# Advanced Data Analysis I 36-401 Fall 2000
# Course Policies and Syllabus

**Professor:**
Howard Seltman
232H Baker Hall
268-3938
hseltman@stat.cmu.edu
office hours: M-F 1-2

**Class Schedule:**
Tuesday & Thursday, 3-4:20 p.m
DH 2105

**Teaching Assistant:**
B. Ricky Rambharat
A60-A Baker Hall
268-6289
ricky@stat.cmu.edu
office hours: Wed. 3:30-4:30, Wean 5202

**Prerequisites:** I will assume that you have had a course in statistical methods, including confidence intervals, hypothesis testing, maximum likelihood estimation, and basic distributions, such as the normal, t, $\chi^2$ and F distributions. If you have taken the CMU courses 36-226, 36-326 or 36-310, then you have the appropriate statistical background. You will need to know some linear algebra, such as might be covered in a one-semester matrix algebra course. You will make extensive use of computer packages during the semester. You will need to use S-plus on the campus Unix system. If I assume that you know something that you have never learned, stop me in class or see me after class as soon as possible.

**Objectives:**

1. To develop skills in exploring data, building and fitting models, investigating model assumptions, and interpreting results from statistical models.
2. To apply the theory of linear regression analysis with applications, in order to investigate model assumptions, and take appropriate action if the assumptions don't hold.
3. To develop skills in the use of a modern statistical software package.
4. To learn to write cogent and concise data analysis reports.

**Textbooks:**

- Regression with Graphics: A Second Course in Statistics, by Lawrence C. Hamilton, Duxbury, 1992.
- Modern Applied Statistics with S-Plus, by W.N. Venables and B.D. Ripley, Springer Verlag, 1997.

**Other Useful Books:** The Basics of S and S-Plus, by Andreas Krause and Melvin Olson, Springer Verlag. Applied Linear Regression, by Sanford Weisberg, 1985, Wiley. Applied Regression Analysis, by N.R. Draper and H. Smith, 1980, Wiley.

**Class Organization:**
Class time will include important or difficult concepts from the textbook (*Regression with Graphics*) in lecture format, demonstration of Splus, additional examples of real data analysis, small group work with homework-type problems, and discussion of data analysis issues. You will be expected to learn the concepts found in both the lectures *and* assigned readings, even if the later are not discussed in class.

**Course Overview:**
This course is intended to be an introduction to the real world of statistics, focusing on perhaps the most widely used tool in statistical science, namely linear regression. From econometrics to medical imaging, linear regression provides a language and a body of techniques for specifying and measuring the quantitative relationships between inputs and outputs in a system. We will look at real data, try various models for the data, assess the validity of assumptions, and *try* to reach conclusions. Computer programs will do most of the calculations, and we will be able to concentrate on the thinking. More broadly, it can be fairly said that most parametric statistical models are in some sense a generalization of the linear regression model. If you understand the issues in linear regression, you understand – at least qualitatively – the issues in all statistical modeling and estimation problems. So we will spend some time on the theory underlying linear regression, as well as some generalizations and extensions, but only so far as it aids your ability to analyze real data.

I will stress the fact that linear regression, like all statistical techniques, involve assumptions. A typical statistical analysis begins with simple graphical and descriptive analyses followed by the application of formal statistical techniques. The analysis must then be followed up by diagnostics that check whether the assumptions have been violated. Most good analyses include a heavy dose of graphical techniques. The golden rule in this course is: Never do a statistical analysis without first plotting the data. Also bear in mind that there is never one right way to do an analysis – typically, many analyses are performed.

You will also get plenty of practice writing reports of the types expected of you in the real world. We will primarily focus on reports for intelligent non-statisticians (e.g. a company officer) and brief documentation of what you did for a statistically competent supervisor.

**Homework and Assessment:**

There will be weekly assignments, a mid-term in-class exam, a group project, and an end-of-term final exam. **Do not schedule your plane home on or before 12/21** until we know what day our final will be! The homeworks will include both mathematical problems and data analyses. I encourage you to discuss the assignments with each other, but the work that you hand in must be your own. This means that each student must perform all analyses on her/his own computer, and must independently write up the analysis. Because homework will be an important component of the final grade, you must not copy mathematical derivations, computer output and input, or written descriptions from anyone or anywhere else, without reporting the source with your work. Plagiarism will be swiftly dealt with to the full extent allowed under

CMU policies. Please review the CMU policies on cheating and plagiarism.

The assignments will make use of computer packages, especially S-plus. I will teach you the minimum you need to know about S-plus – enough to do the work, but not enough to call yourself an expert. **Never** hand in raw computer output. Cut out (either electronically, or with scissors) plots, tables, etc. from the output and include them in your report as needed. Label these clearly by hand or with the computer.

For data analysis problems, writing is the most important part. Assignments will request one or more of the following:

- background material to set the problem in context
- a description of the source and nature of the data as well as any changes that you made to the data (e.g., transformations or elimination of outliers)
- an overview of the final model chosen, including assumptions made and techniques used
- an exposition of why competing models may or may not be appropriate
- a conclusion that explains how the data analysis addresses the original problem, recommends a course of action if appropriate, and summarizes any important reservations you may have about the analysis.

Your write-up should consist of well organized paragraphs that address the points listed above in a manner suitable for the intended audience. Each writing assignment will specify the audience, and we will discuss our assumptions about the audience in class. Try to write in a way that will allow your reader to accurately and efficiently learn what s(he) needs to know.

The project will be the joint responsibility of a small group of students. You will analyze a data set, possibly in conjunction with a subject area expert. For students taking the sequel course, Advanced Data Analysis II (36-402), the project will be a preliminary analysis of the data that you will analyze throughout that course. I will aid you in choosing these projects, but if you have appropriate data from your own research, we should discuss whether that can form the basis for your group project.

Throughout the homework assignments we will include analysis of gas mileage data, performed in stages. The special importance of this component of the homework is that it will serve as a kind of practice project. In fact one homework assignment will be to put all of the pieces that you have worked on together into a single report on the analysis of the gas mileage data. *Be sure to save all of your homework assignments* so that you can write the summary report.

After you choose your group project, weekly homework will consist of only one or two problems and a brief summary of your progress on the project. **You must also schedule a project progress meeting with me between 11/27/00 and 12/6/00.** Any additional consultations with me before or after this meeting are also welcome.

**Grading:**

| | | |
|---|---|---|
| Class Participation 10% | Homework: 35% (2 dropped) | Midterm: 10% |
| Group Project: 20% | Final Exam: 25% | |

The cutoffs for A, B, C, and D will be no higher than 90, 80, 70, 60 percent respectively. Based on multimodality of the distribution, I may lower some of the cutoffs.

There is some inherent subjectivity in grading data analysis and write-ups. Data analysis will be graded based on our judgment of reasonable application of the principles and techniques that you will learn in the course. To aid us, you should use justifying sentences, e.g., "The data were log transformed to reduce the marked positive skew."

Grading of written material is based on how well you organize and present the requested components in a manner suitable for the intended audience. Reports that are excessively long (e.g. too much detail or extraneous matter) and those that are too short (e.g. no justifications of conclusions) will not receive full credit. In the real world, errors in spelling and grammar in your work have a major negative impact on the way you are perceived. If spelling and/or grammar corrections are included on your homework or project, these are intended to point out areas you should work on, but do not affect your grade, although poor organization of ideas in your writing *will* lower your grade.

**Data and web page:** The data, homework assignments, due dates, and supplementary material for this course will be available on the course web page:
http://www.stat.cmu.edu/∼hseltman/401. The best way to reach me is via e-mail at:
hseltman@stat.cmu.edu.

**Approximate Course Outline:** I will more or less follow the Hamilton text, though I will supplement portions of it from other sources.

| Dates | Topics | RwG pages |
|---|---|---|
| 8/29 | Course Introduction, Examples | |
| 8/31 | Introduction to Splus (**Meet in Wean 5202**) | |
| 9/5 | Exploratory Data Analysis | 1-23 |
| 9/7-9/14 | Simple Linear Regression | 29-59 |
| 9/19 | Multiple Regression | 65-85 |
| 9/21 | Splus Skills (**Meet in Wean 5202**) | |
| 9/26-10/3 | Multiple Regression | 65-85 |
| 10/5 | Dummy variables, ANOVA as regression | 85-101 |
| 10/10 | MIDTERM EXAM | |
| 10/12-10/24 | Regression Criticism | 109-136 |
| 10/26-10/31 | Fitting curves, non-linear regression | 145-174 |
| 11/2 | Journal article | |
| 11/7-11/9 | Project description presentations | |
| 11/14 | Journal article | |
| 11/16-11/21 | Robust Regression | 183-212 |
| 11/23 | Happy Thanksgiving! | |
| 11/28-11/30 | Logi(s)t(ic) Regression (*) | 217-242 |
| 12/5 | Monte-Carlo and Bootstrap (*) | 303-326 |
| 12/7 | Random effects and mixed models | suppl. |
| 12/12 | Second project write-up due, Review for Final Exam | |

(*) Project progress meetings with instructor 11/27/00 to 12/6/00.