36-401: Advanced Data Analysis I

36-401

Course Policies and Syllabus

Fall 1998

Instructor:
Joseph B. Kadane
232E Baker Hall
268-8726
kadane@stat.cmu.edu

Office Hours:
TBA
by appointment, or just drop in

Schedule:
Tuesday & Thursday, 3-4:30 p.m.
PH A18C

Grader:
Mario Trottini
229J Baker Hall
268-7831
trottini@stat.cmu.edu

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 or 36-326 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.

I also expect a strong desire to analyze data and a willingness to be an active participant in the class. Although I have an agenda for the class I will be delighted if we digress and talk about statistical problems which you might encounter in other work, especially if you are involved in research. You will make extensive use of computer packages during the semester. You will need to use S-plus on the campus Unix system.

Objectives:

1. To develop skills in exploring data, building and fitting models, investigating model assumptions, and interpreting results from statistical models.

2. To blend 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 literate data analysis reports.
Textbooks:
The Basics of S and S-Plus, by Andreas Krause and Melvin Olson, Springer Verlag.

Other Useful Books:
Applied Linear Regression, by Sanford Weisberg, 1985, Wiley.

Class Organization: Unless we find something better to do, I intend to give formal lectures on most days, but we will freely digress to talk about specific data problems as the need arises. If you are ever confused about anything (computing, statistics, data analysis) feel free to propose a digression.

Overview: This course will cover one of the most commonly used statistical tools – regression analysis. Regression analysis is a method for studying relationships between variables. For example, suppose that the S.A.T. scores and G.P.A. were recorded for a group of students. Regression analysis gives us a way to predict G.P.A. based on S.A.T. score. It does so by fitting a model that relates these two variables. Now suppose that we did not know each student’s S.A.T. score but we did know whether each student had a high score or a low score. We might be interested in investigating whether the average G.P.A. of the students with high S.A.T. scores was different from the average G.P.A. of the students with low S.A.T. scores. This is an example of the type of question that ANOVA techniques try to answer. The second semester of this course will look at some formal techniques for ANOVA, as well as various other statistical methods.

This course will emphasize the conceptual and practical bases of regression. We will discuss some computational or mathematical details, but the prime focus will be looking at and interpreting data. I will stress the fact that these techniques, 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.
Homework and Assessment: There will be weekly assignments, a project, a mid-term in-class exam and an end of term final exam. 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 should be your own. 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. Since the question of what kinds of ideas, help, etc. are and are not permitted causes so much confusion, I take this opportunity to state the policies the course will use.

Plagiarism involves using other people's ideas without attribution. Thus plagiarism is essentially a kind of lie, pretending that you thought of something that you did not. The course will therefore operate under a disclosure policy. You are to disclose, with your written work (homework, project, exam), any conversations, reading material, email, etc. you used to help you, or help you offered to others in the course. Anything you disclose accurately absolves you from a plagiarism issue (because you didn't lie about it). (You don't have to report conversations with me or with the teaching assistant). I may have a grading problem, but you don't have a plagiarism problem. If you are feeling guilty or are in doubt about whether to report something, report it. If you are not feeling guilty about it, report it.

Being careful about disclosure not only relieves you (and me) of plagiarism issues, but it helps me to teach you better. If I think, from the disclosures, that you are either allowing others to rely too much on you, or that you are relying too much on others, we can have a private conversation about it. That way, I can hear from you about your circumstances, and you can hear from me about my views.

All homework, particularly the data analyses, should be written in complete English sentences, in a form that any intelligent person can understand. For data analyses it is especially important that you include a description of the context of the problem and the implications of your answer. Your conclusions and interpretation of the results are the most important part of your write-up. 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. I find it very useful and efficient to write a data analysis report as I do the analysis – cutting and pasting output as I go.

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. Statistical computing packages are very easy to use.
There will be two projects in this course. For each of these projects, you are free to use any books or notes as sources, if you reference them properly. Your written reports should be organized into two parts: a non-technical part, walking a statistically naive reader (e.g., a manager, scientist, or judge) through your results, and highlighting the important assumptions and conclusions, and a technical part, explaining clearly and concisely what analyses and procedures you performed, why you made those choices, what your results were, and what measures you took to diagnose the appropriateness of the model.

The first project will be a single data analysis problem assigned to everyone in the class, shortly after we begin discussing multiple linear regression.

In the second project, you will analyze a data set that is of interest to you. You are encouraged to use data from your own research for this project. Either way, you will need to choose and obtain a data set, and it will be helpful to start thinking about this as soon as possible. By mid-October, you should submit to me a brief proposal (one or two paragraphs) describing the question that you will be addressing and the data set that you will use. If you are unable to find a data set, I will assign one.

**Grading:**
- Class Participation: 10%
- Homework: 30%
- Midterm: 10%
- First Project: 10%
- Second Project: 20%
- Final Exam: 20%

**Data and B-boards:** The data and other material for this course will be available in the Andrew directory `/afs/andrew/stat/data/401`. The Andrew bulletin board `bb+academic.stats.36-401@andrew.cmu.edu` will be used to disseminate timely information about the class. You should also feel free to use the b-board to “talk” amongst yourselves. The best way to reach me is via direct e-mail to `kadane@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.
<table>
<thead>
<tr>
<th>Week</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td><em>Doing and Reporting Statistics</em>, Exploratory Data Analysis, Simple Diagnostics</td>
</tr>
<tr>
<td>3-4</td>
<td><em>Simple Linear Regression</em>, Sums of Squares; ANOVA table, Residuals</td>
</tr>
<tr>
<td>5-6</td>
<td><em>Multiple Regression</em>, dummy variables, Interactions, ANOVA</td>
</tr>
<tr>
<td></td>
<td>MIDTERM EXAM (Week 6)</td>
</tr>
<tr>
<td>7</td>
<td>A few examples, Catch-up, begin first project</td>
</tr>
<tr>
<td>8-9</td>
<td><em>Regression Criticism</em></td>
</tr>
<tr>
<td></td>
<td>First project due (Week 8)</td>
</tr>
<tr>
<td>10</td>
<td><em>Influential Observations</em>, (Supplemental material)</td>
</tr>
<tr>
<td></td>
<td>Second project conferences</td>
</tr>
<tr>
<td>11</td>
<td><em>Nonstandard Situations</em>, Transformations, nonlinear models</td>
</tr>
<tr>
<td>12</td>
<td><em>Robust Regression</em></td>
</tr>
<tr>
<td>13</td>
<td><em>Examples and Special Topics</em></td>
</tr>
<tr>
<td>14-15</td>
<td>Second project write-up due</td>
</tr>
<tr>
<td></td>
<td>FINAL EXAM</td>
</tr>
</tbody>
</table>