

Syllabus for SURV 615 and SURVMETH 685
Statistical Methods I, Fall 2009
(last revised September 8, 2009)

1 Class Details

OVERVIEW OF THE CLASS: The purpose of this class is to learn basic statistical methods through the use of linear model theory and regression. The emphasis will be to understand and apply the methods presented, and develop a feel for how problems in data analysis can be viewed in several different ways. In all cases the emphasis will be on understanding the techniques, rather than deriving their theoretical properties. The student will be expected to apply the techniques on weekly homework assignments and a final project which will be presented to the class.

MEETING TIME: Wednesday 3:30-6:10 PM, 1208 LeFrak (Maryland),
368 ISR (Michigan)

INSTRUCTOR: Stephen M. Miller

OFFICE HOURS: Monday 4:00-5:00 PM, JPSM main office

OFFICE PHONE: *JPSM main number* (301) 314-7911

EMAIL: smiller@survey.umd.edu

TEACHING ASSISTANTS:

- **Maryland** Rebecca Medway, rmedway@survey.umd.edu
- **Michigan** Debanjana Datta, ddatta@umich.edu

CLASS WEBSITE: www.jpsm.umd.edu/surv615

CLASS NOTES: Provided on the class website

REQUIRED TEXTS:

- **Applied Linear Regression**, third edition, by Sanford Weisberg ISBN 978-0471663799

- **Statistical Models Theory and Practice**, revised edition, by David A. Freedman ISBN 978-0521743853

RECOMMENDATIONS:

- **The Little SAS book**, third edition (or later), by Lora D. Delwiche and Susan J. Slaughter
- **Introductory Statistics with R**, by Peter Dalgaard
- The R Project for Statistical Computing www.r-project.org
- Online distribution information: (many others are available, search for your favorite e.g. *statistical distributions*)
 - http://en.wikipedia.org/wiki/List_of_probability_distributions
 - <http://faculty.vassar.edu/lowry/tabs.html>

GRADES:

- Homeworks 67%
- Final Project 33%

2 Class Schedule

All readings associated with a given Class should be completed *before* the class meets (except the first class). The instruction will originate from the University of Maryland (JPSM) unless otherwise noted on the schedule (subject to change depending on the needs of the students, and availability of the instructor).

CLASS 1 (September 9, 2009)

Reporting rules. Basic graphical displays. Review inferences from the univariate Normal distribution.

Reading: Check class website, ALR chapter 1 and review appendix as needed.

Assignment: HW#1 due September 23. **By September 16, send an email to the instructor telling him at which email address(es) you wish to be reached. This will be used by the instructor to send out general messages to all of the students. Also by September 16, send an email to**

your respective TA (Rebecca or Debanjana) with times that you would like them to hold office hours. They will examine all of your requests and attempt to find some mutually agreeable times.

CLASS 2 (September 16, 2009)(Instructor at Michigan)

Analysis of data from a bivariate Normal distribution. Paired t -test, pooled t -tests. Behrens-Fisher problem. Inference about variances. Hypothesis testing involving correlations.

Reading: Check class website

Assignment: None.

CLASS 3 (September 23, 2009)

Simple linear regression. Basic assumptions and theory. Introductory analysis of residuals. Tests of Normality, and tests for skewness and kurtosis.

Reading: Check class website, ALR chapter 2.

Assignment: HW#2 due September 30.

CLASS 4 (September 30, 2009)

Multivariate Normal distribution. Multiple linear regression. Basic assumptions. Matrix notation. Properties of the Least Squares estimates. Interpreting the coefficients. Partial correlation coefficients. ANOVA table.

Reading: Check class website, ALR chapter 3

Assignment: HW#3 due October 7.

CLASS 5 (October 7, 2009)

Prediction (Dummy variable technique). General F -testing. ANOVA table. Linear transformations of predictors. Testing linear restrictions: Plug-in method, testing Coefficients method.

Reading: Check class website, ALR chapter 4

Assignment: HW#4 due October 14.

CLASS 6 (October 14, 2009)(Instructor at Michigan)

Model building with predictor variables. Polynomial models and dummy variables. Analysis of covariance. Lack-of-fit tests with repeated observations.

Reading: Check class website, ALR chapter 6

Assignment: HW#5 due October 21.

CLASS 7 (October 21, 2009)

Regression diagnostics and model assessment. Studentized residuals. Measures of influence. Properties of Ordinary Least Squares when the assumptions are violated. Weighted least Squares, and generalized Least Squares. Lack-of-fit tests (variance known).

Reading: Check class website, ALR chapter 8,9

Assignment: HW#6 due October 28.

CLASS 8 (October 28, 2009)

Transformations of variables in regression. Box-Cox transformations. Variance stabilizing transformations. Transformations to linearity. Testing for heteroscedasticity of error variances, and estimating variance functions. Estimated weighted least squares.

Reading: Check class website, ALR chapter 7

Assignment: HW#7 due November 4.

CLASS 9 (November 4, 2009)

Collinearity and Variable Selection bias. Variable selection methods. Stepwise regression.

Reading: Check class website, ALR chapter 10

Assignment: HW#8 due November 11. **Submit a one page proposal by November 11 of what you intend to do for your Final Project. Be specific about the goals of the research and the data set(s) you intend to use. This proposal will not be graded, but it will be returned with comments and approval or disapproval of the proposal will be given.**

CLASS 10 (November 11, 2009)(Instructor at Michigan)

Instrumental variables. Regression with errors-in-variables. Regression estimation with samples with unequal probability of selection. testing if the sampling weights can be ignored.

Reading: Check class website

Assignment: HW#9 due November 18.

CLASS 11 (November 18, 2009)

Fixed effects linear models. One-way analysis of variance. Balanced and unbalanced designs.

Reading: Check class website
Assignment: HW#10 due December 9.

THANKSGIVING BREAK, NO CLASS (November 25, 2009)

CLASS 12 (December 2, 2009)

Two-way analysis of variance. Balance and unbalanced designs. Cell means model versus the over parameterized model. Missing treatment combinations.

Reading: Check class website
Assignment: None.

CLASS 13 (December 9, 2009)(**Instructor at Michigan**)

Random effects linear model. One-way model. Two-way model, and the nested model. Estimation of variance components. Mixed linear models.

Reading: Check class website
Assignment: None.

CLASS 14 (December 16, 2009)

In-class presentations of Final Projects

Reading: Check class website
Assignment: None.

3 Further Details on Grading

Each graded assignment will receive a numerical score, which corresponds to the following letter grades:

Letter Grade	Numerical Score
A+	[98,100]
A	[93,98)
A-	[90,93)
B+	[87,90)
B	[83,87)
B-	[80,83)
C+	[77,80)
C	[73,77)
C-	[70,73)
D+	[67,70)
D	[63,67)
D-	[60,63)
F	[0,60)

Assignments which are turned in late lose 10 points, unless there is a valid excuse given to the instructor ahead of time. Even when an excuse is given, some points may be subtracted based on the judgment of the instructor. Turning assignments in on time is important, and helps keep you on the proper pace with the course. Falling behind is a very bad idea. The final course grade is determined by assigning the letter grade corresponding to the result of the weighted average of the numerical scores for each of the assignments (67% for the homeworks and 33% for the Final Project).

4 Further Details about Assignments

Assignments that are turned in for grading should be neat and easy to read to so that the instructor can grade them properly. While the emphasis is on learning the statistical techniques, it is also important to learn how to present the results of your analyses. The graded assignment will be returned with written comments. All assignments should be turned in by the *beginning* of the class for which they are due.

5 Further Details about Projects

In class presentations will be about 5 minutes depending on the total number of students (this will be specified later). The paper should be 10-20 pages,

doubled spaced with 12pt font. The paper should have four sections labeled: Introduction, Methodology, Results, and Discussion. The following description of these sections is taken from **Stat Labs, Mathematical Statistics Through Applications** by Deborah Nolan and Terry Speed (by permission of the authors).

Use the Introduction to state the problem you are addressing and your findings. Without giving away all your points, let the reader know where your paper is headed.

- Catch the reader's attention. Start with an example, a quotation, a statistic, a question, or a complaint and use it as a theme that you refer to throughout the paper.
- The Introduction sets the tone for your report. Explain why the problem you are addressing is important. Appear to be interested in the topic.
- Break up a long Introduction into several paragraphs. One huge paragraph at the outset of a paper can put readers off.
- Avoid such phrases as "I will discuss" or "this report will examine." Better to just dive right in.

Use the Methodology section to describe your data and how they were collected. This information helps the reader assess the appropriateness of your analysis and the significance of your findings.

- Describe the subject(s) under study. Be as specific as possible. Make clear who was included in the study and who was not.
- Outline the procedures used for collecting the data. For example, if the data are from a sample survey, then the reader may need to know the sampling method, the interview process, and the exact wording of the questions asked. Also address any problems with the data such as nonresponse.
- Explain how the variable measured can be used to address the scientific question of interest. Clearly distinguish between the main outcome of the study and the auxiliary information. Be sure to provide the units in which the responses were measured.

Use the Results section to present your findings. Limit the presentation to those results that are most relevant to your argument and most understandable to the reader.

- Be parsimonious in your use of supporting tables and graphs. Too much extraneous information overloads the reader and obscures the importance of your main thesis. The reader is often willing to accept a brief statement summarizing your additional findings, especially if the material presented is well displayed and to the point. When preparing your data displays, follow the guidelines that appear later in the appendix. Each display must be discussed in the prose.
- Limit the use of statistical jargon. Save the most technical material for an Appendix where you show the advanced reader your more sophisticated ideas and more complicated calculations.
- Include in your report any findings that point to a potential shortcoming in your argument. These problems should be considered in the Discussion section.
- If you include a figure from another paper, cite the original source in your figure caption.

Use the Discussion section to pull together your results in defense of your main thesis.

- Be honest. Address the limitations of your findings. Discuss, if possible, how your results can be generalized.
- Be careful not to overstate the importance of your findings. With statistical evidence, we can rarely prove a conjecture or definitively answer a question. More often than not, the analysis provides support for or against a theory, and it is your job to assess the strength of the evidence presented.
- Relate your results to the rest of the scientific literature. Remember to give credit to the ideas of others. Consider the following questions:
 - Do your results confirm earlier findings or contradict them?
 - What additional information does your study provide over past studies?

- What are the unique aspects of your analysis?
- If you could continue research into the area, what would you suggest for the next step?

The following description of data displays is also taken from **Stat Labs, Mathematical Statistics Through Applications** by Deborah Nolan and Terry Speed (by permission of the authors), and summarizes material from "How to Display Data Badly," H. Wainer, *The American Statistician* **38**: 137-147, 1984.

The aim of good data graphics is to display data accurately and clearly, and the rules for good data display are quite simple. Examine data carefully enough to know what they have to say, and then let them say it with a minimum amount of adornment. Do this while following reasonable regularity practices in the depiction of scale, and label clearly and fully.

The following list provides guidelines for how to make good data displays.

- *Density*- Holding clarity and accuracy constant, the more the information displayed the better. When a graph contains little information, the plot looks empty and raises suspicions that there is nothing to be communicated. However, avoid adding to the displays extraneous graphics such as three-dimensional bars, stripes, and logos. Chart junk does not increase the quantity of information conveyed in the display; it only hides it.
- *Scale*-
 - Graph data in context; show the scale of your axes.
 - Choose a scale that illuminates the variation in the data.
 - Do not change scale in mid-axis.
 - If two plots are to be compared, make their scales the same.
- *Labels*- Captions, titles, labels, and legends must be legible, complete, accurate, and clear.
- *Precision*- Too many decimal places can make a table hard to understand. The precision of the data should dictate the precision reported. For example, if weight is reported to the nearest 5 pounds then a table presenting average weights should not be reported to the nearest 1/100 of a pound.

- *Dimensions*- If the data are one-dimensional, then use a visual metaphor that is one-dimensional. Increasing the number of dimensions can make a graph more confusing. Additional dimensions can cause ambiguity: is it length, area, or volume that is being compared?
- *Color*- Adding color to a graph is similar to adding an extra dimension to the graph. The extra dimension should convey additional information. Using color in a graph can make us think that we are communicating more than we are.
- *Order*- Sometimes the data that are to be displayed have one important aspect and others that are trivial. Choose a display that makes it easy to make the comparison of greatest interest. For example: (a) ordering graphs and tables alphabetically can obscure structure in the data that would have been obvious had the display been ordered by some aspect of the data; (b) Stacking information graphically indicates the total but can obscure the changes in individual components, especially if one component both dominates and fluctuates greatly; (c) comparisons are most easily made by placing information all on one plot.