Linear Regression Models
W4315

Instructor: Dr. Frank Wood

Required Text: Applied Linear Regression
Authors: Kutner, Nachtsheim, Neter

Not Registered Yet?

Fill out the form at
http://tinyurl.com/mqfq95

Additional books we will draw material from in this course:

- Pattern Recognition and Machine Learning, by Christopher M. Bishop. Springer, 2006.
- Bayesian Data Analysis, Second Edition, by Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin, Chapman & Hall/CRC Texts in Statistical Science

Course Description

Theory and practice of regression analysis, Simple and multiple regression, including testing, estimation, and confidence procedures, modeling, regression diagnostics and plots, polynomial regression, colinearity and confounding, model selection, geometry of least squares. Extensive use of the computer to analyze data.

Course website http://bit.ly/9QoeyL

or http://www.stat.columbia.edu/∼fwood/w4315/

Philosophy and Style

- Easy first half.
- Very hard second half.
- Frequent, long digressions from the required book.
- Understanding == proof (derivation) *plus* implementation.
- Practice makes perfect.
- Frequentist *and* Bayesian perspectives taught.

If you are looking for a pure applied, pure frequentist treatment of regression as a diagnostic tool and/or you've never programmed before, seek another section.

Available sections

- MW 9:10-10:25am, 702 Hamilton Hall (Lindquist)
- MW 6:10-7:25pm, 209 Havemeyer Hall (Yang)

Goals

- Deep theoretical understanding
  - Book provides only recipes
  - Much detail missing
  - Can always look up recipes in future
- Ability to implement/code *all* regression functionality
  - Different levels of understanding
  - Not enough to simply be able to apply formula and use pre-built regression software

Fair warning:

I am not perfect in attaining my goals...

Feedback from prior years

Frank was often better at relaying the information when he cared about the material. For example, the sections on Bayesian Statistics were explained more clearly and thoroughly than the earlier sections. Frank answered questions well in person, but rarely responded to emails.

Prof. Wood is clearly a brilliant, fascinating individual. Everyone in the class, I think, likes him on a personal level and his teaching style. However, overall, this course was a waste of $4,000. I paid money for an applied course in linear regression models. Instead, basically, the course was taught from a theoretical standpoint.

Feedback from prior years

> *Frank definitely knows what he's doing; the material can sometimes be a little rough and it's not always clear what his motivations are in trying to explain some topics, but I wouldn't want to take this class with any other professor.*

Feedback from prior years

Never had I had such an effective statistics instructor at Columbia. Prof Wood explained the material so well, yet very quick, and got discouraged when he spent more time expected explaining certain topics... even though it was really to the benefit of the class. I honestly wish he could teach me all of the material I have to encounter before graduation and then maybe it would make sense. My only criticism is that as energetic as he was in class, he was quite intimidating, which I think was a disservice to most of the class when he tried to facilitate class participation. Only about 15% of the class regularly participated as a result.

About me

- Computer Science PhD, 2007, Brown University
- Postdoc in Machine Learning, Gatsby Unit, University College London
- Sports gambling consulting.
- Former entrepreneur.

My research

- Inference for nonparametric Bayesian models.
- Compression.
- Natural language data modeling.

My website: http://www.stat.columbia.edu/∼fwood

Course Outline

First half of the course is on the traditional view of linear regression. The first half of the first half is a formal, theoretical review of single variable regression and its classical, frequentist treatment.

- Roughly 1 chapter per week
- 3-5 weeks, linear regression
  - Least squares
  - Maximum likelihood, normal model
  - Tests / inferences
  - ANOVA
  - Diagnostics
  - Remedial Measures
  - Linear algebra review
  - Matrix approach to linear regression

Course Outline Continued

The second half of the first half covers multiple regression and the various topics that arise from including multiple predictor variables into models.

- 3-4 weeks multiple regression
  - Multiple predictor variables
  - Diagnostics
  - Tests

Midterm

Course Outline Continued

The remainder of the course will deviate from the book and may be ordered differently than what is shown here. In general we will retain a focus on models that are linear in the parameters, but will look at nonlinear models and Bayesian treatments of linear models.

- 3-4 weeks on Bayesian regression
  - MCMC
  - Bayesian linear regression
  - Gaussian process regression
  - Projects

- 3-4 weeks on generalized regression
  - Polynomial regression
  - Logistic regression
  - Neural networks
  - Generalized linear models

Requirements

- Calculus
  - Derivatives, gradients, convexity
- Linear algebra
  - Matrix notation, inversion, eigenvectors, eigenvalues, rank, quadratic forms
- Probability
  - Random variables
  - Bayes Rule
- Statistics
  - Expectation, variance
  - Estimation
  - Bias/Variance
  - Basic probability distributions
- Programming

Projects (homework and final)

- Software
  - For homework – Matlab.
  - For final project, don't care:
    - R
    - Matlab
    - S-Plus
    - SAS
    - Minitab
    - Excel
    - java, c++, c, assembly, . . .

Grading

- Bi-weekly homework (35%)
  - Due every other week
    - no late homework accepted
  - None allowed to be missing
  - No homeworks due during final project preparation period.
- Participation (5%) (up a half grade if I know you by the end, down a half grade if not)
- Midterm examination (25%)
- Final project (35%)
- Curve

Office Hours / Website

- $http://www.stat.columbia.edu/\sim fwood$
- Office hours : TBA
- Office : Room 1017
- TA : Wei Wang
  - TA office hours TBD
  - Email: ww2243@columbia.edu
  - Office: 901 SSW

Why regression?

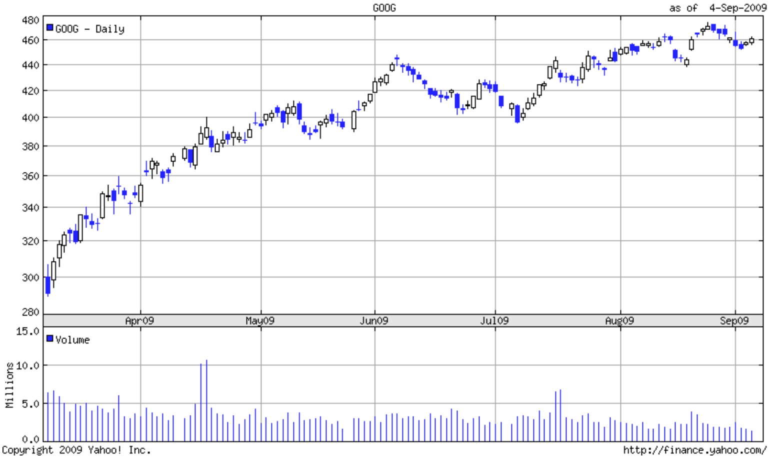
- Want to model a functional relationship between an "predictor variable" (input, independent variable, etc.) and a "response variable" (output, dependent variable, etc.)
  - Examples?
- But real world is noisy, no $f = ma$
  - Observation noise
  - Process noise
- Two distinct goals
  - Tests about natural of relationship between predictor variables and response variables
  - Prediction

History

- Sir Francis Galton, 19^{th} century
  - Studied the relation between heights of parents and children and noted that the children "regressed" to the population mean
- "Regression" stuck as the term to describe statistical relations between variables

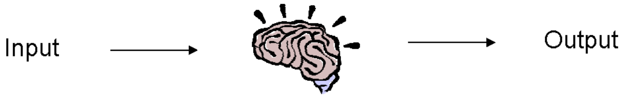
Example Applications

Trend lines, eg. Google over 6 mo.



Others

- Epidemiology
  - Relating lifespan to obesity, smoking habits, and/or other patient *features*.
- Science and engineering
  - Relating physical inputs to physical outputs in complex systems
- Grander



$$\text{output} \sim f(\text{input})$$

Aims for the course

- Given something you would like to predict and some number of covariates
  - What kind of model should you use?
  - Which variables should you include?
  - Which transformations of variables and interaction terms should you use?
- Given a model and some data
  - How do you fit the model to the data?
  - How do you express confidence in the values of the model parameters?
  - How do you regularize the model to avoid over-fitting and other related issues?
- Not be boring