An Intense Course in Data Analysis Using Multi-Level Regression Models

Dr. Diane Lambert*
Google
New York, NY 10011
dlambert@google.com

June 15 - June 24, 2010

Course Description

Data is the raw material of knowledge, and computing, graphics, and statistical models are the tools that statisticians use to extract information from data. This course will expand on that theme through a series of lectures and team laboratory projects related to building, interpreting and validating regression models, especially those with binary responses or multiple levels of randomness.

Students will be expected to have had a first course in probability and statistics, including an introduction to linear regression, and to be willing to try new ideas on real data – this is a hands-on data analysis course.

The statistical computing language R is the language that statisticians use to both analyze data and share new statistical methods. R will be used extensively throughout the course. Students who participate in the data analysis labs will be expected to have R loaded on their laptops before the first class and to know the basics of R, including reading data into R, making simple plots, like histograms and scatter plots, computing means, standard deviations and quantiles, and fitting simple linear regression models using the `lm` function. R can be downloaded for free for windows, mac, and linux computers at **http://cran.r-project.org/**. Just follow the installation directions there, choosing the default option whenever there is a choice.

There are several primers on R to help you get started, such as

http://www.cran.r-project.org/doc/contrib/usingR.pdf

http://cran.r-project.org/doc/manuals/R-intro.pdf

http://www.statmethods.net

http://www.stat.washington.edu/cggreen/rprimer/

*With many thanks to Professor Bin Yu, University of California, Berkeley, for encouraging me to teach such a course, to Professor Ming Jiang and Professor Yuan Yao of Peking University for arranging all the details of my visit, to Google for giving me the time to teach the course, and to The MOE-Microsoft Key Laboratory of Statistics and Information Technology and the Beijing International Center for Mathematical Research, for making the course possible.

You do not have to know everything in these documents, but you should know the material, such as means, quantiles, and simple plots, mentioned in the previous paragraph.

I encourage you to try R as you read one of the primers. As Professor Emerson of Yale University says, you can't break R and you can't break your computer using R, so there is no harm in trying it. You may have data from past courses or data of your own to play with in R, but if not there are some data sets available in R itself. Typing `quakes` and `airquality` at the prompt in R will show you two of the data sets that come with R, for example.

In addition to the course, Dr. Lambert will give a seminar titled "Quality Control at Google Scale" that will be open to anyone. Loosely speaking, quality controls usually implies continuously improving industrial products by experimenting with small, local changes in engineering and management processes. At the core of quality improvement lie measurement, experimentation, and learning followed by implementation. This talk will show how Google uses these well-established quality principles (along with huge amounts of data) to improve seach and ads for users, advertisers, and publishers.

About Dr. Lambert

Dr. Lambert is a Research Scientist at Google in New York City. Before joining Google, she was head of the Statistics Department at Bell Laboratories and a Bell Labs Fellow. Previously, she was a tenured member of the faculty of the Statistics Department of Carnegie Mellon University. She is a fellow of the American Statistical Association and the Institute of Mathematical Statistics. She has much experience working with real data in applications as diverse as telecommunications fraud detection, network monitoring, and online advertising.

Tentative Course Schedule

Tue June 15 linear regression
    morning lecture, afternoon lab

Wed June 16 more on linear regression
    morning lecture, afternoon lab

Thu June 17 logistic regression
    morning lecture

Fri June 18 logistic regression
    afternoon lab

Mon June 21 seminar on Statistics at Google

Tue June 22 introduction to multilevel linear models
    lecture + lab

more on multilevel linear models
    Wed June 23 lecture + lab

multilevel logistc regression
    Thu June 24 lecture + lab