Linear Regression Models
W4315

Instructor: Dr. Frank Wood

Required Text: Applied Linear Regression Models
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.

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.

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.

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 generalized regression
  - Polynomial regression
  - Logistic regression
  - Neural networks
  - Generalized linear models

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

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
  - One allowed to be missed
  - Completing all is "extra-credit"
- Participation (5%)
- Midterm examination (25%)
- Final project (35%)
- Curve

Office Hours / Website

- $http : //www.stat.columbia.edu/ \sim fwood$
- Office hours : Thursday 7:30-9pm subject to change
- Office : Room 1011
- TA : Heng
  - TA office hours TBD

Why regression?

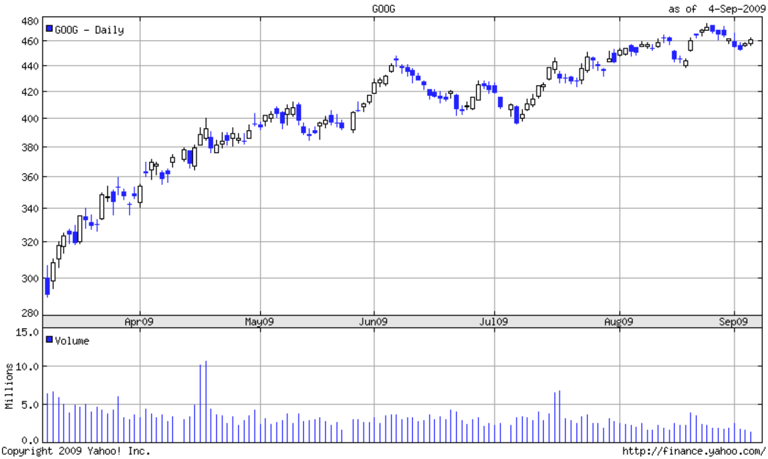
- Want to model a functional relationship between an "predictor variable" (input, independent variable, etc.) and a "response variable" (output, dependent variable, etc.)
  - Examples?
- But real world is noisy, no $f = ma$
  - Observation noise
  - Process noise
- Two distinct goals
  - Tests about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

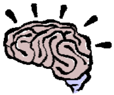
Example Applications

Trend lines, eg. Google over 6 mo.



Others

- Epidemiology
  - Relating lifespan to obesity or smoking habits etc.
- Science and engineering
  - Relating physical inputs to physical outputs in complex systems
- Grander

Input $\longrightarrow$  $\longrightarrow$ Output

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

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?