36-402: Advanced Data Analysis II
Spring 1998

`Student Center 201                TR 9:30--10:50`

Course Policies and Syllabus

Vital Information

Instructor:	*Teaching Assistant:*
Brian Junker 232C Baker Hall 268-8873 `brian@stat.cmu.edu`	Ashish Sanil Student Center 218 268-1889 `ashish@stat.cmu.edu`
Office Hours:	*Office Hours:*
Please feel free to drop in. If I am busy, I will schedule a better time.	Also by appointment.

Required Text

The main text for this course is

- Venables, W. N. and Ripley, B. D. (1994). *Modern applied statistics with Splus.* New York: Springer-Verlag.

which you probably already own, since it was required for 36-401. I will also use excerpts from

- Mathsoft, Inc. (1997). *S-PLUS 4: Guide to Statistics.* Seattle, WA: Author.

Some other books that are useful for this course are listed at the end of the syllabus.

Prerequisites

I assume that:

- You have (more or less) mastered the material in 36-401, especially the material about regression models and diagnostics.
- You can deal with Unix on Andrew. The primary computer package for the course is SPLUS, but we will also be using other packages such as SAS and Minitab. All of these are available on the campus Unix systems (if you can find them on PC's and Mac's on campus, you are welcome to use them there instead/too).
- You have a strong desire to analyze data and a willingness to participate in class.

Course Description

Advanced Data Analysis (ADA) I and II are key courses in the Statistics undergraduate program.

In ADA I (36-401) you should have learned to apply the theory of linear regression analysis—including confidence intervals and hypothesis testing, basic distributions such as the normal, t, χ^2 and F distributions, basic facts and techniques for lienar regression, including the material about regression models and diagnostics—in order to investigate model assumptions, and take appropriate action if the assumptions don't hold. In addition you should have learned something about using SPLUS, and about writing literate data analysis reports.

In ADA II (36-402), we will further develop those skills in exploring data, building and fitting models, investigating model assumptions, interpreting results from statistical models, and report writing. We will begin by talking about some topics that are closely related to linear regression, and move farther and farther from linear regression as the semester progresses.

We will emphasize the conceptual basis of the modeling and data analysis techniques, not the computational details. However, when difficult theory is needed to understand something we will use it. 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 plotting the data.** Also bear in mind that there is never just one right way to do an analysis—typically, many analyses are performed, even by a single person looking at a single data set to answer a single question.

Course Work

In this course there will be approximately weekly "finger exercises" and three or four "mini-projects." I encourage you to work together on the finger exercises; rules for collaboration on the projects are different and I will hand those rules out with the project assignments.

There will also be reading assignments, from Venables and Ripley, and from other material that I hand out or make available on-line. This reading is very important; there won't be enough time for me to go over everything in detail in class.

All assignments and projects in this course will make use of computer packages. I will go over the main points of the computer packages in class, but teaching yourself computing details that I don't have time to cover in class is also part of this course. Nowadays, statistical computing packages have excellent on-line help facilities and are generally very easy to use.

There will be no midterm or final exams.

Grading

Mini-Projects:	60%
Finger Exercises:	20%
Classroom Participation:	20%
Total	100%

Each mini-project will involve taking some data and question(s), analyzing the data to answer the question(s), and writing a report on your analyses and conclusions, using LaTeX, MS-Word, WordPerfect, or some other good word processing software.

Homework and project writeups should discuss the results of your analyses. Your conclusions and interpretations are the most important part of your write-up. **Never hand in raw computer output.** Cut and paste (physically with Elmer's or electronically) plots, tables, etc., from the stat package output and include them in your report as needed. Include a short summary that someone not versed in statistics could read.

For at least the first mini-project, I will provide the data and the question(s). As the course progresses I want the mini-projects to revolve around data and questions that you bring in.

So start thinking about where you are going to get data!

Possible Topics

Here is a menu of topics for the course. I also indicate the relevant chapters in Venables and Ripley, and about how long I expect each topic to take.

From the eight topics listed below we need to select about 4–5 topics for the semester.

1. *ANOVA/ANCOVA and designed experiments.* Linear regression gives us tools for thinking about, designing and analyzing experiements and other studies. Historically, though, experimental design and ANOVA were developed separately from multiple linear regression (because computing used to be expensive), and so the language and approaches to problems are somewhat different from what you learned in ADA I.

   Venables and Ripley, Chapter 6, plus supplementary materials; about 4 weeks

2. *Generalized linear models, logistic regression and categorical data analysis.* Normal-errors linear regression deals with continuous response variables that can take on any positive or negative values. This is the adaptation of linear regression models to deal with responses that are categories (yes/no, good/better/best, red/green/blue), counts, or waiting times.

   Venables and Ripley, Chapter 7, plus supplementary materials; about 4 weeks

3. *Robust statistical methods and non-linear regression*. How can one deal with outliers "automatically" in linear regression? What if the response variables and predictors are all continuous, but the response function is very nonlinear, like $y = a \cdot e^{bx} + \epsilon$ (such nonlinear relationships arise frequently in physics/chemistry/biology applications, for example).

   Venables and Ripley, Chapters 8 and 9; 2–4 weeks, depending on interest.

4. *Random effects and mixed effects models.* In standard linear regression models there is only one source of error or random variation. What do you do when there are several sources of error? For example, in studies of public education, the performance of an individual student is the sum of his/her own performance, plus influences from teachers, the school, the school district, the state, etc., so different students (from different teachers, schools, etc.) should have different "error terms" in the linear regression model. How do we deal with this?

   Venables and Ripley, Chapter 10; 2–4 weeks, depending on interest.

5. *Modern regression methods*, including additive models and projection-pursuit. This would also be a natural place to discuss some topics in *machine learning* (neural networks, reinforcement learning and classification, etc.)

   Venables and Ripley, Chapters 11 and 17; 2–4 weeks, depending on interest.

6. *Continuous multivariate analysis*. Data where there are multiple continuous-valued responses. Possible topics include, Discriminant Analysis, Principal Components, Factor Analysis, Multivariate ANOVA, Clustering Methods, Multivariate Graphics.

   Venables and Ripley, Chapters 13–14; 2–4 weeks depending on interest.

7. *Survival Analysis.* How to look at a clinical trial and try to understand, say, how effective a drug is. Possible topics include Survival Analysis and Hazard Rates, Censoring, Parametric and Non-parametric models, Goodness of Fit.

   Venables and Ripley, Chapter 12; 2–4 weeks depending on interest.

8. *Time series.* Data recorded in time where an important concern is understanding the serial correlation of the data. We might explore both the time and frequency domain.

   Venables and Ripley, Chapter 15; 2–4 weeks depending on interest.

Plan of Action

I will begin the course talking about ANOVA/ANCOVA, designed experiments, and generalized linear models. Where we go from there can depend largely on your interests.

Computing, Data Sets, Web Page, Communications

- Data sets, functions, and other files for this class are kept in `/afs/andrew/stat/data/402`; access the same area on the Web via `http://www.stat.cmu.edu/~brian/402/`.
- Both the TA (`ashish@stat.cmu.edu`) and the instructor (`brian@stat.cmu.edu`) read email regularly. Please feel free to send us email with questions, comments, etc., anytime. Also, please feel free to drop by our offices or schedule special appointments by email anytime.

Other Useful Texts

The following texts will occasionally be useful in this course. Most are in the E&S Library, in the DeGroot Library, on the Web, or on my bookshelves. Amazon books (`http://www.amazon.com`) is a good way to order a copy quickly, if you like.

Books about Data Analysis and Modeling

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis and model building.* New York: Wiley.

Cryer, J. D. (1986). *Time series analysis.* Boston: Duxbury Press.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian data analysis*. London and New York: Chapman and Hall.

Johnson, R. A. and Wichern, D. W. (1992). *Applied multivariate statistical analysis*, 3rd edition. Prentice-Hall, Englewood Cliffs NJ.

Mitchell, T. (1997). *Machine learning.* Burr Ridge, IL: McGraw Hill.

Myers, R. H. (1990) *Classical and modern regression with applications.* (second or current edition). Boston: PWS-Kent (Duxbury).

Neter, J., Wasserman, W. and Kutner, M. H. (1990). *Applied linear statistical models, 3rd Ed.* Homewood, IL: Irwin.

Books about SPLUS

Spector, Phil. (1994). *An Introduction to S and Splus.* Belmont CA: Duxbury Press.
This text gives a gentle introduction to S and Splus, with relatively few rich statistical examples. If you can't figure out how to do something in Splus, often browsing through this book will give you the hint you need. Almost the same info is available on the Web from `http://lib.stat.cmu.edu/` in two formats: `sguide.ps1` is formatted in full-sized 8.5×11 inch pages, and `sguide.ps2` is formatted in 4.25×5.5 inch pages, to save paper.
This was a recommended text for 36-401 last fall, so you probably already own a copy.

Mathsoft, Inc. (1997). *S-PLUS User's Guide, Version 4.0* Seattle, WA: Author.
Contains much of the same material that is in Spector's book above.

Mathsoft, Inc. (1997). *S-PLUS Programmer's Guide, Version 4.0* Seattle, WA: Author.
Technical reference for Splus. We probably won't need it, but I mention it just in case.

Other Useful Books.

Cody, R. P. and Smith, J. K. (1997). *Applied statistics and the SAS programming language.* Upper Saddle River, NJ: Prentice-Hall.
This is sold on the main floor [not the basement] of the bookstore in the University Center, in the computer books section.

Young, Margaret L. and Levine, John. (1995). *UNIX for dummies: quick reference, 2nd Edition.* Foster City, CA: IDG Books Worldwide.
This is sold on the main floor [not the basement] of the bookstore in the University Center, in the computer books section.