanged, some reduced the categories (e.g. combining some into 'Other'), some of you did other things. However, nearly all of you left this as a multicategory ($k > 2$), nominal variable.

So why is this a significant issue with respect to the goal of preparing the data for future use in an analysis? Simply put, this variable in it's current form is useless. To provide an example, let's say you coded this variable as White = '1', Black = '2', Asian = '3', and Other = '4.' Further say that you wanted to calculate a mean. What would a mean of 1.5 mean? Or say that you found a positive correlation between this variable and education. Could you say that as race increases so does education? No, clearly this would be meaningless. Generalizing from this, we simply cannot use a nominal variable with multiple categories in a meaningful way in (almost) any quantitative analysis.

Problem 3: Your Description of Your Variables Was Woefully Insufficient (Or Completely Absent)

Many of you failed to thoroughly "provide a descriptive summary of the variables." There were two levels to this problem. The more mild case provided some discussion in the text and/or in a figure/table of what the variables looked like. However, the discussion did not go beyond simply telling me what the raw data looked like. This was more problematic when, in combination with Problem 2, your data was really just the original data. However, in and of itself, this is a relative minor problem.

The worst cases of this issue, however, were when there was a figure or table summarizing the data, but no discussion at all. In other words you provided me with a *potentially* useful table/figure **but** left me as the reader to figure out what it means and why it's there. They are just there. There is not even a reference to the table/figure in the text to tell me what it was supposed to be. A table or figure that is not reference in the text does not belong in a paper. Thus, if you have a table or figure and it is important, you should talk about it and tell the reader what it means. If you do not think it important enough to talk about, then it is not important enough to include in the paper. The assignment asked you to "provide a descriptive summary of the variables." Your tables/figures *could have been* a good way to do this had you taken the time to explain what they told us.

So what would a professional quality write-up have done? It would provide a clear descriptive account of each variable ideally in the form of an aesthetically pleasing figure, a good table, and/or clearly presented in the text. It would then *tell me what the data summary means!* For example, an incomplete paper may tell me nothing at all about the distribution of the variable used for gender. A poor paper, might show the break down of the proportions in each category with no discussion at all just a figure or table that is never referenced in the text. A mildly problematic paper might provide a descriptive like the poor paper, but add some brief discussion that doesn't go beyond simply summarizing the table or figure with words that add nothing else. A good paper will take the next step and tell me what that descriptives mean. In this example, assume there were 60% women and 40% men in your data. You might then say something about how this distribution cause lead to issues with inference as the gender distribution of respondents does not match the actual population of interest.

Very Common, But More Minor, Problems

- All variables should **always** have a meaningful name. For dummy variables this is the '1' value. What does gender mean? Does a '1' on a variable called gender mean those respondents have more gender than those coded '0'? Of course not, this makes no sense. Instead the variable should be named Man or Woman (whichever is the '1' value). This is usually a theoretical call for interpretation purposes.
- Female(s) and male(s) are adjectives. Never use them as nouns. Woman and man are nouns. Thus, you would never say "the sample was 50% females." You would instead say, "the sample was 50% women."
- Always use an Oxford, or serial, comma. Do **NOT** omit these.
- Dummy variables should **always** be coded 0/1 **NOT** 1/2 (or any other number pair).
- Do not use the word "utilize" when you mean "use."
- It is ok to use first-person, singular pronouns (e.g. I, me, my, etc). If you are writing something that is not coauthored, do not use first-person, plural pronouns (e.g. we, our, etc.). You are not multiple people. Also, do not use passive voice to avoid using first-person pronouns (e.g. "the variable was measured" vs "I measured the variable"). It is stylistically abhorrent.
- We do not use bullet pointed lists in research papers. We write things out in paragraphs.
- Colors in graphs for presentations can be for useful in improving the visual aesthetics. However, for papers you should always use greyscale for your graphs as thing are

usually printed in black and white and colors can become indistinguishable when printed in black and white.

- **Never** incude `install.package()` commands in your replication file! Or, at least, comment them out.
- **Never** incude any commands that save files (e.g. figures) locally in your replications file. Or, at least, comment them out.
- When making replication code to share, I (and most others) prefer to have all `library(...)` commands at the top of the file.
- When using publicly available data, always provide a footnote providing the source (e.g. url) where it can be accessed.
- Many of you provided a (way too) lengthy description of the ANES. The ANES and most other widely used, publicly available datasets are well-known and it is unnecessary to provide an overly lengthy description of the dataset itself.