36-707: Applied Regression Analysis
Fall 2001

TTh, 12:00--1:20pm, Old Student Center 201
http://www.stat.cmu.edu/~brian/707

Course Information

Instructor:
Brian Junker, Statistics
132E Baker Hall
268-8874
brian@stat.cmu.edu

Office Hours:
T, Th 10–11
(or by appointment).

Grader:
Hoa Nguyen
229J Baker Hall
268-7831
htnguyen@stat.cmu.edu

Office Hours:
T 2–3pm

Prerequisites

I will assume you have had a course in statistical methods, including significance testing, confidence intervals, maximum likelihood estimation, and distribution theory for the normal, t, F, and χ^2 distributions. Also, you will need to know some linear algebra, such as might be covered in a one-semester undergraduate matrix algebra course. Here at Carnegie Mellon, 36-226 or 36-326, or concurrent registration in 36-705, would suffice.[1] You will need to learn the UNIX system and Splus for most data analysis problems in this course, and I'll also assume you have had to write term papers in English at the undergraduate level[2].

Required Texts

Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998). *Applied regression analysis: a research tool (second edition).* NY: Springer-Verlag. Homepage: http://www.stat.ncsu.edu/publications.

Strunk, W. and White, E. B. (2000). *The elements of style (fourth edition).* Boston: Allyn and Bacon.

Venables, W. N., and Ripley, B. D. (1999). *Modern applied statistics with S-plus, third edition.* NY: Springer-Verlag.

Recommended Texts

Krause, A., and Olson, M. (2000). *The basics of S and S-plus, second edition.* NY: Springer-Verlag.

Kopka, H. and Daly, P. W. (1999). *A guide to LaTeX2e: document preparation for beginners and advanced users, third edition.* Reading, MA: Addison-Wesley.

[1] If you have already taken 36-401, please talk to me before continuing in 36-707.

[2] If you are not a native speaker of English, please see the chapter on English as a foreign language in Higham's *Handbook of writing for the mathematical sciences.*, which is available at Carnegie Mellon's Hunt Library.

Course Description

This is a one-semester course on linear regression, intended to be an introduction to the "real world" of statistics. 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. Linear regression provides a language and a body of techniques for specifying and measuring relationships between quantitative variables found in every field of study. Moreover, 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 discussing the theory of linear regression, but applications will never be far from sight.

A statistician's job is never done when the last decimal is calculated and the last graph is drawn, however. Statistical work is almost always done in collaboration with a researcher in another field—perhaps economics, medicine, psychology or engineering—and communicating a data analysis effectively to others is as critical as doing the data analysis well to begin with. When you do communicate well, your work influences the research in your collaborator's field, and you get credit for a job well done. When you do not, your work goes unnoticed and unused. So we will also do a variety of reading and writing exercises in the course, to develop a taste for scientific writing that is informative and persuasive, and we will practice scientific writing by performing and reporting on complete data analyses.

The course will begin with a discussion of exploratory methods, which are informal techniques for summarizing and viewing data. We will then consider simple linear regression, a powerful tool for predicting one variable from another and the basis of much scientific inference. A brief review of linear algebra will precede our discussion of multiple regression, where we have more than one variable available to help us predict a response of interest. Then we will consider the underlying assumptions and look at ways of checking whether the data support these assumptions. Finally, some extensions of the basic model will be examined.

Course Objectives

- Learn some basic tools of exploratory data analysis, including graphical displays.
- Develop model–building skills, including evaluation of assumptions and interpretation of model–fitting results for linear regression models.
- Understand the basic mathematical theory underlying linear regression models.
- Practice carrying out statistical data analysis using a modern statistical software environment.
- Improve the ability to communicate technical information in a clear and organized fashion.

Grading Scheme

Homework is optional, projects are not. You will find however that to get good grades on the projects it will help to do and get feedback on the homework. Class participation—doing the reading, answering and asking questions in class, etc.—also matter a lot but do not figure directly into your numerical grade for the class. For each of you, I will use whichever of the following schemes is more advantageous:

Homework:	30%	Homework:	0%
Project I:	30%	Project I:	45%
Project II:	40%	Project II:	55%

Homework

I will assign approximately one problem set per week as homework. Technical exercises will involve either practice using Splus to implement what we learn in class, or developing and exploring the mathematical material that I will present in class. In addition there will be reading and writing exercises, since these are also skills that a good data analyst must have.

You may work with other students on these problems or refer to other sources if you would like. The computations and writeup of your assignment, however, must be your own. Please note that the written interpretation and conclusions from a data analysis are at least as important as generation of data summaries, statistics, tests, etc.

Unless specifically requested, **never** *submit raw computer output pages*. Instead, cut out the appropriate parts of the output and neatly tape it onto your homework paper (or better yet, use LaTeX, with `verbatim` for text output and `psfig` for figures, to put approporiate parts of the output files into your writeup). Please label all output, plots, variables, etc., appropriately.

Projects

There will be two major projects in this course, in lieu of exams:

- The *first project* will be a single data analysis problem assigned to everyone in the class, shortly after we begin discussing multiple linear regression. You will be able to use whatever books and notes you wish, but you may not discuss this project with anyone but me.
- For the *second project* you will analyze a data set that is of interest to you. You are encouraged to find data yourself for this project, or use data from your own past research experiences. 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 memo describing the question that you want to address and the data set that you will use. If you are unable to find a data set, I will assign one.

We will talk more later about the format and contents of the data analysis reports that you will turn in for these projects, and of the memo that you will turn in for the second project.

Your grade on each data analysis report will be very subjective, and will depend on selecting and adhering to a logical and readable format for the report; on the balance of inventiveness and appropriateness in the selection of methods and conclusions in your report; on the correct use of whatever exploratory, graphical, and/or fitting technique you use; and on the readability and understandability of the report when technical material is deleted.

Computing, Data Sets, Web Page, Communications

- Splus will be used for almost all computer assignment in this class. Splus is available on the department workstations, on Andrew, and on PC/Windows.
- I will assume everyone in the course has an account on the Statistics Department workstations, or on Andrew. If not, let me know.

- As far as possible, all materials for this course will be made available on the course web page, `http://www.stat.cmu.edu/~brian/707`. You may also wish to consult the web pages for the undergrad regression course, `http://www.stat.cmu.edu/~brian/401`.
- The surest way to contact us anytime outside of class is via email:
  - Instructor: `brian@stat.cmu.edu`
  - TA: `htnguyen@stat.cmu.edu`

  Also please feel free to drop by our offices or schedule special appointments with either of us.

Some other useful books

In addition to the books below, I could list twenty more books on the theory and practice of linear regression; most of these are in the stacks at the **E&S** library. In addition there are more books on writing for engineers and social scientists in the stacks of the **Hunt** library; and several other writing texts and references for engineers on permanent reserve in the **E&S** library.

On Applied Regression:

Draper, N. R. and Smith, H. (1980). *Applied regression analysis, 2nd ed.* New York : Wiley. At **E&S**.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* SAGE Publications, Thousand Oaks CA. Homepage: `http://davinci.socsci.mcmaster.ca/`.

Hamilton, L. C. (1992). *Regression with Graphics: A Second Course in Applied Statistics, 1st Edition.* Duxbury Press.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* New York : Wiley. At **E&S**.

Weisberg, S. (1985). *Applied linear regression, 2nd ed.* New York : Wiley. At **E&S**.

On Writing:

Alley, M. (1996). *The craft of scientific writing, third edition.* Springer-Verlag, New York. ISBN 0-387-94766-3. Home page: `http://fbox.vt.edu/eng/mech/writing/csw.html`.

Higham, N. J. (1993). *Handbook of writing for the mathematical sciences.* Philadephia, PA: Society for Industrial and Applied Mathematics. At **Hunt**.

Holloway, B. R. (1999). *Technical writing basics: a guide to style and form.* Upper Saddle River, NJ: Prentice-Hall.

Graduate students in the Department of Statistics also have access to the department's **Degroot Library**, which contains textbooks, reference books and archived journals on many standard topics in Statistics. You may browse and read these books in the Degroot Library, but it is considered bad etiquette to take them out of the library (except for quick trips to the department's xerox machine).

Syllabus and Plan of Action

Following is a rough outline of the *technical* topics we will cover, with estimated readings from Rawlings and other sources. "Rawlings" is the required text on applied linear regression; "MASS" is the Venables and Ripley text on Splus; and "Strunk and White" is the writing style guide. "K/O" stands for the recommended text on Splus by Kraus and Olson. These readings will be supplemented as needed throughout the semester.

I: Preliminaries [2 weeks?]

S/Splus Basics; Exploring Data ..... Handouts; see also MASS, pp 1-68; and K/O Ch's 3, 4, 5
The Role of Statistics ..... Handouts
Writing ..... Strunk & White; Handouts
Simple Regression ..... Rawlings Ch 1

II: Basic Linear Regression [4 weeks?]

Multiple Regression ..... Rawlings Ch 3 [Review Rawlings Ch 2]; Handouts
Checking the Assumptions ..... Rawlings, Ch 10, Sects 11.1, 12.1–12.4
First project assigned
Dummy Variables and F-tests ..... Handouts
Regression Theory ..... Rawlings Ch's 4, 6; Handouts
Memo proposing second project due

III: Diagnostics and Fixes [3 weeks?]

Regression Diagnostics ..... Rawlings Ch 11; Handouts
Collinearity, Model Selection ..... Rawlings Ch 7, Handouts
First project report due
WLS, GLS, Bootstrap ..... Handouts, Rawlings Sect 12.5
Draft of second project report due

IV: Special Topics [4 weeks?]

One- and Two-Way ANOVA ..... Rawlings Ch 9; Handouts
Nonlinear Models ..... Rawlings Ch 14
Logistic Regression ..... Handouts
Taste of Bayesian Regression ..... Handouts
Random Effects and Mixed Models ..... Handouts, Rawlings Ch 18
??? ..... ???
Second project report due

Mileposts for first and second projects are approximate.