

# Statistics 49S: Fundamentals of Modern Statistical Modeling and Data Analysis Spring 2009, Final Project

For the final project, you will apply the knowledge and skills that you have learned in this course to analyze a data set that interests you. This handout describes the project and demonstrates guidelines for writing the paper.

1. *When is the project due?*

Turn in the written paper by **April 26**. You will present your work in a short talk at the end of the semester.

You can turn in rough drafts before turning in the paper, and I will return them to you with comments. The last day for turning in rough drafts is **April 17**.

2. *What should the project be about?*

The project should be a statistical analysis of a question that interests you. It should involve building, estimating, and interpreting statistical models for genuine data. The question and associated data can be from any academic area. I recommend using existing data sources, so that you do not have to collect your own (which can be very time consuming to do well). All projects must be approved by the instructor.

3. *What happens if I can't find a topic or relevant data?*

Come talk to me, and we'll explore ideas.

4. *What do I turn in?*

Turn in a typed, double-spaced paper that has a maximum of twelve pages of main text. Use 12 point font. The paper should include any graphical and tabular displays that are essential for your analyses within the main text. If you have figures or tables that you want readers (me) to have access to, but you don't want to include them in the main text, you can include them in an appendix. Formal bibliographic references are required. The paper should also include any code that you used for programming.

Tables and figures in the main text count towards the twelve page limit. The bibliography, computer code, and any figures or tables in the appendix do not count towards the twelve page limit.

See the end of this handout for guidelines on paper organization.

5. *How will the project be graded?*

The criteria that I use to grade the project include:

- **Consistency:** Did you answer your question of interest?
- **Clarity:** Is it easy for your reader to understand what you did and the arguments you made?

- **Relevancy:** Did you use statistical techniques wisely when addressing your question?
- **Interest:** Did you tackle a challenging, interesting question (good), or did you just collect and publish descriptive statistics (bad)?

Some suggestions for scoring highly on these criteria, and suggestions to keep in mind whenever you write anything, include the following:

- Know your audience. In this case, you should be writing for fellow STA 49S students. You may want to have others in the class read your paper to make sure that they understand what you are doing.
- State your question up front, and use statistics to help answer it. The statistics should not drive the question; the question should drive the statistics.
- Don't just collect data and publish them; rather, have a specific question in mind. Otherwise, you wind up being hard pressed to come up with something challenging and interesting.
- Most importantly, talk to your instructors for advice. You can ask me, for example, about your methods and analyses, and ask professors in the subject you are covering about background and other issues that can help improve your analyses.
- Be selective with tables and figures to help clarity.

If you are using a technique that we covered in class, you don't have to explain that technique. That hurts clarity. If you are using a technique that we did not cover in class, you should definitely explain that technique. That is clarity!

6. *Can I work in a group?*

No, each person should work individually. I encourage you all to discuss what you are doing with classmates. This will improve your final products.

7. *Can I use data that I am using for a project in another class?*

It depends on the particulars of your proposed project. Come talk to me.

8. *What are the guidelines for paper organization?*

The guidelines are explained in the next two pages. The format of the guidelines mimics how you should organize the paper. You should include the same sections, and follow the same bibliographic and referential style.

# Guidelines for Writing a Scientific Paper With Data Analyses

Jerry Reiter

## 1 Introduction

What are relevant guidelines for writing scientific papers? More appropriately, what are the guidelines for the STA 49S final project? This handout attempts to answer these questions.

In the Introduction, get right to the main topic of the paper. This captures readers' attention. You can put some relevant background information here, but don't spend too much space on it. You want just to give readers a taste of what's to come. Also, it's a good idea to outline your conclusions in this section, so that readers know what to look for as they read.

## 2 Background

This section should contain background information for your readers. Most likely, readers are not going to know the relevant issues of your problem. Explain them here. Also, define any technical terms needed for the remainder of the paper. Don't include technical terms if you don't use them later. That hurts clarity.

This is the appropriate section for references related to the background of your problem and discussion of other, related analyses. For example, Reiter (2000) wrote a paper on causality and statistics that all of you interested in statistics should read. Notice that I refer to the paper with the author's last name and the date in parentheses. The actual bibliographic reference comes at the end of the paper. You don't need to reference standard techniques like linear regression, although when I write for a nonstatistical audience I provide references to texts (e.g., Ramsey and Schafer, 1998) so that readers can learn about the methods in my analyses.

## 3 Data

Here you should describe how the data were collected; describe the variables used in the models; and, discuss issues such as missing data or confidentiality restrictions on the data. It's usually good practice to provide at least means and standard deviations of all variables in your model. That way, readers can get a handle on the type of data that you're using.

Don't include the actual data in the paper. It's nice to make your data publicly available somewhere, such as on the Web. That way, others can use the data you collected, including teachers of statistics courses. You don't have to do that for this project.

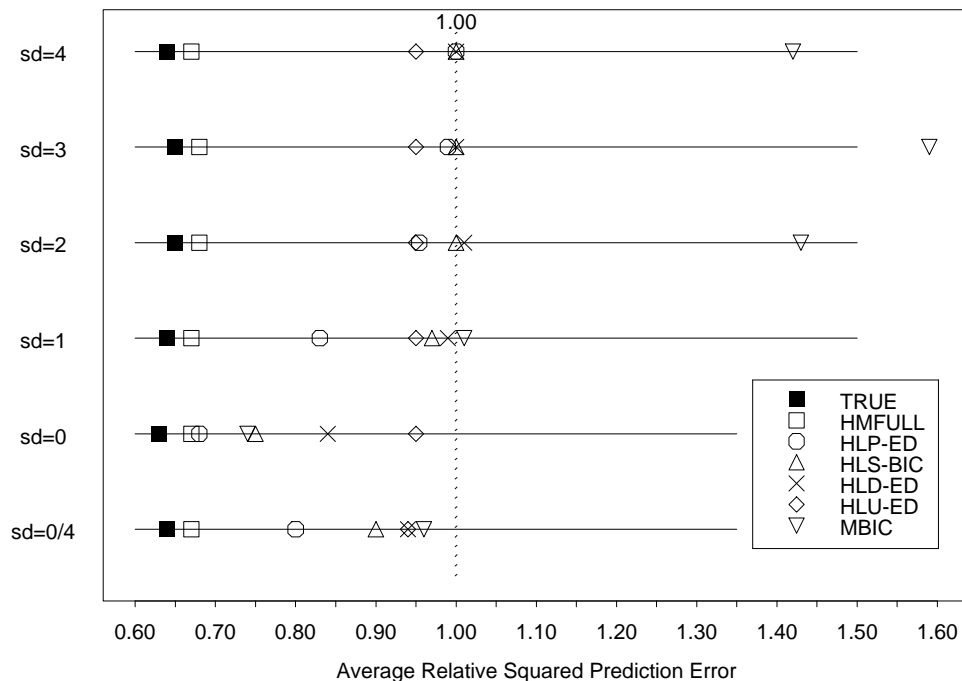


Figure 1: The  $\bar{\mathcal{A}}^{(p)}$  for local HMsPs in the simulation study. The local medial information pooling procedures can increase accuracy relative to FULL, particularly when many coefficients are close to zero, and are generally more effective than MBIC.

## 4 Analyses

Here you describe the analyses and results. You should include the form of the statistical model as a display, e.g., for a regression write

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \quad (1)$$

$$\epsilon_i \sim N(0, \sigma^2). \quad (2)$$

You should define  $y_i$ ,  $x_{i1}$ ,  $x_{i2}$ , and  $x_{i3}$  in words so readers understand the terms in your model.

For Bayesian inference, include figures showing the posterior distributions of key parameters. Alternatively, you can show tables with five number summaries (5th percentile, 25th percentile, median, 75th percentile, and 95th percentile) of the posterior distributions. For maximum likelihood inference, include either figures showing the likelihood or include a table of point estimates with 95% confidence intervals.

Discuss any model checking that you did. For example, you can say something like, “Plots of the residuals versus each individual predictor did not indicate any gross violations of the linear regression assumptions.” There is no need to include the plots if they are uneventful. Should you find a possible violation that you can’t remedy, then you should include the plots and discuss the evidence that led you to suspect that violation.

Any figures included in the text should be clearly labeled. For example, Figure 1 comes from my thesis. Notice that I describe the main conclusion about Figure 1 in the caption. You should do the same. Do not include any graphical or tabular displays in the main text

unless they are relevant for your arguments and are referred to in the paper. If you have other graphical displays or tables that you want readers (including me) to see but don't want to count towards the 12 page limit, you can include them in an appendix. But, make sure that these additional pages are useful for making your points. There's little that is more annoying to a reader than trying to figure out without guidance how graphs or tables relate to a paper. If you want to know what this graphical display means in the context of my thesis research, please come by office hours. I love talking about my thesis research!

Don't forget to discuss the conclusions about the question of interest that you can draw from the analyses.

## 5 Discussion

In this section, you discuss the broad implications of your research. You can discuss issues that you'd like to explore further, interesting relationships among variables that are not quite central to answering your question, or even reservations about the analyses that you think may require more complex modeling.

## References

- Ramsey, F. L. and Schafer, D. W., (1998), *The Statistical Sleuth: Second Edition*, Pacific Grove, CA: Duxbury.
- Reiter, J. P., (2000), "Using statistics to determine causal relationships," *The American Mathematical Monthly* **107**, 24–32.