Bios 6312: Modern Regression Analysis

Spring 2016

Project Assignment

March 29, 2016

General Comments:

Students will complete the project in small groups of 3-4 people.

The data set and description are posted on Blackboard under the *Project* content area.

Each group of students will submit a short paper describing the results of a statistical analysis addressing the relevant scientific questions. The paper will then be reviewed by the instructor, and a revised paper will be resubmitted. Grading of the project will be primarily based on the revised paper and response to reviewers' comment, though cases of egregious nonperformance on the first submission will lower the grade on the project.

For electronic submissions, use the PDF file format to ensure that your tables and figures will appear legibly on all computers. Many word processors will save as a PDF directly, but, if needed, there are free converters on the web (e.g. http://www.freepdfconvert.com/) for creating PDF files from various file formats.

Each group will also submit with the revised paper (1) a copy of the dataset (in Stata format, complete with variable labels and, where needed, value labels); and (2) a copy of the documented Stata code.

Due Dates:

Note that all deadlines are strict. Under all but extreme circumstances, failure to meet a deadline is synonymous with failing the project and an incomplete grade in the course.

- 5:00 PM on Monday, April 18. Electronic submission of your first draft to the instructor at robert.e.johnson@vanderbilt.edu .
- Your first draft with comments will be returned within one week after your electronic submission. Thus, if you turn in your first draft before the due date, you will receive it back earlier and have more time to work on revisions for final submission. Drafts will be processed and returned in the order received.
- 5:00 PM on Monday, May 2. Electronic submission of the final draft to the instructor at robert.e.johnson@vanderbilt.edu . Also include a point-by-point response to the reviewer comments.
- Present your analysis to the class on Tuesday, May 3, 9-12noon, in the usual classroom.

Ground Rules:

- You are not to discuss your data analysis or paper with anyone other than the members of your group, the course instructor or course TA.
- Submit one report for your entire group. You should not copy information you obtain from other works into your report. This prohibition extends to the documentation of the dataset which was provided. Use your own words.
- If a member of the group disagrees strongly with something in their group report, they are invited to separately submit a short document detailing their disagreement and how they would have proceeded differently.

Requirements for the Manuscript:

Your paper should be 4 to 12 single sided sheets of paper, not counting figures and tables. It may contain at most five tables and at most four figures (though the figures may have multiple panels to display different endpoints). You may not use fonts less than 10 points for the main text.

In this report, you should describe the results of your analysis and the conclusions you would reach from those results. This report should look like a formal report to a statistically naïve client (i.e., the researcher who brought you the data and/or involved you in the analysis) or an interested reader of the scientific/medical literature. Because a statistical analysis aims to answer a scientific question, you should organize your report in the manner which is customarily used in science.

- *Abstract/Summary:* Provide a concise description of the question, the data used to try to answer it, and the conclusions of your analysis. Give a brief description of the study design/sampling scheme. Give the most pertinent estimates, confidence intervals, and P values. **Note that estimates and confidence intervals regarding the main question of interest are also important when there is no statistically significant effect.** Don't give too much detail here, but do note any significant problems that were encountered. The basic goal is to have all the key information in your summary, and the rest of your report is the supporting detail.

- *Background*: Provide a description of the scientific motivation for the analysis. Use your own words rather than copying the description provided or from some other source. By providing your understanding of the problem, the client may be able to correct any misconceptions that you had about the science. You don't have to go into great detail here, but do give all the facts that entered into your decision process during the analysis. Generally this will include a statement about the overall goal you are trying to address (e.g., the disease and the public health impact of the disease), the current state of knowledge (e.g., conclusions reached in previous studies), and the specific aims of the current study. In general, to conduct an adequate statistical analysis, you need to understand about 10% of the underlying science, not every detail.

- *Questions of Interest*: List the specific questions that your client posed as well as the questions that you answered. Highlight discrepancies between the two categories of questions.

- *Source of the Data*: Describe the source and how the data were gathered, if known. Describe the variables that are available and their meaning for the analysis. Highlight patterns of missing data, if any, as well as possible confounding by measured or unmeasured variables. This should not be a detailed presentation of descriptive statistics, however. That will come under *Results*.

- *Statistical Methods*: Describe the methods used for the analysis at two levels. 1) Give a low-level technical description of the analysis for the client to use in the manuscript. Include references for non-standard techniques. Describe the methods used for assessing the appropriateness of your models. Describe the software used. 2) Explain the basic philosophy behind the analysis techniques in layman's terms. If you transform the data, then give a justification for the transformations. Describe the use of P values and confidence intervals if they play an important role in your analysis.

- *Results*: Provide the pertinent results of your analyses. Do not include all the dead-end analyses you might have done unless they provide insight into the question. Do lead the client up to the analyses gradually.
    - Start off with descriptive statistics. The goal is to describe the basic characteristics

of the sample used to address the question (materials and methods), as well as to present simple descriptive statistics (non-model based) that address the questions. Tables and plots are the key tools. If there are any characteristics of the data that present technical problems that needed to be addressed in the modeling (validity of any assumptions), try to present descriptive statistics illustrating those issues. The basic idea is to presage all the issues you will talk about when presenting the models used in statistical inference, insofar as possible with simple descriptive statistics.

  - Then go to the major analyses used to answer the primary questions. Present summaries of the statistical inference obtained from these models (point estimates, CI, P values). Highlight any particular issues that materially affected the models used to answer the question (confounding, interactions, nonlinearities, etc.) Tables can often be used to good effect here.
  - Leave exploratory analyses (if any) for last and highlight the exploratory nature of those analyses.
  - Present the results of your analyses in tables and publishing quality figures. DO NOT INCLUDE TEXT OUTPUT DIRECTLY FROM STATA, unless you use a program to produce publication-ready results . When possible, use words instead of cryptic variable names. Use forms of estimates that have some meaning to a statistically naïve researcher. Thus, for example, if you log transform your response, present results in the non-transformed form. Present confidence intervals rather than the values of Z, t, F, or chi squared statistics.

- *Discussion*: Discuss the conclusions which you feel can be drawn from the analyses. Suggest directions for future studies and analyses. Highlight the limitations of the data and your analyses. Sometimes particularly speculative analyses are reported here.

The major theme of the above is to write to the scientific community rather than to a statistician. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to anyone.

Also, lead your reader to all the proper results. You spent a long time analyzing the data. Now provide a brief tour through the high points of your work. Statistical diagnostics, which take a lot of our time, can most often be summarized in a single sentence ("Similar trends were observed at other time points." or "We found no evidence to suggest that the final model did not fit the data adequately.") You are reporting your major results and impressions of the data. If the client wanted to see every detail, he/she would have to do the analysis himself/herself.

It is probably most useful to first consider the tables and figures you will present. In an observational study such as this, the following might be provided:

- Table 1: Descriptive statistics for the subjects whose data is included in the analysis.

- Table 2: Descriptive statistics for the associations between the predictor(s) of interest and potential confounders/precision variables in the sample. This might include both univariate and multivariate associations. Statistical significance is of interest insofar as such associations have not previously described in the population, but it may well be of interest to report associations that are not significant *vis a vis* confounding.

- Figure 1: A graphical display of the most interesting associations between the predictor(s) of interest and potential confounders/precision variables. This could either be primarily descriptive (e.g., by showing the (possibly jittered) data) with superimposed smooths, or it could be primarily inferential (by showing point estimates with standard error bars or confidence intervals). With time to event data, it is not uncommon to display the survival curves, which also serves to depict the range of the data.

- Table 3: Inferential statistics presenting univariate associations between the predictor(s) of interest and the outcome variable. For ease of presentation, I might also include univariate associations between potential confounders/precision variables and the outcome variable.

- Figure 2: A graphical display of the association between the outcome variable and the predictor(s) of interest. Again, this can be primarily descriptive or inferential.

- Table 4: Inferential statistics presenting multivariate (adjusted) associations between the predictor(s) of interest and the outcome variable.

- Figure 3: A graphical display of multivariate associations as indicated.

- Table 5: The results of any exploratory analyses of particular interest. This might include investigations of associations within subgroups or analyses directed toward assessing effect modification.

Note that you need not follow this exact scheme. It is provided as a guide to help you understand expectations.