Soc 73994, Homework #1: Preliminary Data Analysis & Setup

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised August 23, 2024

All answers should be typed and submitted through Canvas. Normally homeworks will just be viewed by the TA, but I will view and comment on this assignment too. Be sure your response includes your name, the date, and a clear title, e.g. Homework # 1. If there is a huge amount of output, you may want to be selective in what you copy and paste into your assignment (but make sure you include enough so it is clear what commands you executed, e.g. you might show all the commands but only parts of the output).

The purpose of this assignment is to get a data set ready that you can potentially use later in the semester, either for your paper or at least for other homework assignments. You will do some basic descriptive analyses with the data. The website handout on Basic Dataset Exploration may be extremely helpful to you.

1. Choose a data set. Get it into Stata dta format if it isn't already.
    a. Ideally this will be the data set you want to use or might use for your final paper. This will help to get you off to a running start.
    b. If that isn't possible it could be a data set you have used before, possibly with different or recoded variables.
    c. You could use a data set from a source like the General Social Survey or the European Social Survey. The syllabus lists several places you can get data from. This may require some work on your part so I would be hesitant to do this unless you think you might use such a data set for your paper.
    d. You can use one of the data sets from Long & Freese, so long as you don't just copy things from their book. I think I have all these data sets stored at `https://www3.nd.edu/~rwilliam/statafiles/`. Use commands like `use https://www3.nd.edu/~rwilliam/statafiles/binlfp2, clear`. The index of Long and Freese has an entry for "datasets described" which lists the data sets used in the book.
    e. If worse comes to worst, you can use a data set that comes with Stata. From within Stata, give commands like `help dta_contents` or `help q_base`. I often use the `nhanes2f` data set (which has a lot of medical variables) and the `lbw` (low birth weight) data. (Please, please don't use the horrible auto data set though). You can load these data sets via commands like `webuse nhanes2f, clear`.
    f. Ideally you not only get the dataset, you get documentation that goes with it. A copy of the actual questionnaire can be invaluable. Codebooks and other materials typically show how variables were computed and how the data should be weighted. They also usually say who collected the data, why they did so, and what the virtues and limitations of the data set are.
    g. Briefly describe the data set and your variables so people know what it is. At least briefly explain why this data set is good and why it is appropriate for what you want to study. Just because somebody collected data does not mean it is good!

   h. If you aren't that familiar with the data yourself you will probably want to use things like the `des` or `codebook` commands.

2. Identify two or more variables that might be used as dependent variables in a categorical data analysis. (The more types of possible variables you can identify, the better, since many assignments require a different type of categorical variable.) These might include
   a. A binary variable (employed/unemployed; married/not married)
   b. An ordinal variable (strongly agree/agree/neutral/disagree/strongly disagree; high/medium/low)
   c. A multinomial unordered variable (takes the bus, drives, walks)
   d. A count variable (# of articles published; number of visits to the doctor)
   e. A continuous variable that could be recoded (for example, income might be recoded into high/medium/low; # of doctor visits could be recoded 0 = none, 1 = 1 or more). Indeed I would recommend trying to find at least one continuous variable since, via recoding, it could be used for many different purposes.
   f. Note that other variables might also be recoded for homework purposes, e.g. you might want to dichotomize an ordinal or count variable and use it for logistic regression. Things like the `recode` command can be used. Just be careful you do it right though, e.g. don't accidentally recode missing values to 1s or 0s unless you are sure that is what you want to do.
3. Identify at least three or four variables that might be used as independent variables. Things like sex, race, and income are obvious choices. Ideally you already have some great theory motivating your analysis but if not just pick variables that seem reasonable.
4. Run descriptive analyses of your variables.
   a. For variables with few categories you probably want to use the `tab1` command or, better yet, the user-written `fre` command (available from SSC).
   b. For continuous variables you probably want to use the `sum` command, possibly with the `detail` option.
   c. You may also want to do some bivariate analyses, e.g. crosstab your dependent variable with sex. For continuous and dichotomous variables you could run correlations.
   d. Be careful to identify variables that might need recoding or cleaning up, e.g. "9" might be used to identify missing data, and will need to be recoded. Either fix the problem or if you aren't sure yet how to fix it, at least note that a problem exists. Explain how and why you did any recoding of the data so others can assess if you did it right. Don't just assume the data are clean, because often it isn't! Novices often overlook problems in their data because they've never had to pick their own datasets before; they have just analyzed data that was given to them.
   e. Don't just let the data "speak for themselves," e.g. present a bunch of output and expect the reader to make sense of it. Explain, in words, what the results are, and how they should be interpreted.
5. Discuss the potential pros and cons of your analyses. For example,
   a. Is there a large amount of missing data in any of your variables?

b. Are some of the categories of the dependent variable very thin, e.g. do only a few people strongly disagree?
c. Do you think there are some variables you might need to recode for later analysis, e.g. combine categories?
d. Do you have good measures of the concepts that you think are the most critical for what you want to do? If you don't have everything you want, do you still have enough to write a good paper?
e. Indicate whether you think this data set is worth analyzing further.

If you like this data set you are free to continue using it, at least for homeworks. If you were thinking about using this data set for your paper and you now realize that it seems to have all sorts of problems, or if you just decide you like another data set better, you are free to switch to something else. Better to realize the limitations of a data set one or two weeks into the semester rather than one or two weeks before the paper is due!

I am guessing this will be a very easy assignment for many people. At least some of you already have a dataset and/or are used to running basic analyses. However, if you don't have much experience with Stata or with data sets, this may be extremely difficult. Feel free to contact either the TA or me if you have problems.