Stat GU4205/5205 Lecture 1

Jingchen Liu

Department of Statistics
Columbia University

- Textbook: *Applied Linear Regression Models*, fourth edition, Kutner, Nachtsheim, and Neter
- Teaching assistants: Guanhua Fang
- Course design: statistics MA
- Office hours: TBD

Syllabus

- Prerequisites: multivariate calculus, linear algebra, Stat GU4203, and GU4204
- Course material: simple linear regression, multiple linear regression, model diagnostics, model selection, and all other related statistical issues.
- Statistical inference: point estimation, hypothesis testing, confidence interval
- Basic statistical concepts
- Software: R http://cran.r-project.org/

Syllabus

- Prerequisites: multivariate calculus, linear algebra, Stat GU4203, and GU4204
- Course material: simple linear regression, multiple linear regression, model diagnostics, model selection, and all other related statistical issues.
- Statistical inference: point estimation, hypothesis testing, confidence interval
- Basic statistical concepts
- Software: R http://cran.r-project.org/

Syllabus

- Prerequisites: multivariate calculus, linear algebra, Stat GU4203, and GU4204
- Course material: simple linear regression, multiple linear regression, model diagnostics, model selection, and all other related statistical issues.
- Statistical inference: point estimation, hypothesis testing, confidence interval
- Basic statistical concepts
- Software: R http://cran.r-project.org/

Syllabus

- Prerequisites: multivariate calculus, linear algebra, Stat GU4203, and GU4204
- Course material: simple linear regression, multiple linear regression, model diagnostics, model selection, and all other related statistical issues.
- Statistical inference: point estimation, hypothesis testing, confidence interval
- Basic statistical concepts
- Software: R http://cran.r-project.org/

Syllabus

- Prerequisites: multivariate calculus, linear algebra, Stat GU4203, and GU4204
- Course material: simple linear regression, multiple linear regression, model diagnostics, model selection, and all other related statistical issues.
- Statistical inference: point estimation, hypothesis testing, confidence interval
- Basic statistical concepts
- Software: R   http://cran.r-project.org/

Grading

- Homework: 30%
- Midterm: 30%
- Final: 40%

About statistics

- Statistics is "a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data" – Wikipedia
- Statistical modeling: capturing the pattern of data for interpretation and prediction

About statistics

- Statistics is “a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data” – Wikipedia
- Statistical modeling: capturing the pattern of data for interpretation and prediction

Single variable

- Random versus deterministic
- Probability: the chance of raining tomorrow is 40%.
- Distribution: characterizing the behavior of a random variable (object).

Single variable

- Random versus deterministic
- Probability: the chance of raining tomorrow is 40%.
- Distribution: characterizing the behavior of a random variable (object).

Single variable

- Random versus deterministic
- Probability: the chance of raining tomorrow is 40%.
- Distribution: characterizing the behavior of a random variable (object).

Daily log-return, December 31, 2013

$$\log(S_{today} / S_{yesterday})$$

IBM: 0.6%, AAPL: 0.6%, GS: 0.8%, BAC: 0.1% ...

Graphical illustration

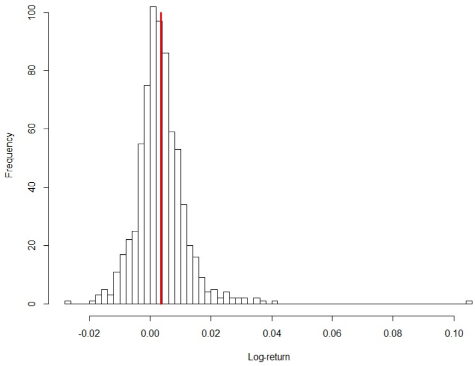


Figure: Histogram of a single random variable

Basic statistics

$$x_1, \ldots, x_n$$

- Sample mean:

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Sample variance:

$$s^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- Sample standard deviation: s

Basic statistics

$$x_1, \ldots, x_n$$

- Sample mean:

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Sample variance:

$$s^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- Sample standard deviation: s

Basic statistics

$$x_1, \ldots, x_n$$

- Sample mean:

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Sample variance:

$$s^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- Sample standard deviation: s

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Basic statistics

- Median
  - $x_1, \ldots, x_{2n+1}$: $m = x_n$
  - $x_1, \ldots, x_{2n}$: $m = \frac{x_n + x_{n+1}}{2}$
- Quantile

Basic statistics

- Median
  - $x_1, ..., x_{2n+1}$: $m = x_n$
  - $x_1, ..., x_{2n}$: $m = \frac{x_n + x_{n+1}}{2}$
- Quantile

Basic statistics

- Median
  - $x_1, ..., x_{2n+1}$: $m = x_n$
  - $x_1, ..., x_{2n}$: $m = \frac{x_n + x_{n+1}}{2}$
- Quantile

Basic statistics

- Median
  - $x_1, ..., x_{2n+1}$: $m = x_n$
  - $x_1, ..., x_{2n}$: $m = \frac{x_n + x_{n+1}}{2}$
- Quantile

Two variables

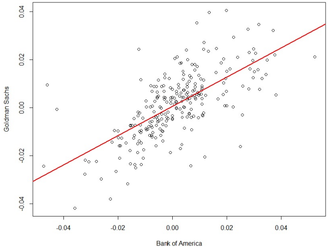


Figure: 2013 Bank of America versus Goldman Sachs

- Deterministic versus random
- Predictability

Two variables

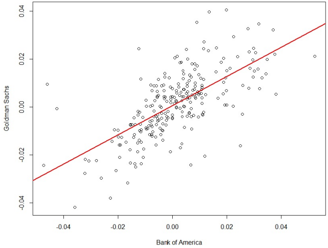


Figure: 2013 Bank of America versus Goldman Sachs

- Deterministic versus random
- Predictability

Two variables

$$(x_1, y_1), \ldots, (x_n, y_n)$$

- Covariance

$$C_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Correlation

$$\rho_{x,y} = \frac{C_{x,y}}{s_x s_y}$$

Two variables

$$(x_1, y_1), \ldots, (x_n, y_n)$$

- Covariance

$$C_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Correlation

$$\rho_{x,y} = \frac{C_{x,y}}{s_x s_y}$$

Two variables

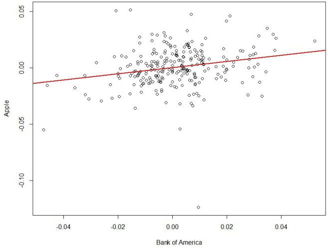


Figure: 2013 Bank of America versus Apple

A nonlinear example

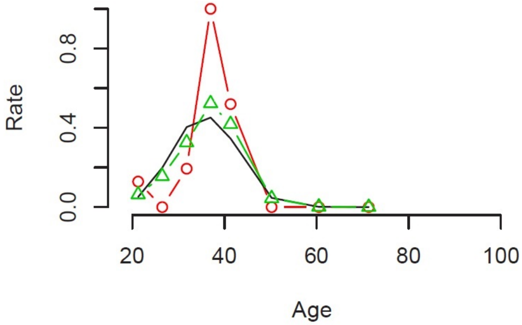


Figure: Major depression prevalence rate against age

Another nonlinear example

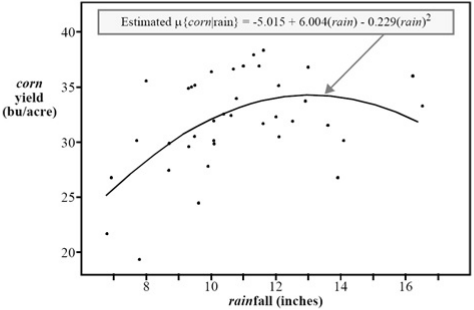


Figure: Corn yield against rain fall

Regression models

- Separating the predictable part from the random part
- Basic setting:
  - Predictor (covariates, independent variable): X
  - Response variable (dependent variable): Y

Regression models

- Separating the predictable part from the random part
- Basic setting:
  - Predictor (covariates, independent variable): X
  - Response variable (dependent variable): Y

Regression models

- Separating the predictable part from the random part
- Basic setting:
  - Predictor (covariates, independent variable): X
  - Response variable (dependent variable): Y

Regression models

- Separating the predictable part from the random part
- Basic setting:
  - Predictor (covariates, independent variable): X
  - Response variable (dependent variable): Y

The use of regression models

- Causality versus association
  - Clinical trial
  - Genetic association study
  - Economics
- Prediction
- Signal detection, variable selection

The use of regression models

- Causality versus association
  - Clinical trial
  - Genetic association study
  - Economics
- Prediction
- Signal detection, variable selection

The use of regression models

- Causality versus association
  - Clinical trial
  - Genetic association study
  - Economics
- Prediction
- Signal detection, variable selection

The use of regression models

- Causality versus association
  - Clinical trial
  - Genetic association study
  - Economics
- Prediction
- Signal detection, variable selection

The use of regression models

- Causality versus association
  - Clinical trial
  - Genetic association study
  - Economics
- Prediction
- Signal detection, variable selection

The use of regression models

- Causality versus association
  - Clinical trial
  - Genetic association study
  - Economics
- Prediction
- Signal detection, variable selection

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

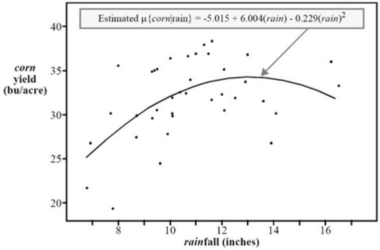
About causality

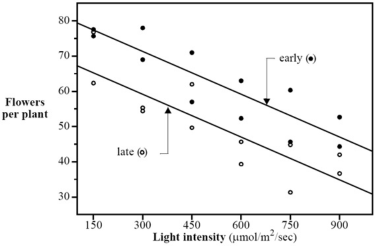
- Experimental study
- Observational study

About causality

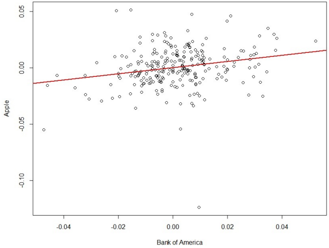
- Experimental study
- Observational study

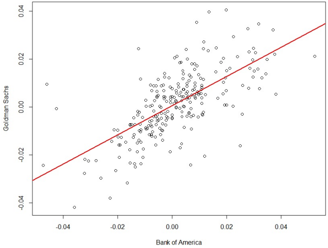
Experimental study versus obesrvational study



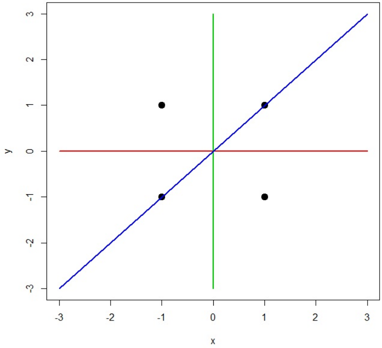


Line fitting





Simple example



Least squares estimator

$$(x_1, y_1), \ldots, (x_n, y_n)$$

- Fitting a straight line $y = \beta_0 + \beta_1 x$
- Sum of squares of residuals

$$SS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Least squares estimate

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Least squares estimator

$$(x_1, y_1), \ldots, (x_n, y_n)$$

- Fitting a straight line $y = \beta_0 + \beta_1 x$
- Sum of squares of residuals

$$SS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Least squares estimate

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Least squares estimator

$$(x_1, y_1), \dots, (x_n, y_n)$$

- Fitting a straight line $y = \beta_0 + \beta_1 x$
- Sum of squares of residuals

$$SS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Least squares estimate

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Another estimator

- Least squares estimate

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|.$$

- Sample mean and median.

Another estimator

- Least squares estimate

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|.$$

- Sample mean and median.

The rationale

- Prediction
- From the viewpoint of design, x is preset and y is observed.
- Probabilistic interpretation

The rationale

- Prediction
- From the viewpoint of design, x is preset and y is observed.
- Probabilistic interpretation

The rationale

- Prediction
- From the viewpoint of design, x is preset and y is observed.
- Probabilistic interpretation

- Derivation of least squares estimator

Least squares estimator for simple linear regression

- The slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Interpretation

Least squares estimator for simple linear regression

- The slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Interpretation

Least squares estimator for simple linear regression

- The slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- The intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Interpretation

Least squares estimator for simple linear regression

- Another representation

$$\hat{\beta}_1 = \rho_{x,y} \frac{s_y}{s_x}.$$

- The fitted regression line

$$(x - \bar{x}) = \rho_{x,y} \frac{s_y}{s_x} (y - \bar{y})$$

Least squares estimator for simple linear regression

- Another representation

$$\hat{\beta}_1 = \rho_{x,y} \frac{s_y}{s_x}.$$

- The fitted regression line

$$(x - \bar{x}) = \rho_{x,y} \frac{s_y}{s_x} (y - \bar{y})$$