Linear Regression Models
W4315

Instructor: Dr. Yang Feng

Required Text: Applied Linear Regression Models (4th Ed.)
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?

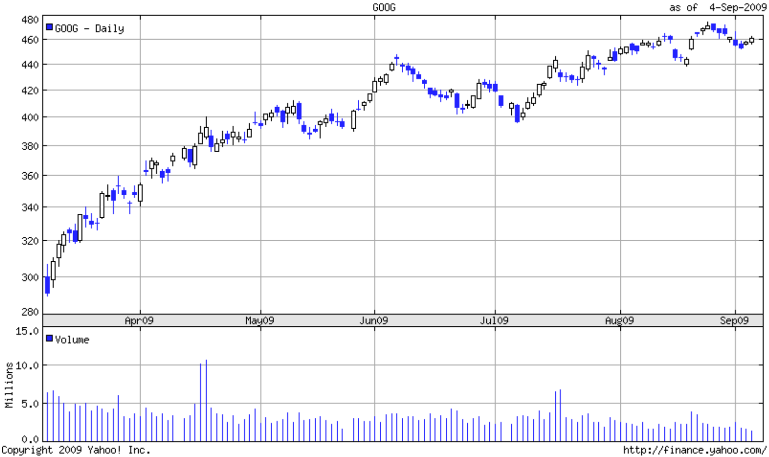
- Want to model a functional relationship between an “predictor variable” (input, independent variable, etc.) and a “response variable” (output, dependent variable, etc.)
  - Examples?
- But real world is noisy, no $f = ma$
  - Observation noise
  - Process noise
- Two distinct goals
  - Tests about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

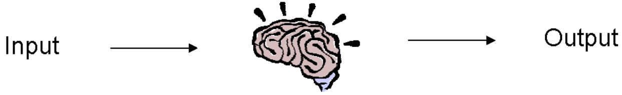
Example Applications

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?

Questions?

Good time to ask now.

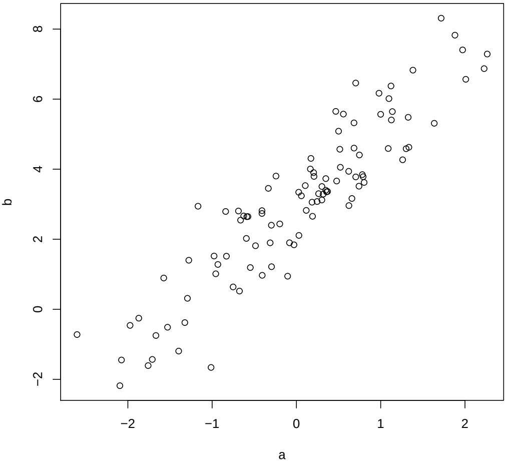
Linear Regression

- Want to find parameters for a function of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Distribution of error random variable not specified

Quick Example - Scatter Plot



Formal Statement of Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i value of the response variable in the i^{th} trial
- β_0 and β_1 are parameters
- X_i is a known constant, the value of the predictor variable in the i^{th} trial
- ϵ_i is a random error term with mean $\mathbb{E}(\epsilon_i)$ and variance $\text{Var}(\epsilon_i) = \sigma^2$
- $i = 1, \ldots, n$

Properties

- The response Y_i is the sum of two components
  - Constant term $\beta_0 + \beta_1 X_i$
  - Random term ϵ_i
- The expected response is

$$\begin{aligned}
\mathbb{E}(Y_i) &= \mathbb{E}(\beta_0 + \beta_1 X_i + \epsilon_i) \\
&= \beta_0 + \beta_1 X_i + \mathbb{E}(\epsilon_i) \\
&= \beta_0 + \beta_1 X_i
\end{aligned}$$

Expectation Review

- Definition

$$\mathbb{E}(X) = \mathbb{E}(X) = \int XP(X)dX, \; X \in \mathcal{R}$$

- Linearity property

$$\begin{aligned} \mathbb{E}(aX) &= a\,\mathbb{E}(X) \\ \mathbb{E}(aX + bY) &= a\,\mathbb{E}(X) + b\,\mathbb{E}(Y) \end{aligned}$$

- Obvious from definition

Example Expectation Derivation

$$P(X) = 2X, 0 \leq X \leq 1$$

Expectation

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^1 XP(X)dX \\
&= \int_0^1 2X^2 dX \\
&= \frac{2X^3}{3}|_0^1 \\
&= \frac{2}{3}
\end{aligned}$$

Expectation of a Product of Random Variables

If X,Y are random variables with joint distribution $P(X, Y)$ then the expectation of the product is given by

$$\mathbb{E}(XY) = \int_{XY} XYP(X, Y)dXdY.$$

Expectation of a product of random variables

What if X and Y are independent? If X and Y are independent with density functions f and g respectively then

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{XY} XYf(X)g(Y)dXdY \\
&= \int_X \int_Y XYf(X)g(Y)dXdY \\
&= \int_X Xf(X)[\int_Y Yg(Y)dY]dX \\
&= \int_X Xf(X)\,\mathbb{E}(Y)dX \\
&= \mathbb{E}(X)\,\mathbb{E}(Y)
\end{aligned}$$

Regression Function

- The response Y_i comes from a probability distribution with mean

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i$$

- This means the regression function is

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X$$

Since the regression function relates the means of the probability distributions of Y for a given X to the level of X

Error Terms

- The response Y_i in the i^{th} trial exceeds or falls short of the value of the regression function by the error term amount ϵ_i
- The error terms ϵ_i are assumed to have constant variance σ^2

Response Variance

Responses Y_i have the same constant variance

$$\begin{aligned}\text{Var}(Y_i) &= \text{Var}(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= \text{Var}(\epsilon_i) \\ &= \sigma^2\end{aligned}$$

Variance (2^{nd} central moment) Review

- Continuous distribution

$$\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \int (X - \mathbb{E}(X))^2 P(X) dX, \; X \in \mathcal{R}$$

- Discrete distribution

$$\text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \sum_i (X_i - \mathbb{E}(X))^2 P(X_i), \; X \in \mathcal{Z}$$

Alternative Form for Variance

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}((X-\mathbb{E}(X))^2) \\
&= \mathbb{E}((X^2-2X\,\mathbb{E}(X)+\mathbb{E}(X)^2)) \\
&= \mathbb{E}(X^2)-2\,\mathbb{E}(X)\,\mathbb{E}(X)+\mathbb{E}(X)^2 \\
&= \mathbb{E}(X^2)-2\,\mathbb{E}(X)^2+\mathbb{E}(X)^2 \\
&= \mathbb{E}(X^2)-\mathbb{E}(X)^2.
\end{aligned}$$

Example Variance Derivation

$$P(X) = 2X, 0 \leq X \leq 1$$

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
&= \int_0^1 2XX^2 dX - (\frac{2}{3})^2 \\
&= \frac{2X^4}{4}|_0^1 - \frac{4}{9} \\
&= \frac{1}{2} - \frac{4}{9} = \frac{1}{18}
\end{aligned}$$

Variance Properties

$$\begin{aligned}
\mathsf{Var}(aX) &= a^2\,\mathsf{Var}(X) \\
\mathsf{Var}(aX+bY) &= a^2\,\mathsf{Var}(X) + b^2\,\mathsf{Var}(Y)\,\mathrm{if} X \perp Y \\
\mathsf{Var}(a+cX) &= c^2\,\mathsf{Var}(X)\ \mathrm{if} a, c \text{ both constant}
\end{aligned}$$

More generally

$$\mathsf{Var}(\sum a_i X_i) = \sum_i \sum_j a_i a_j \mathrm{Cov}(X_i, X_j)$$

Covariance

- The covariance between two real-valued random variables X and Y, with expected values $\mathbb{E}(X) = \mu$ and $\mathbb{E}(Y) = \nu$ is defined as

$$Cov(X, Y) = \mathbb{E}((X - \mu)(Y - \nu))$$

- Which can be rewritten as

$$\begin{aligned} Cov(X, Y) &= \mathbb{E}(XY - \nu X - \mu Y + \mu\nu), \\ Cov(X, Y) &= \mathbb{E}(XY) - \nu\,\mathbb{E}(X) - \mu\,\mathbb{E}(Y) + \mu\nu, \\ Cov(X, Y) &= \mathbb{E}(XY) - \mu\nu. \end{aligned}$$

Covariance of Independent Variables

If X and Y are independent, then their covariance is zero. This follows because under independence

$$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y) = \mu\nu.$$

and then

$$Cov(XY) = \mu\nu - \mu\nu = 0.$$