Linear Regression Models
W4315

Instructor: Dr. Yang Feng

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

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.
- Hard second half.
- Some digressions from the required book.
- Understanding $==$ proof (derivation) *plus* implementation.
- Practice makes perfect.

About me

- Operations Research & Financial Engineering PhD, 2010, Princeton University
- First time teaching a course...

My research

- High-dimensional Statistical Learning
- Variable Selection
- Nonparametric and Semi-parametric Statistics
- Bioinformatics

Course Outline

First half of the course is single variable linear regression.

- Least squares
- Maximum likelihood, normal model
- Tests / inferences
- ANOVA
- Diagnostics
- Remedial Measures

Course Outline (Continued)

Second half of the course is multiple linear regression and other related topics .

- Multiple linear Regression
  - Linear algebra review
  - Matrix approach to linear regression
  - Multiple predictor variables
  - Diagnostics
  - Tests
- Other topics (If time permits)
  - Principle Component Analysis
  - Generalized Linear Models
  - Introduction to Bayesian Inference

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

Software

R will be used throughout the course and it is required in all homework. An **R** tutorial session will be given on Sep 22.
Reasons for **R**:

- Completely free software
- Available on various systems, PC, MAC, Linux, $\cdots$

Grading

- Bi-weekly homework (25%)
  - Due every other week
    - no late homework accepted
  - Lowest score will be dropped

- Exams are open book and open notes.
- In Class Midterm exam (30%), Wednesday, Oct 20, 2010 (tentatively).
- In Class Final exam (45%).
- Curve

Office Hours / Website

- `http://www.stat.columbia.edu/~yangfeng`
- Course Materials and homeworks with the due dates will be posted on the course website.
- Office hours : Wednesday 2-4pm subject to change
- Office Location : Room 1012, SSW Building (1255 Amsterdam Avenue, between 121st and 122nd street)
- TA : Qinghua
  - TA office hours TBA

Why regression?

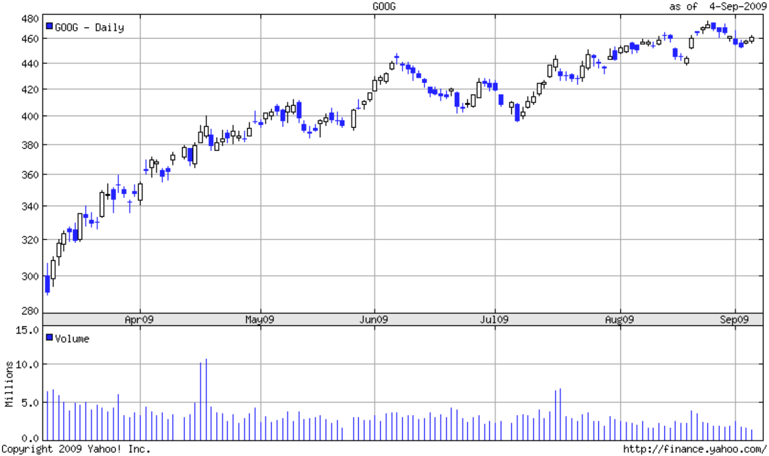
- Want to model a functional relationship between an "predictor variable" (input, independent variable, etc.) and a "response variable" (output, dependent variable, etc.)
    - Examples?
- But real world is noisy, no $f = ma$
    - Observation noise
    - Process noise
- Two distinct goals
    - Tests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

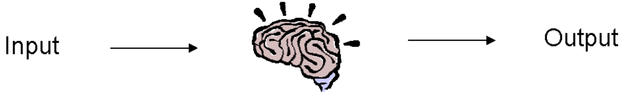
Example Applications

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

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?