BST 760: Style Guide

Patrick Breheny

March 27, 2013

1 Structure

In this course, you will turn in three written projects: one involving logistic regression, one involving Poisson regression, and a final project involving data and an analysis approach of your choice. All projects should contain the following sections:

- **Methods:** (5 points) This should be a fairly boring, standard description of the methods and software you employed to fit the models, carry out tests, construct confidence intervals, etc.
- **Results:** This will be the main portion of your project, and should be broken down into three subsections:
  - **Descriptive:** (20 points) This section should provide basic descriptive statistics summarizing the sample size, outcome variable, and explanatory variables. It is usually a good idea to provide some descriptive statistics for the unadjusted associations between outcome and explanatory variables here as well.
  - **Model choice:** (20 points) This would not typically be a section in an actual published research article, but because making good choices about how to model the data is one of the primary learning outcomes of this course, special emphasis is given to it in BST 760. In this section, (a) clearly state the model that you feel is the best choice upon which to base inferences concerning the relationships being investigated, (b) describe the logic behind this model and the process by which you chose it, and (c) mention any other models that might seem to be reasonable choices and why you *did not* choose them. Components (a), (b), and (c) do not need to be presented in this order; use whatever structure for this section seems clearest to you.
  - **Model results:** (40 points) Present and explain the results of the model you chose in the previous section. **Do not simply present coefficients and p-values**. Decide on relevant quantities of interest that pertain to the research question, report them along with their confidence intervals, and comment objectively on the scientific/medical relevance of these estimated effects. Clearly communicating the results of a complicated model is not trivial, and often requires careful (and creative!) thought.
- **Conclusion:** (10 points) This section should summarize your main results in a mostly qualitative manner: it's usually a good idea to re-emphasize your 1 or 2 most important quantitative

results, but you should in no way try to repeat all of the numbers in the results section. There is no need to overstate your case here: if you feel the study was inconclusive regarding certain research questions, say so. This section should be brief (one or two paragraphs).

The remaining 5 points are for turning in (electronically) your code for carrying out the analysis you describe. This code should be reproducible (*i.e.*, I should be able to run it and obtain the numbers in your results section) and should not appear anywhere in the written document. Documenting your code is always nice, but I will not take away points for a lack of documentation in this course.

For the final project, also include an introduction section (10 points) briefly explaining the background, how the data were collected, and what research questions you hope to answer using the data.

2 Content vs. communication, clarity vs. insight

In the results section, half of the points will be awarded to "content", meaning the numerical results presented (or, in the model choice section, the logic behind the model selection process), and the other half will be awarded to "communication", meaning the quality of the writing describing those results.

"Communication" is further broken down into what I will call "clarity" and "insight", each with 50% of the "communication" points. Clarity involves concise writing that makes clear points with correct descriptions and accurate interpretations. Insight involves pointing out interesting patterns in the data and creatively putting those observations into words. Long, meandering paragraphs making vague and incorrect statements will score low on clarity; dry, mechanical restatements of numbers that already appear in tables will score low on insight.

Thus, for example, the Descriptive Results section is worth 20 points. Of those 20 points, 10 go to the actual descriptive statistics provided, 5 go to the clarity with which the meaning of those numbers are communicated, and 5 go towards providing insight into what the numbers tell us about the sample.

Note that clarity and insight are very much at odds with one another – the narrower and more technical your descriptions of the results, the easier it is to score high on clarity but the harder it is to score high on insight. Conversely, telling a bunch of stories inspired by your data will probably raise your insight score but certainly risks losing points for clarity. Striking the proper balance between the two (finding interesting patterns in the data while refraining from over-interpretation) is perhaps the primary challenge of statistical writing.

To help strike this balance, it is often a good idea to explicitly separate specific numeric results from verbal explanations. Here are three ways of doing so:

- **Separate sentences:** We find that the odds ratio comparing males to females is 2.5, with a 95% confidence interval of (1.9, 3.3). Thus, men are at considerably higher risk of coronary heart disease than women.
- **Parentheses:** Men are at considerably higher risk of coronary heart disease than women (OR: 2.5, 95% CI 1.9-3.3).
- **Use of tables/figures:** As shown in Table 2, men are at considerably higher risk of coronary heart disease than women.

3 Other comments

- **p-values:** It is entirely possible to carry out a thorough analysis and write an excellent report without including a single p-value. No points are explicitly given to the reporting of p-values; whether you include them or not is up to you. However, if you do choose to include them, you must interpret and describe them appropriately. There are more-or-less agreed-upon conventions for describing, in words, the degree of evidence against the null hypothesis that a p-value represents:

	.10	.05	.025	.01	.001
Evidence against H_0	borderline	moderate	substantial	strong	overwhelming

  For example, if the test for whether the regression coefficient for age has a p-value of 0.08, it would be appropriate to say that you have borderline significant evidence that age is associated with your outcome. If $p = 0.04$, there is moderately significant evidence that age is associated with the outcome. If $p = 10^{-8}$, age is clearly associated with the outcome and there probably isn't much point in discussing the p-value other than to assure readers that you have definitive evidence that age is associated with the outcome.

- **Significant digits:** Generally speaking, 2 significant digits is enough when reporting most results – additional digits are often unimportant and distracting. For example, saying that the odds ratio is 2.5 is better than saying that the odds ratio is 2.4762. Unless your sample size is in the millions, it's somewhat absurd to think that you can estimate an odds ratio to the fourth decimal place.

- **Creative use of graphics:** I will also award up to 5 bonus points for creative use of graphs to illustrate results. These can appear in any of the "Results" subsections. Note that quantity is not equivalent to quality – including pages of automatically produced residual plots, QQ plots, etc., will not benefit you, and indeed will almost certainly lower the "clarity" portion of your grade. All your graphs should have a purpose, and should be discussed in your report.

  Two important (perhaps debatable) rules to keep in mind are: (1) Anything you could put in a table, you could also put in a figure, and it would probably be better as a figure, and (2) Just about any statistical concept or observation you have about the data, you could probably think of a way to illustrate it with a graph.