ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

15-Jan-2018

Week 1

Week 1: Introduction and Simple Linear Regression

Week 1: Introduction and Simple Linear Regression

Introduction

- Instructor Information
- Prerequisites
- Software
- Textbook
- Evaluation

Instructor Information

- Instructor: Choi, Yunjin
- Office: s16/06-115
- Phone: 6516-8949
- email: stachoiy@nus.edu.sg (preferred)

Prerequisites

- Calculus
- Linear Algebra
- Statistics

(at undergraduate level)

Software: R

- it's a freeware!
- can be downloaded from http://cran.r-project.org
- An introduction to R is available at http://cran.r-project.org/doc/manuals/R-intro.pdf

Textbook

- Applied Linear Regression Models
  by Kutner, Nachtsheim and Neter. Fourth edition. McGraw-Hill

Evaluation

- Homework (10%): no late submission
- Midterm (30%):
  tentatively on 5th of March or 12th of March (in-class)
- Final Exam (60 %): scheduled on 5th of May (Sat) at 1:00pm
  close book exam, one A4 size papers (on both sides) are allowed. A non-programmable calculator is allowed and necessary.
- Grading policy: once you make a regrade request, not only the particular question of your interest but your entire exam will be subject to regrading. Please consider seriously before you make a request as your exam score might go down. When you would like to make a request referring to other student's grading, you need to bring the whole exam paper that you are referring to. The entire range of the referred exam paper is also subject to regrading, and its score might go down.

Evaluation

- Homework (10%): no late submission
- Midterm (30%):
  tentatively on 5th of March or 12th of March (in-class)
- Final Exam (60 %): scheduled on 5th of May (Sat) at 1:00pm
  close book exam, one A4 size papers (on both sides) are allowed. A non-programmable calculator is allowed and necessary.
- Grading policy: once you make a regrade request, not only the particular question of your interest but your entire exam will be subject to regrading. Please consider seriously before you make a request as your exam score might go down. When you would like to make a request referring to other student's grading, you need to bring the whole exam paper that you are referring to. The entire range of the referred exam paper is also subject to regrading, and its score might go down.

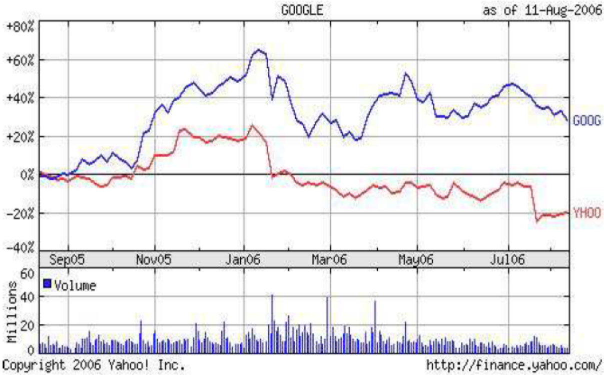
NOTE: BOTH EXAMS ARE REQUIRED TO PASS THE COURSE!

What is “Regression Analysis”?

- Want to model a functional relationship between a “predictor variable” (independent variable, usually denoted by X) and a “response variable” (dependent variable, usually denoted by Y).
- Goals
  - Description: to figure out the relationship between a predictor variables and responses
  - Prediction: to predict the unknown future response given the new observed predictors.

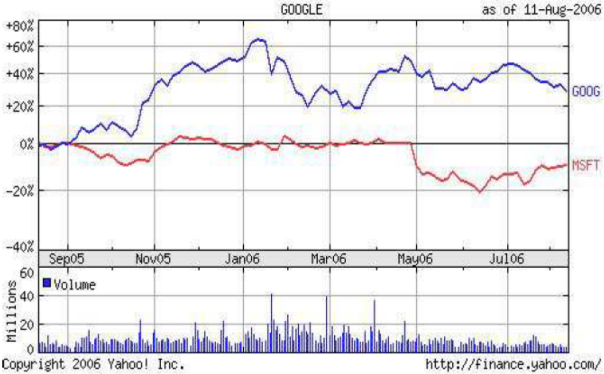
Example

GOOGLE and YAHOO



Example

GOOGLE and MICROSOFT



Other Applications

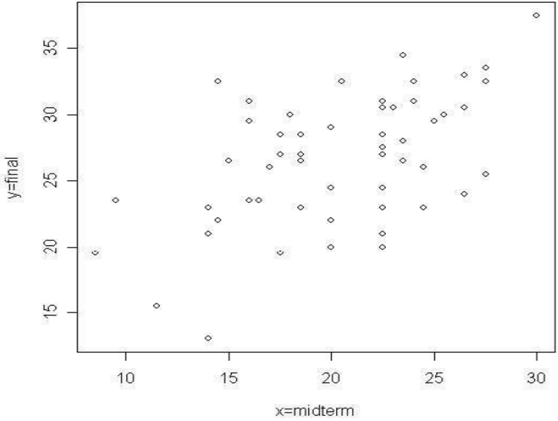
- Exam scores
  - investigate relationship between midterm score and final exam score
- Education and income
  - investigate relationship between education level and income
- Heights of father and son
  - relate heights of father and son

Example 1

Table: Grade book from a statistical course “ST500”

ID	midterm	final	hw	total
1	24.5	26.0	28.5	79.0
2	22.5	24.5	28.2	75.2
3	23.5	26.5	28.3	78.3
4	23.5	34.5	29.2	87.2
		.		
		.		
		.		
54	26.5	33	27.5	87.0
55	23.5	28	24.3	75.8

Example 1 continued
Graphical display–scatter plot



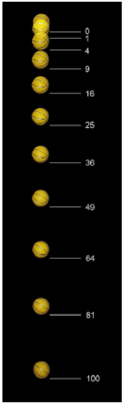
Week 1: Introduction and Simple Linear Regression

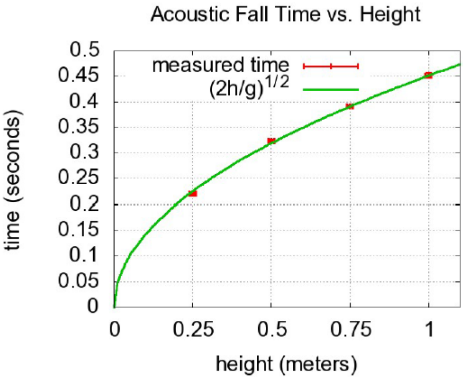
Comparison with a famous experiment in Physics

How long does it take
for an object to fall from d meter?

$$t = \sqrt{2 * d/9.8}$$

In an ideal world...





Functional Relationship vs. Statistical Relationship

- Functional Relationship
  $Y = f(X)$
- Statistical Relationship
  $Y = f(X) +$ some variation/uncertainty/unknown
  $Y = f(X) + \epsilon$, ϵ is a random variable with a p.d.f and $E[\epsilon] = 0$

Let’s revisit the Physical experiment

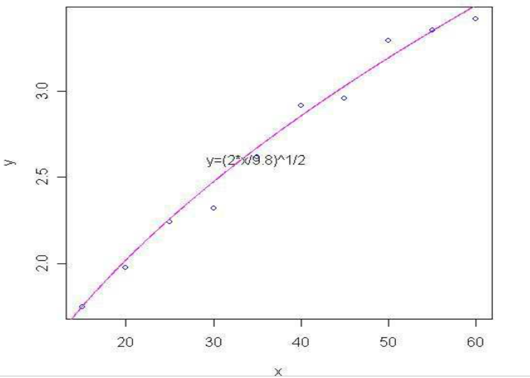
Table: Falling Body Data

No	Y	X
1	1.744571	15
2	1.973905	20
3	2.239605	25
4	2.317665	30
5	2.617552	35
6	2.915219	40
7	2.955634	45
8	3.293491	50
9	3.355540	55
10	3.420441	60

Scatter plot

$$y_i = \sqrt{2 * d/9.8} + \epsilon_i, i = 1, ..., 10$$

ϵ - measurement error, air resistance, etc.



Simple Linear Regression (SLR) model

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

- Y_i is the value of the response variable in the ith trial
- X_i is a known constant (fixed),
  the value of the predictor variable in the ith trial
- ϵ_i is a random error term
  $E[\epsilon_i] = 0$, $Var[\epsilon] = \sigma^2$, $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$

Simple Linear Regression (SLR) model

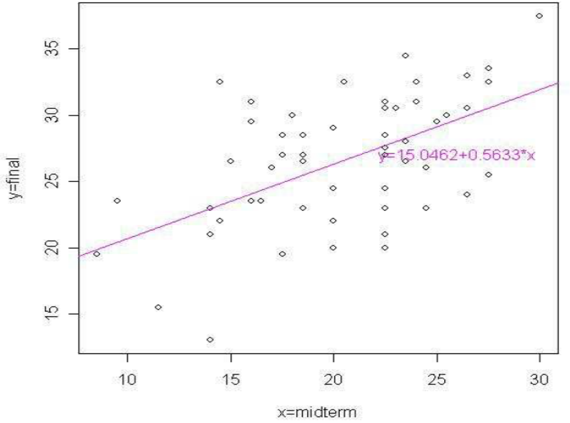
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Regression function: $f(X) = \beta_0 + \beta_1 X$
- Regression parameters: β_0 and β_1

  β_0–intercept: the expected value of Y when $X = 0$

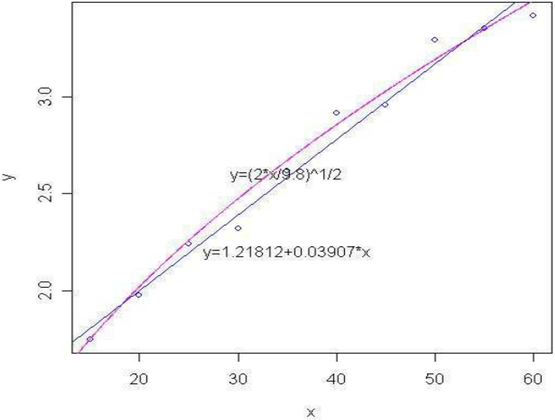
  β_1–slope: the change of $E[Y]$ when 1 unit increase in X

Back to score example



Physical Experiment Data

Simple linear regression–first-order model



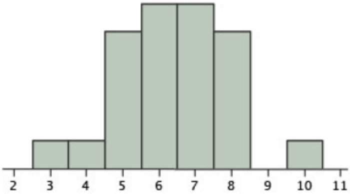
Regression Effect

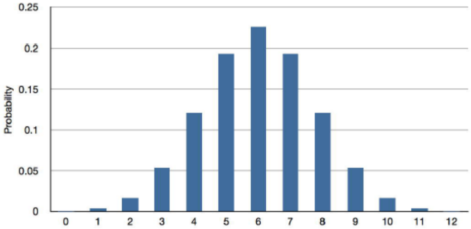
In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test and the top group will on average fall back. This effect is known as the regression toward the mean (a.k.a regression effect)

Week 1: Introduction and Simple Linear Regression

Example 1

- Outcome (purely by chance): 12 coin flipping prediction (25 subjects).
- The best prediction of the re-test for the subject scored highest?





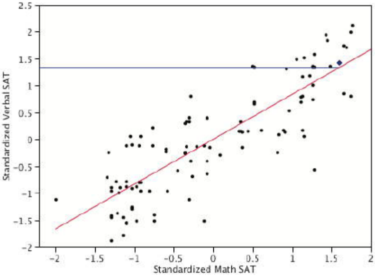
Week 1: Introduction and Simple Linear Regression

Example 2

- Outcome (half chance+half skill): 6 coin flipping prediction + 6 TURE/FALSE math tests (25 subjects).
- The best prediction of the re-test for a subject scored e.g., 10/12?

Example 3

- Outcome (chance+skill): Verbal SAT versus MATH SAT score
- $Y_{\text{Verbal}} = rX_{\text{Math}}$
- $r = 0.835$



Regression/correlation vs. causality/causation

- Causation: football weekend → heavier traffic/more food sales
- Regression $\neq$ Causation
    - midterm scores/final scores
    - ice cream sales/number of shark attacks on swimmers
    - Skirt lengths/ stock prices (as skirt lengths in a country get shorter, stock prices go up)
    - The number of cavities in elementary children/ vocabulary size

Week 1: Introduction and Simple Linear Regression

Which drug is better?

Drug A and B both treat high blood pressure.

- Study 1: Examine all the patients in a hospital who has taken A or B for a certain amount of time and make the comparison based on average fitness of each group.
- Study 2: Conduct an experiment on 100 voluntary patients with high blood pressure. Randomly assign 50 of them to take drug A while the rest of them take drug B.

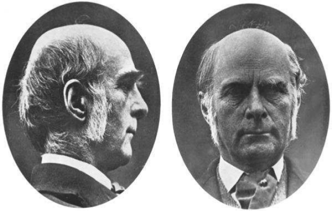
Observational Data vs. Experimental Data

- Study 1: observational study, no control over X which is an explanatory variable.
  Need extra caution to investigate cause and effect.
- Study 2: experimental study, can control Xwhich is an explanatory variable.

History

Sir Francis Galton (1822-1911)

He first introduced statistical concepts regression/correlation

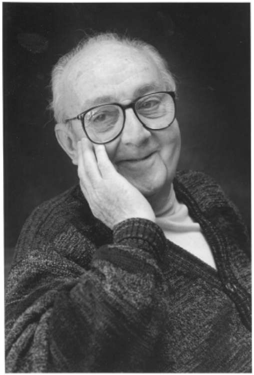


- Studied the relation between heights of parents and children and noted that the children regressed to the population mean. He developed a mathematical description of this regression tendency.
- The term "regression" persists to this day to describe statistical relations between variables.

Week 1: Introduction and Simple Linear Regression

George E. P. Box (1919-2013)

Essentially, all models are wrong, but some are useful.



Independence

- Independence: two random variables x and y are said to be independent ($x \perp y$) if
  - The joint probability density $=$ the product of the two marginal densities of x and y (x and y are continuous random variables); or
  - The joint probability mass function$=$ the product of the two marginal probability mass functions of x and y (x and y are discrete random variables).
- Property of independence: if $x \perp y$, then $f(x) \perp g(y)$, where $f(\cdot)$ and $g(\cdot)$ are both continuous functions.

Uncorrelated vs. Independent

- Uncorrelated: $Cov(X, Y) = 0$
- If independent, then uncorrelated.
  i.e., If $X \perp Y$, then $Cov(X, Y) = 0$
- BUT, the reverse does not hold:
  $Cov(X, Y) = 0$ does not imply $X \perp Y$.
- NOTE: Uncorrelated + Bivariate normal ⇒ Independent
  If $Cov(X, Y) = 0$ and X and Y are both normally distributed,
  then $X \perp Y$.

Useful formulas of $E(\cdot)$, $Var(\cdot)$, and $Cov(\cdot)$

Let $x, x_1, \cdots, x_n$ be **random variables**, and $a, a_1, \cdots, a_n, b, b_1, \cdots, b_n$ be **real numbers**.

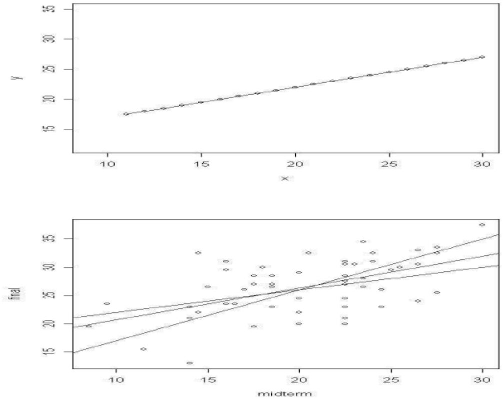
- $E(ax+b) = aE(x) + b$
  $E\left(\sum_{i=1}^n (a_i x_i + b_i)\right) = \sum_{i=1}^n a_i E(x_i) + \sum_{i=1}^n b_i$
- $Var(ax+b) = a^2 Var(x)$
- If $x_1, \cdots x_n$ are **independent**, then
  - $Var(\sum_{i=1}^n a_i x_i + b) = \sum_{i=1}^n a_i^2 Var(x_i)$, and
  - $Cov(\sum_{i=1}^n a_i x_i, \sum_{i=1}^n b_i x_i) = \sum_{i=1}^n a_i b_i Var(x_i)$

Week 1: Introduction and Simple Linear Regression

Simple Linear Regression (SLR) model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- X_i fixed
  $E(\epsilon_i) = 0,\ Var(\epsilon_i) = \sigma^2,\ Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$
- Regression function: $\beta_0 + \beta_1 X$
- Regression parameters: β_0 and β_1
  β_0–intercept, β_1–slope

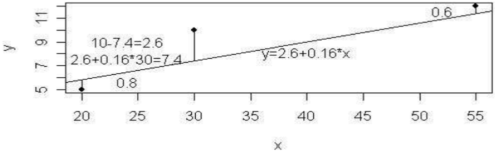
Estimation of β_0, β_1–finding the best line

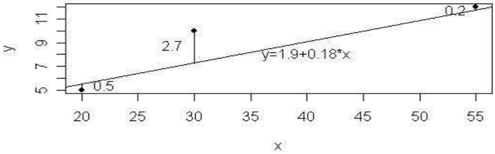


Week 1: Introduction and Simple Linear Regression

Which line is better?

Y	5	12	10
X	20	55	30





Least squares principle

- Use $\beta_0 + \beta_1 X_i$ to approximate Y_i :

  Deviations: $Y_i - \beta_0 - \beta_1 X_i$

- LS principle - find best (β_0, β_1) by minimizing

  $$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

  $$0.8^2 + 2.6^2 + 0.6^2 = 7.76 > 0.5^2 + 2.7^2 + 0.2^2 = 7.58$$

- Did we use LS principle before? (Hint: center)

Mathematical solution - least squares estimator of β_0, β_1

- Partial derivatives:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)$$
$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}X_i(Y_i - \beta_0 - \beta_1 X_i)$$

- LS estimator b_0, b_1 solves **normal equations**:

$$-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$
$$-2\sum_{i=1}^{n}X_i(Y_i - b_0 - b_1 X_i) = 0$$

Week 1: Introduction and Simple Linear Regression

Least squares estimator in Simple Linear Regression

- $b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad b_0 = \bar{Y} - b_1\bar{X} \; (\bar{Y} = b_0 + b_1\bar{X})$
- Properties - unbiasedness

  $E\{b_1\} = \beta_1, \; E\{b_0\} = \beta_0$
- The regression line goes through $(\bar{X}, \bar{Y})$

Week 1: Introduction and Simple Linear Regression

Estimated (fitted) regression line

- Regression line - $\beta_0 + \beta_1 X$
- Estimated (fitted) regression function - $b_0 + b_1 X$

  $E\{b_0 + b_1 X\} = \beta_0 + \beta_1 X$
- Observations Y_i versus fitted values $\hat{Y}_i = b_0 + b_1 X_i$
- Residuals $e_i = Y_i - \hat{Y}_i$ (same as ϵ_i?)

Bias Variance Trade-off

- For an estimator $\hat{\theta}$ of a parameter θ, the mean squared error (MSE) of $\hat{\theta}$ is as follows: $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$.
- MSE can be decomposed: $MSE(\hat{\theta}) = Var(\hat{\theta}) + bias(\hat{\theta})^2$

Properties of the solution

Assuming $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ for all i, and $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, b_0, and b_1 are the **Best Linear Unbiased Estimator (BLUE)**, that is

- Unbiased: $E(b_0) = \beta_0$
- Best: b_0 and b_1 have minimum variance among all unbiased linear estimators

((1.11) in the text book)

Week 1: Introduction and Simple Linear Regression

Properties of the LS estimators

- $\sum_{i=1}^{n} e_i = 0$
  $(\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = \sum_{i=1}^{n} Y_i - nb_0 - b_i \sum_{i=1}^{n} X_i = 0)$
- $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$
- $\sum_{i=1}^{n} X_i e_i = 0$
  $(\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = \sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0)$
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$
  $\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n}(b_0 + b_1 X_i) e_i = b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} X_i e_i = 0$
- The regression line always pass through the point $(\bar{X}, \bar{Y})$

Return to score example - what does “regression” mean?

- Regression line in original space: $\hat{Y}_i = b_0 + b_1 X_i$
- Regression line after mean centering: $\hat{Y}_i - \bar{Y} = b_1(X_i - \bar{X})$

  Remark: $\hat{Y}_i = b_0 + b_1\bar{X} + b_1(X_i - \bar{X}) = \bar{Y} + b_1(X_i - \bar{X})$
- Regression line after standardization [3]: $\frac{\hat{Y}_i - \bar{Y}}{\mathrm{sd}(Y)} = b_1^* \frac{X_i - \bar{X}}{\mathrm{sd}(X)}$

  where $\mathrm{sd}(X) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$, $\mathrm{sd}(Y) = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$

  $$b_1^* = \rho(X, Y) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

[3] easy to verify this, and $b_1 = \rho(X, Y)\mathrm{sd}(Y)/\mathrm{sd}(X)$

Estimating the variance

- Single Population:
  Estimate the population variance by the sample variance s^2 where
  $s^2 = \frac{\sum_1^n (Y_i - \bar{Y})^2}{n-1}$
- Regression Model:
  - Estimate σ^2, the variance of each observation by MSE:
    $s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_1^n (Y_i - \hat{Y}_i)^2}{n-2}$
    where SSE stands for error sum of squares
    $(SSE = \sum_1^n (Y_i - \hat{Y}_i)^2 = \sum_1^n e_i^2)$
  - MSE is an unbiased estimator of σ^2
    i.e., $E[MSE] = \sigma^2$
  - Note the denominator $n - 2$. It represents the associated "degrees of freedom"

Week 1: Introduction and Simple Linear Regression

Simple linear Regression with normally distributed errors

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, ..., n$

- No assumption on forms of probability distributions of ϵ_i

- What could we do if we assume ϵ's are distributed as $N(0, \sigma^2)$

- Reasons for normal assumption: justifiable in many real applications

Week 1: Introduction and Simple Linear Regression

Review of Maximum Likelihood Estimator

- Example 1: Bag1 (three red balls and seven black balls)
  Bag2 (five red balls and five black balls)
  From one of these bags randomly pick a red ball. Which bag?
  $\frac{3}{10} < \frac{5}{10}$

- Two independent observations from the same probability distribution (either $N(0,1)$ or $N(1,2)$). Two observations are $-1, 2$. Which distribution?

$$\frac{1}{\sqrt{2\pi \times 1}} e^{\frac{(-1-0)^2}{2}} \frac{1}{\sqrt{2\pi \times 1}} e^{\frac{(2-0)^2}{2}} > \frac{1}{\sqrt{2\pi \times 2}} e^{\frac{(-1-1)^2}{2\times 2}} \frac{1}{\sqrt{2\pi \times 2}} e^{\frac{(2-1)^2}{2\times 2}}$$

Week 1: Introduction and Simple Linear Regression

Maximum Likelihood Estimation for single population

- $Y_1, \ldots, Y_n$ independently from $N(\mu, 1)$

- Maximum likelihood estimate of μ: maximize likelihood function:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{(Y_i - \mu)^2}{2}}$$

- Maximize log-likelihood function

$$\ell(\mu) = \log L(\mu) = \sum_{i=1}^{n} \{- \log \sqrt{2\pi} + \frac{(Y_i - \mu)^2}{2}\}$$

- Differentiation: $\frac{\partial \ell(\mu)}{\partial \mu}$

Maximum Likelihood Estimation for simple linear regression

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, ..., n$ independently from $N(0, \sigma^2)$
- Maximize log-likelihood function

  $\ell(\beta_0, \beta_1, \sigma^2) = \log L(\beta_0, \beta_1, \sigma^2) =$

  $-\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$

Maximum Likelihood Estimator $\hat{\beta}_0, \hat{\beta}_2, \hat{\sigma}^2$

- $$\frac{\partial \ell}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$
  $$\frac{\partial \ell}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i)$$
  $$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$
- $$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$
  $$\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$
  $$-n\sigma^2 + \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$
  $$\rightarrow \hat{\beta}_0 = b_0, \hat{\beta}_1 = b_1, \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2}{n}$$

Reading Materials: Some Important Distributions

Normal Distribution

- Normal distribution $N(\mu, \sigma^2)$: probability density function (pdf)

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(x-\mu)^2/(2\sigma^2)\}.$$

- If $x \sim N(\mu, \sigma^2)$, then $(x-\mu)/\sigma \sim N(0, 1)$.

χ^2 Distribution

- Assume $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d.) random variables from $N(0,1)$.
- Definition: $x_1^2 \sim \chi_1^2$ and $\sum_{i=1}^n x_i^2 \sim \chi_n^2$, where n is called the degrees of freedom (d.f.).
- Property 1: if $y_1 \sim \chi_n^2$ and $y_2 \sim \chi_m^2$, and $y_1 \perp y_2$, then $y_1 + y_2 \sim \chi_{n+m}^2$.
- Property 2: if $y \sim \chi_n^2$, then $E(y) = n$.

t Distribution

- If $x \sim N(0,1)$ and $y \sim \chi^2_n$ and they are independent, then
$$x/\sqrt{y/n} \sim t_n,$$
the t distribution with degrees of freedom n.
- if $n > 1$, the expectation of t_n distribution is 0.

F Distribution

- If $y_1 \sim \chi^2_n$ and $y_2 \sim \chi^2_m$, and they are independent, then
$$\frac{y_1/n}{y_2/m} \sim F_{n,m},$$
the F distribution with degrees of freedom m and n.