STA302/1001 - Methods of Data Analysis I

(Week 01 - Lecture A)

Wei (Becky) Lin

May 15-19, 2017

About me

- Wei(Becky) Lin
- PhD and MSc degrees in statistics and a BSEc degree in computer science.
- Now an assistant professor, in teaching stream at UTSG.
- Research interests: likelihood inference, statistical computing and graphics/data visualization, machining learning, survey data analysis, health data analysis.

Notes about syllabus

- **Course syllabus** is available on blackboard, please read it carefully.

  `http://portal.utoronto.ca`

- **Classes**
  - Section L0101: Tuesday/Thursday: 2-5pm in in **MC102**.
  - Section L5101: Monday/Wednesday 6-9pm in in **MC102**.
  - **Office Hours**
  - Me: Monday 11:00am-12:30pm, SS6011 (starts from 2nd week)
  - TAs: Every Wednesday: 3-5pm and OH before assignments/MT/Final.
- **Textbook(s)**
  - *Applied Linear Regression Models*, 5th edition by Kutner, et al.
  - Reference (recommended)
    - *A Modern Approach to Regression with R* by Simon J. Sheather.
    - *Applied linear regression* 4th edition by Sanford Weisberg.

Notes about syllabus

- All course material (syllabus, lecture slides, practice problems and solutions) will be posted on portal
- We will use **Piazza Discussion Board**. This will serve as an on-line forum for questions of general interest (course material, practice problems, etc)
- For all other inquiries come to office hours or speak to me before/after lecture
  - Please do not send me an email if the information can be found on portal or in lecture notes or discussion board.
- If an urgent matter arises, I may contact the entire class by e-mail. In order to receive these message, please make sure that you use your mail.utoronto.ca account and check your mail at least once a day.

Most HW and MT are online marking based.
Do check your UT email often.

Notes about syllabus

- **Computing**: R and R studio software are used for assignments and you need to be able to write R code and interpret R output on the midterm and exams.
  - R and R studio are available for free.
  - We are using basic package in R for this course.
  - R studio is an add-on that make R easier to use for beginner.
  - Please follow the instruction in Week 00 slides to install R, Rstudio and Rmarkdown on your computer after this class. Make sure your could knit to get test.pdf.
  - See the course syllabus to get reference on learning R.
- **Background**: you better have knowledge of following topics
  - Basic probability, at least know Normal, student t, F distribution.
  - Random variables (expectation, variance, covariance, correlation).
  - Point estimate (unbiasedness, MVUE, consistency, BLUE and etc).
  - Maximum likelihood estimation procedure and property of MLE.
  - Inference for mean and variance.
  - First year calculus, good knowledge about matrix and linear algebra.

Marking Scheme

Evaluation

	Weight	Date	Time	location
Midterm	24%	June 5th (L5101), June 6th (L0101)	6-8pm (L5101), 2-4pm (L0101)	EX200 EX200
Assignment 1	3%	Thursday, May 25th	L0101/5101: due 11pm	Crowdmark
DataCamp	3%	June 23rd.	L0101/5101: due 11pm	DataCamp
Assignment 2	10%	Thursday, June 1st	L0101/5101: due 11pm	Crowdmark
Assignment 3	10%	Thursday, June 22nd	L0101/5101: due 11pm	Crowdmark
Final Exam	50%	By A&S on June 9	TBA	TBA

Important dates

- **Midterm (24%):** Monday, June 5th, 6-8pm at EX200 (L5101). Tuesday, June 6th, 2-4pm at EX200 (L5101). Miss MT with Med. notes: weight of it will be shifted to Final.
- **Assignments (26%)**
  - A1: due May 25th (3%)
  - A2: due June 1st (10%).
  - A3: due June 22nd (10%).
  - Complete 3 courses on DataCamp (3%).
  - Late submission penalty: 10% per day.
- **Final exame (50%)**: timetable for F section code courses is available and posted by Art&Sci on June 9th.

Do and Do Not

{Do}

- Attend lecture and take notes.
- Practice problems after every class.
- Practice proofs on your own.
- Write your assignment independently.
- ...

Be Honest !

{Do Not}

- Don't copy, and don't let anyone copy from you.
- It is academic dishonesty to present someone else's work as your own, or to allow your work to be copied for this purpose.
- The person who allows her/his work to be copied is equally guilty, and subject to disciplinary action by the university.
- ....

Course Objective

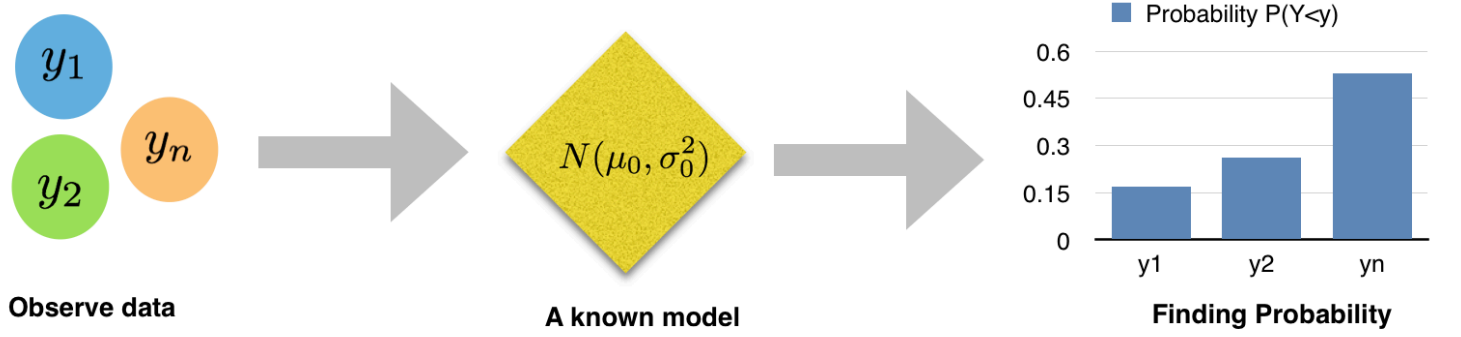
- Course covers a large part of the theory and gain practical skills of developing linear regression models for inference, prediction and interpreting the results.
  - Least squares / MLE estimation.
  - Inference for regression parameters.
  - Model diagnostics and remedial procedure.
  - Multiple linear regression
  - Model building.
- Practical data analysis using R.
  - You will learn basic R to do data analysis in this course.
  - You will learn R markdown to write your assignment.(The lecture slides of this course are created by R markdown too. ^_*)

Connection to pre-requisite course

- Introduction to probability (eg. STA257:learn several distributions, know how to find mean, variance, etc)

Probability:

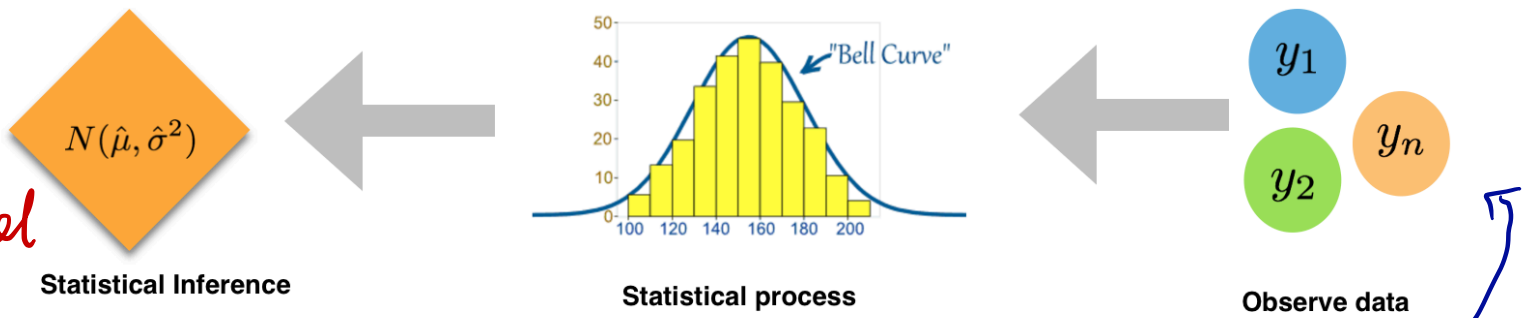
data → Prob.



- Introduction to statistical inference (eg, STA261: know how to estimate model parameter θ, CI,hypothesis testing, etc)

Statistics:

data → infer. model



① $Y_1, \cdots Y_n \sim N(\mu, \sigma^2)$

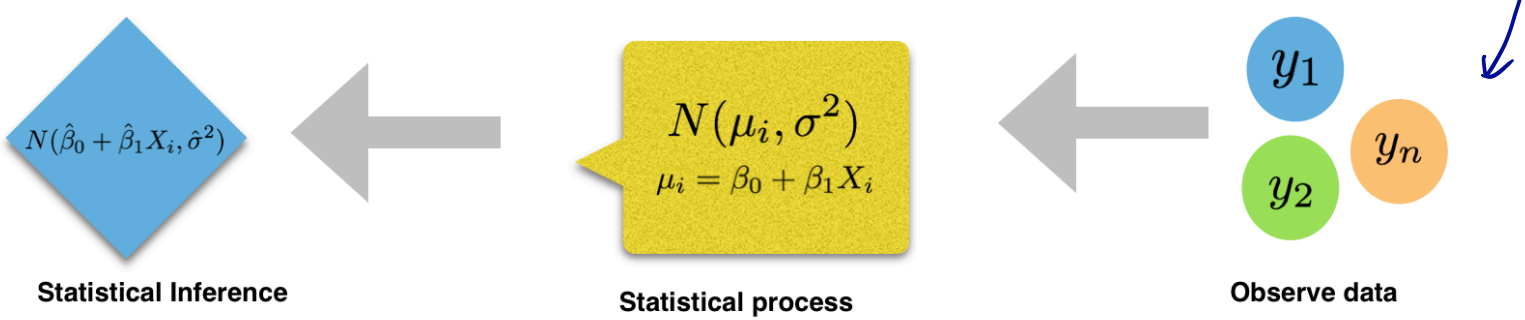
② $\hat{\mu}, \hat{\sigma}^2$

Connection to pre-requisite course

$\beta_0 + \beta_1 x_i$

- $Y_i \sim N(\mu_i, \sigma^2)$
- (1) Given $(X_1, Y_1) \cdots (X_n, Y_n)$
- (2) $\hat{\beta}_0, \hat{\beta}_1, \sigma^2$

- STA302: methods of data analysis I (the major topic is on linear regression)



- Basically, we will carry the same topics that we have in STA261, but only assume that $E(Y|X) = \beta_0 + \beta_1 X$ where β_0, β_1 are assumed to be some constant but unknown.
- Estimation and inference.


ARE YOU
READY?

Chapter 1: Linear Regression with One Predictor Variable

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots n.$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Week 01- Learning objectives & Outcomes

SLR { Population regression line: $Y_i = \beta_0 + \beta x_i + \varepsilon_i$, $i = 1, \cdots n$ (RU)

Estimated regression line: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = b_0 + b_1 x_i$, $i = 1, \cdots n$

- Distinguish between a functional relationship and a statistical relationship.
- Know the Gauss-Markov conditions for simple linear regression.
- Understand the least squares (LS) method.
- Know how to derive and obtain the LS estimates b_0, b_1.
- Show LS estimators b_0 and b_1 are BLUE.
- Recognize the difference between a population regression line and the estimated regression line.
- Interpret the intercept b_0 and slope b_1 of an estimated regression equation.
- Understand the unknown σ^2 and how to get its unbiased estimator.

SLR { unknown population parameters: $\beta_0, \beta_1, \sigma^2$

use data, we find their estimate: $\hat{\beta}_0 = b_0$, $\hat{\beta}_1 = b_1$, $\hat{\sigma}^2 = \frac{SSE}{n-2}$

└ Simple Linear Regression.

What is regression?

- Regression means "going back"
- Linear regression/linear models: a procedure to analyze data
- Historically, *Francis Galton* (1822-1911) invented the term and concepts of regression and correlation.
  - He predicted child's height from fathers height
    - Sons of the tallest fathers tended to be taller than average, but shorter than their fathers.
    - Sons of the shortest fathers tended to be shorter than average, but taller than their fathers.
  - He was deeply concerned about "regression to mediocrity".
  - A brief history of Linear Regression and more about Galton, `http://www.amstat.org/publications/jse/v9n3/stanton.html`
- Regression analysis is a statistical method to summarize and study the relationships between variables in a data set.

Types of relationships

Response and predictor variables

- One variable, denoted Y, is regarded as the response (or outcome, or dependent) variable
  - the variable whose behaviour that we want to study and predict
- The other variable, denoted X, is regarded as the predictor (or explanatory, or independent) variable.
  - variable used to help us study

Relationship between Y and X

- Functional (or deterministic) relationships
  - $Y = f(X)$, where f() is some function. eg. Circumference=$\pi \times$ diameter.
- Statistical Relationship
  - $Y = f(X) + \epsilon$,where ϵ is the random error term. eg. SLR model.

↑ f: could be linear/nonlinear, this course: linear.

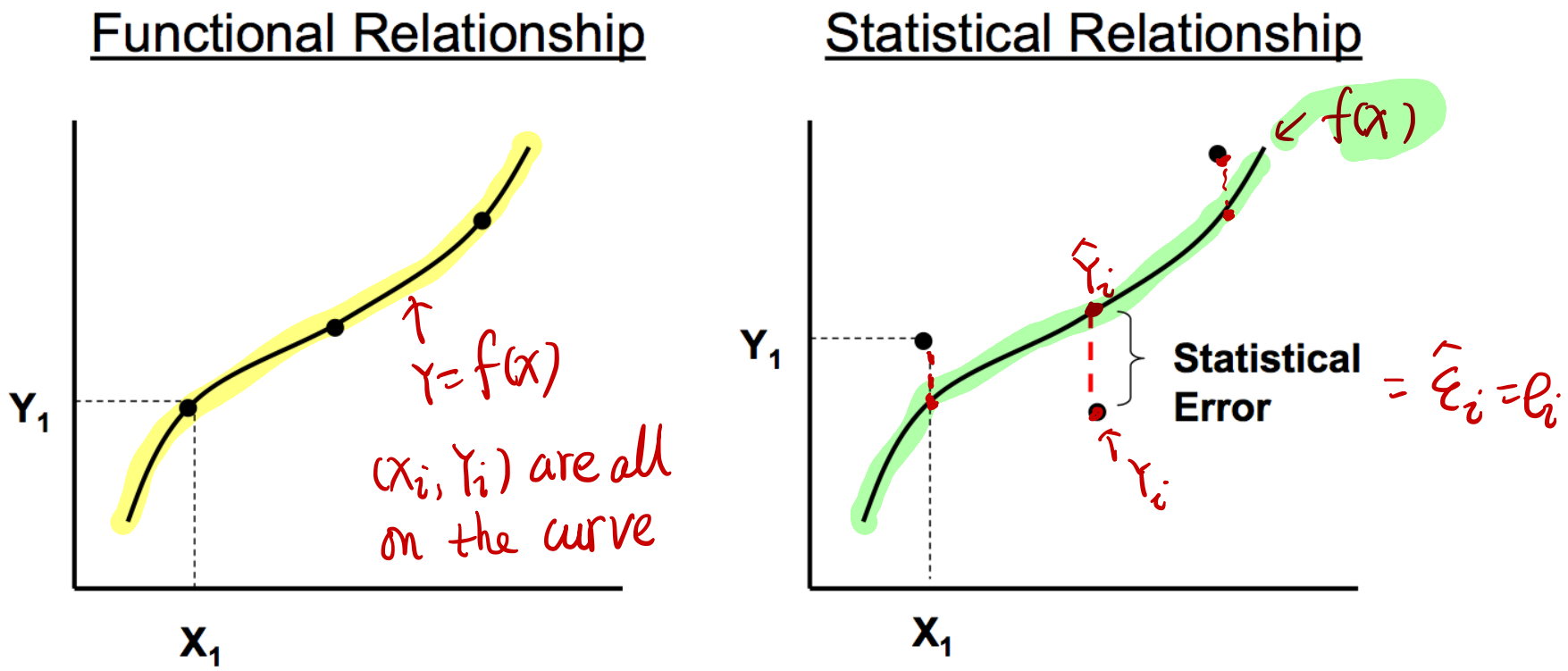
What a data looks like?

i	X	Y
1	0	6.95
2	1	5.22
3	2	6.46
4	3	7.03
5	4	9.71
6	5	9.67
7	6	10.69
8	7	13.85
9	8	13.21
9	9	14.82

For $i = 3, (X_3, Y_3) = (2, 6.46)$. For a real data, usually you don't have the index i column as given in the table.

Types of relationships

- Scatter plots of data pair (Y_i, X_i)



- For each of these **functional relationships**, the equation, $Y = f(X)$, exactly describes the relationship between the two variables. We are not interested in the functional relationship in this course.
- Instead, we are interested in **statistical relationships**, in which the relationships between the variables is not prefect.

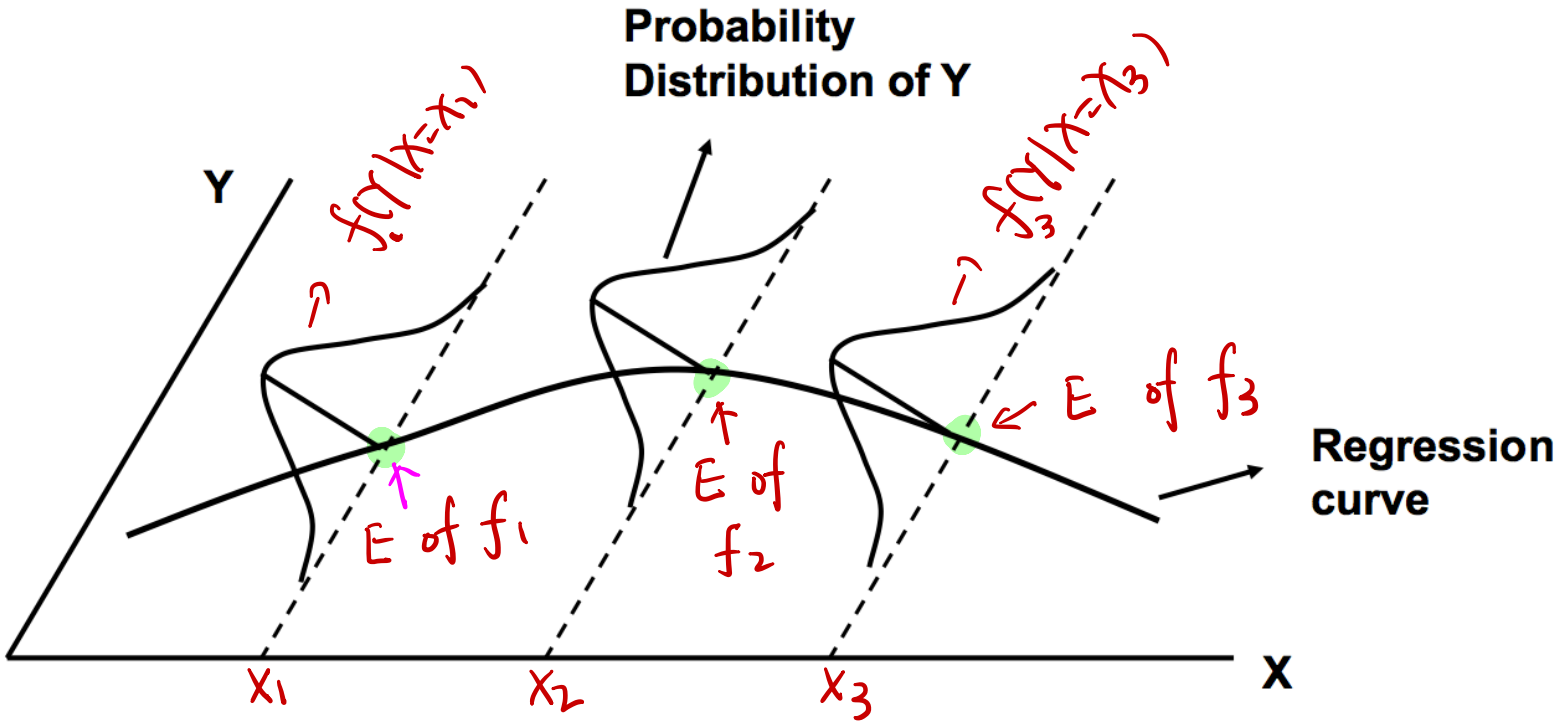
Regression Models

- Regression model describes the statistical relationship between the response variabel Y and one or more predictor variable(s)
  - The response variable Y has a tendency to vary with the predictor variable X in a systematic fashion.
  - The data are scattered around the regression curve.
- Regression model assumes a distribution for Y at each level of X.
- When the relationship between Y and X is linear, we call it linear regression.
  - In linear regression model, if it concerns study of only one predictor, then we have simple linear regression (SLR) model.
  - In contrast, we have multiple linear regression (MLR).

SLR: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

MLR (with 3 X_s): $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$

Regression model (non-linear)

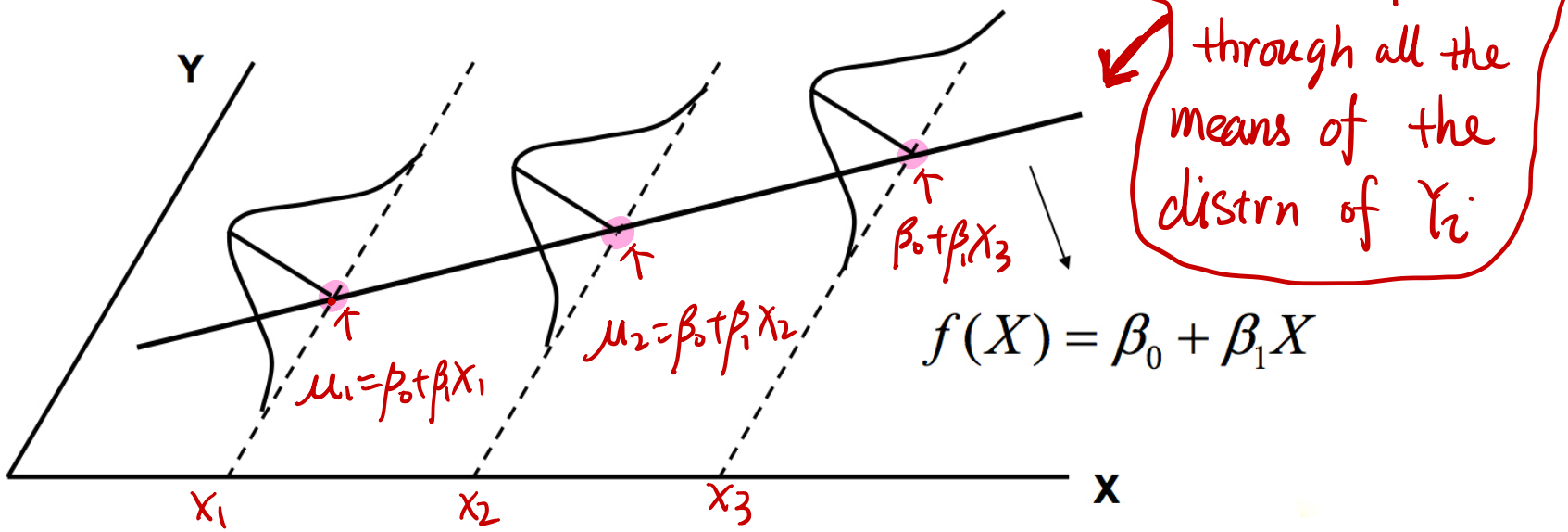


1. There is a probability distribution of Y for each level of X.
2. The means of these distributions of Y at different levels of X follow the regression curve.

$E(Y|X=X_i) = \mu_i = E \text{ of } f_i$

Simple linear Regression

- It concerns about the statistical relationship between Y and one X.
- The regression curve is a straight line.



The relationship is termed as linear if it is linear in parameters (β_0, β_1) and nonlinear, if it is not linear in parameters.

eg. { Linear: $E(Y|X) = \beta_0 + \beta_1 X$
nonlinear: $E(Y|X) = \beta_0 + e^{\beta_1} X$

Simple Linear Regression (SLR)

- Formal model form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

  - Y_i is the value of response variable in the i^{th} trial (random but observable).
  - X_i is the predictor in the i^{th} trial(a known constant).
  - β_0 is the intercept of the regression line (model parameter: assume constant but unknown).
  - β_1 is the slope of the regression line (model parameter: assume constant but unknown).
  - ϵ_i is the error term (random and unobservable)

- In summary

R/C	Known	Unknown
Random	Y	ϵ
Constant	X	$\beta_0, \beta_1, \sigma^2$

SLR example 1: hourly wage (Y) and education years (X)

Variables

- Y: hourly wage(pound)
- X: years of education

Intuitively, we expect that Y increases as X increases.

Parameter interpretation

- β_0: Y-intercept, it gives the starting salary
- β_1: slope, it gives hourly wage raise

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

$E(Y|X) = \beta_0 + \beta_1 X$

For $X=0$, on avg, $Y = \beta_0$

Increase X by 1 unit, on avg, Y increases by β_1

SLR example 1: hourly wage (Y) and education years (X)

$E(Y)=\beta_0+\beta_1 X$ ↙ $Y=\beta_0+\beta_1 X+\varepsilon$ ↙

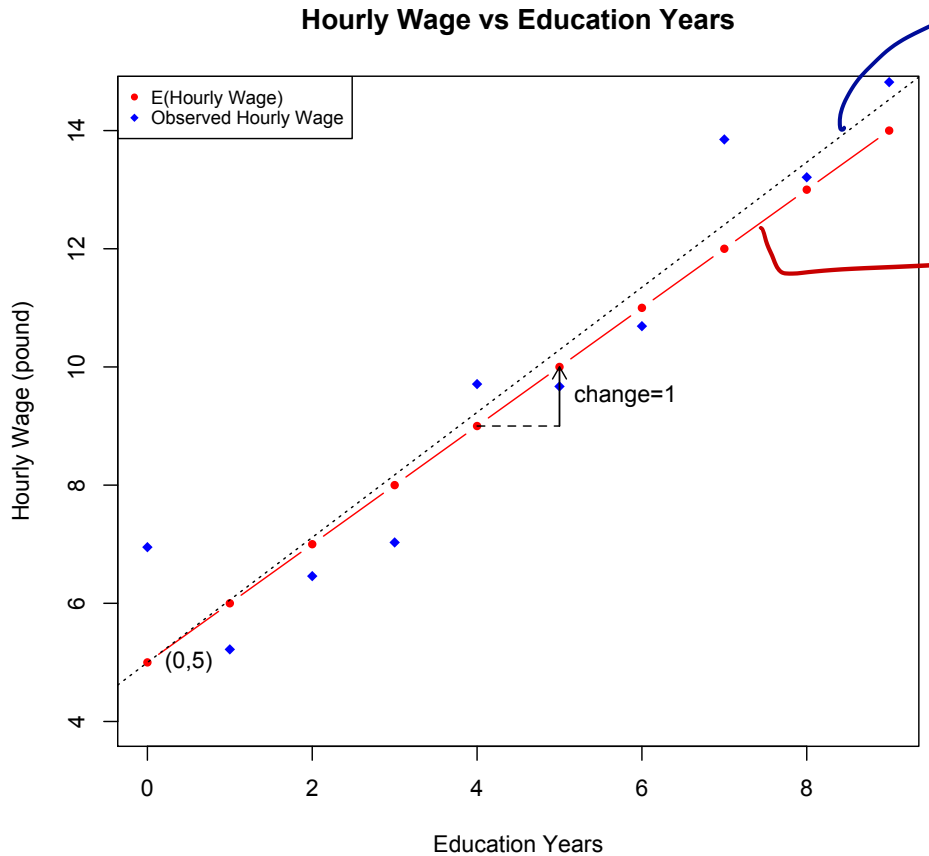
EducYrs	E(Y)=E($HWage_T$)	Y=$HWage_O$
0	5	6.95
1	6	5.22
2	7	6.46
3	8	7.03
4	9	9.71
5	10	9.67
6	11	10.69
7	12	13.85
8	13	13.21
9	14	14.82

↑ hypothetical column since we don't know β_0, β_1

→ observed Y, not perfect, might have measurement error (ε_i)

- EducYrs (X): years of education;
- $HWage_T$ (true E(Y)): the true expected hourly wage (pound).
- $HWage_O$ (observed Y): the observed hourly wage (pound)

SLR example 1: hourly wage (Y) and education years (X)



$\hat{Y}_i = b_0 + b_1 X_i$
(Estimated line)

$E(Y_i) = \beta_0 + \beta_1 X_i$
(True line)
(population line)

The observed Y goes up and down around the population regression line. In real world, we don't observed the true error term (ϵ), instead we have data (EducYrs, HWage_O). We aim to reveal the true relationship between Y and X using the data we observed. That is, how to use observed data to estimate β_0, β_1?

b_0 b_1

True vs Estimated model

Assume we have a data set of size n : $(Y_i, X_i), i = 1, \ldots, n$.

(*) True regression model (or population regression model)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = f(X) + \epsilon_i, \quad f(X) = \beta_0 + \beta_1 X_i$$

(*) Estimated regression model (or sample regression model)

$$\hat{Y}_i = b_0 + b_1 X_i = \hat{f}(X), \quad \hat{f}(X) = b_0 + b_1 X_i$$

- Point estimators of β_0, β_1 are denoted by b_0, b_1 respectively.
- The estimate of Y_i (for given X_i) is denoted by $\hat{Y}_i$.
- The estimate of ϵ_i (for given X_i) is denoted by e_i

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$$

View it as prediction error in data mining/machine learning.

This implies that

$$Y_i = \hat{Y}_i + e_i = (b_0 + b_1 X_i) + e_i$$

$Y_i = (Y_i - \hat{Y}_i) + \hat{Y}_i = e_i + \hat{Y}_i = e_i + (b_0 + b_1 X_i)$

(*) $\epsilon_i = Y_i - E(Y_i)$

$e_i = Y_i - \hat{E}(Y_i) = Y_i - \hat{Y}_i$

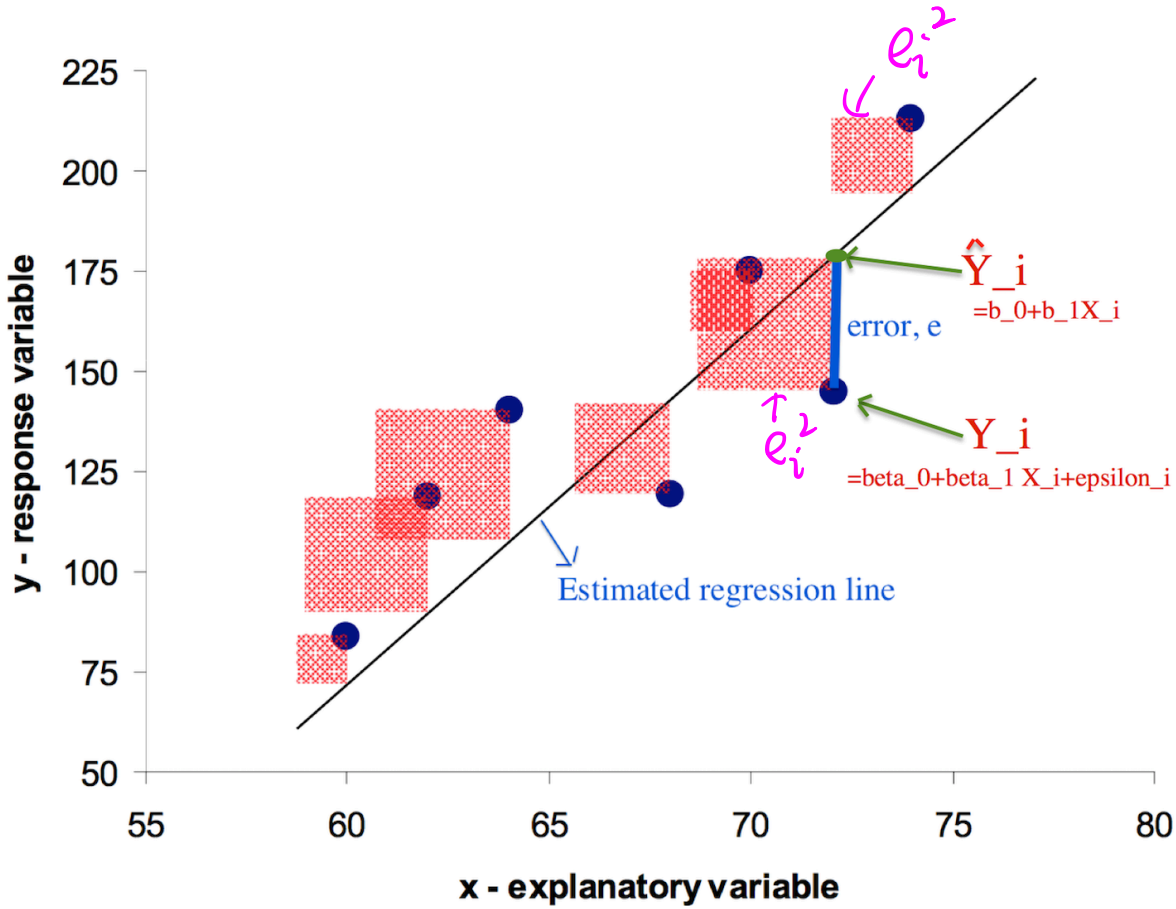
*Note that e_i and ϵ_i are different!!!

True vs Estimated model

- Difference between $\hat{Y}_i = b_0 + b_1 X_i$ and $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$.
- Note that we never observe ϵ_i, but we could estimate it by e_i.

$$Y_i = \hat{Y}_i + e_i = \hat{f}(X) + \text{estimated error}_i,$$

where $e_i = Y_i - \hat{Y}_i$. = residual



$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

↑ Estimate β_0, β_1 by LS method

Estimation by Least Squares method

Gauss-Markov Assumptions

- **Gauss-Markov Assumptions**:
  1. Dependent variable (DV) is linear in parameter and can be written as : $Y = \beta_0 + \beta_1 X + \epsilon$
  2. $E(\epsilon_i) = 0$. ϵ_i is R.V. with mean 0.
  3. $V(\epsilon_i) = \sigma^2$, this homoskedasticity implies that the model uncertainty is identical across observations.
  4. $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. ϵ_i and ϵ_j are uncorrelated:
- X is assumed to be constant, ie, X is uncorrelated with the error term ($Cov(X_i, \epsilon_i) = 0$).
- $cov(\epsilon_i, \epsilon_j)$=0 does not guarantee ϵ_i and ϵ_j are independent. But if they are independent, their covariance must be 0.

$Cov(\epsilon_i, \epsilon_j)=0 \not\Rightarrow \epsilon_i \perp \epsilon_j$

- **Above assumptions imply**:
  - $E(Y_i|X_i) = \mu_i = \beta_0 + \beta_1 X_i$, that is $f(X) = \beta_0 + \beta_1 X$
  - $V(Y_i|X_i) = V(\mu_i + \epsilon_i) = V(\epsilon_i) = \sigma^2$
  - $Cov(Y_i, Y_j|X_i) = E\{(Y_i - \mu_i)(Y_j - \mu_j)\} = E(\epsilon_i \epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$

We often drop $|X$ notation in above because X is non-random.

Gauss-Markov Assumptions (be contd)

- Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ GM-②

- $E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i = \mu_i$

  ($\beta_0 + \beta_1 X_i$: constant, not RVs)

- $V(Y_i) = V(\beta_0 + \beta_1 X_i + \varepsilon_i) = V(\mu_i + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$

- $cov(Y_i, Y_j) = E\{(Y_i - \mu_i)(Y_j - \mu_j)\}$

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

$\Rightarrow Y_i - \beta_0 - \beta_1 X_i = \varepsilon_i$

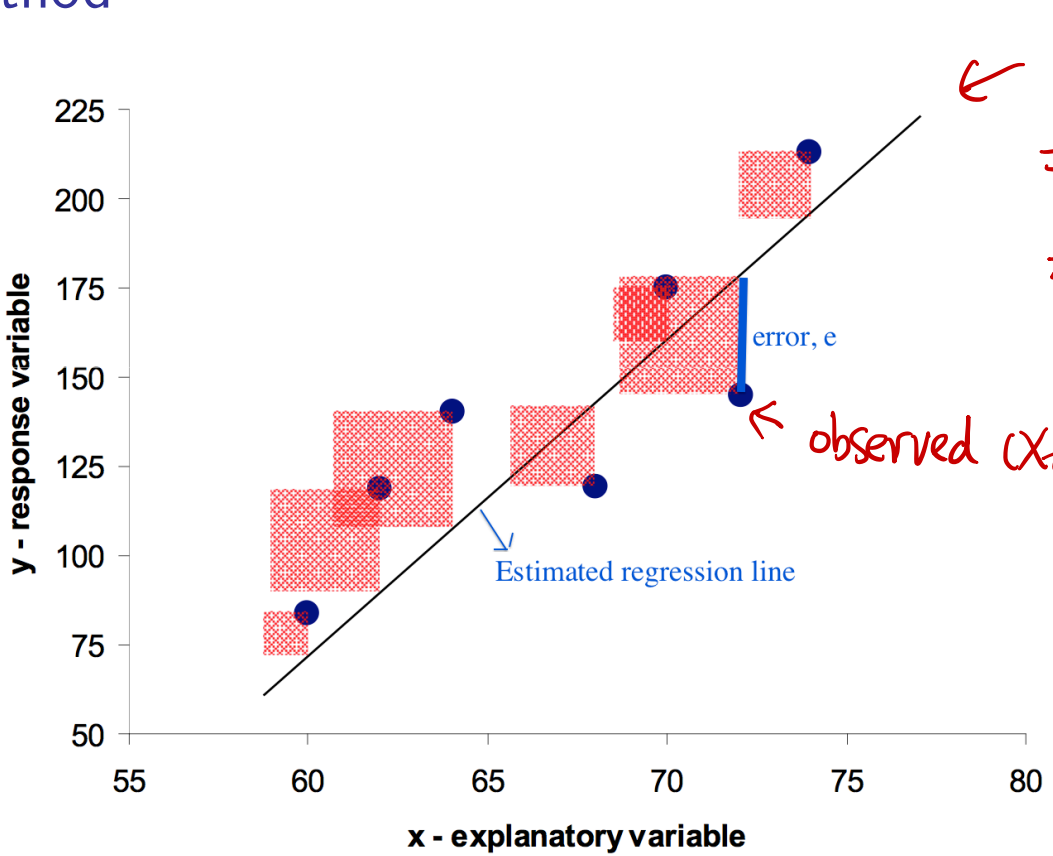
$= E\{(Y_i - \beta_0 - \beta_1 X_i)(Y_j - \beta_0 - \beta_1 X_j)\}$

$= E(\varepsilon_i \varepsilon_j)$

$= E(\varepsilon_i \varepsilon_j) - E(\varepsilon_i) E(\varepsilon_j)$ ($E(\varepsilon_i) = 0$, $E(\varepsilon_j) = 0$)

$= cov(\varepsilon_i, \varepsilon_j) = 0$ by GM-④

Least Square Method



← fitted reg. line
= estimated reg. line
$\Rightarrow \hat{Y}_i = b_0 + b_1 X_i$

← observed (X_i, Y_i)

- The equation of the estimated model (or best fitting line) is: $\hat{Y}_i = b_0 + b_1 X_i$
- We need to find the values b_0, b_1 that make the sum of the squared prediction error the smallest it can be. That is, find b_0 and b_1 that miniminze the objective function Q.

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Least Square Estimates b_0, b_1

$$Q = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$$

Minimizing Q gives

$\hat{Y}_i = b_0 + b_1 X_i$
$= \bar{Y} + b_1(X_i - \bar{X})$

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1\bar{x} \tag{2}$$

$$b_1 = \hat{\beta}_1 = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{3}$$

where

$$\bar{X} = \frac{1}{n}\sum_1^n X_i, \quad \bar{Y} = \frac{1}{n}\sum_1^n Y_i, \quad S_{xy} = \sum_1^n (x_i - \bar{x})(y_i - \bar{y}), S_{xx} = \sum_1^n (x_i - \bar{x})^2$$

Substituting b_0 in the estimated model, it can be rewritten as

$$\hat{Y}_i = b_0 + b_1 X_i = \bar{Y} + b_1(X_i - \bar{X}),$$

$= \bar{Y} - b_1\bar{X}$

this also implies

$$Y_i = \bar{Y} + b_1(X_i - \bar{X}) + e_i \leftarrow Y_i = (Y_i - \hat{Y}_i) + \hat{Y}_i = e_i + \hat{Y}_i$$

i.e. The estimated regression line always goes through the point data point $(\bar{X}, \bar{Y})$.

Proof

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

$$\frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0 \tag{4}$$

$$\frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) X_i = 0 \tag{5}$$

These lead to the **Normal equations:**

(4) →

$$\sum_{i=1}^{n} Y_i = n b_o + b_1 \sum_{i=1}^{n} X_i$$

(5) →

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

$$\Rightarrow \quad b_1 = \frac{S_{xy}}{S_{xx}}, \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

The normal equations can be solved simultaneously for b_0 and b_1 given in equations (2) and (3) respectively.

proof

The Hessian matrix which is the matrix of second order partial derivatives in this case is given as

$$H = \begin{pmatrix} \frac{\partial Q}{\partial \beta_0^2} & \frac{\partial Q}{\partial \beta_0 \beta_1} \\ \frac{\partial Q}{\partial \beta_0 \beta_1} & \frac{\partial Q}{\partial \beta_1^2} \end{pmatrix} = 2 \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}$$

- The 2 by 2 matrix H is positive definite if its determinant and the element in the first row and column of H are positive.
- The determinant of H is given by $|H| = 4n \sum (x_i - \bar{x})^2 > 0$ given $x \neq c$(some constant).
- So H is positive definite for any (β_0, β_1), therefore Q has a global minimum at (b_0, b_1).

For students with firm maths.

Review on Positive Definite matrix

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

- In general, a symmetric matrix is Postive Definite (P.D.) iff all its eigenvalues are positive.

For a 2 by 2 symmetric matrix,

- Since $det(A) = \lambda_1 \lambda_2$, it is necessary that the determinant of A be positive. On the other hand, if $det|A| > 0$, then either both eigenvalues are positive or negative.
- $tr(A) = \lambda_1 + \lambda_2$, if $det(A) > 0$ and $tr(A) > 0$ then both eigenvalues must be positive.
- However, $det(A) = ac - b^2 > 0$, then a and c must have the same sign. Thus $det(A) > 0, tr(A) = a + c > 0$ is equivalent to the condition that $det(A) > 0$ and $a > 0$.

For students with firm maths.

Equivalent formula for b_1

(A) $S_{xx} = \sum_1^n (X_i - \bar{X})$ = a constant

(B) $\sum_1^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - \sum_{i=1}^n \bar{X}$
$= n\bar{X} - n\bar{X} = 0$
$\uparrow$ $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$

- (6)→(7):

$\because \sum_{i=1}^n (X_i - \bar{X})\bar{Y}$
$= \bar{Y} \sum_{i=1}^n (X_i - \bar{X})$
$= \bar{Y} \cdot 0 = 0$

$$b_1 = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_1^n (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}} \tag{6}$$

$$= \frac{\sum_1^n (X_i - \bar{X}) Y_i}{S_{xx}} \tag{7}$$

- (7)→(8):

$\frac{\sum_{i=1}^n (X_i - \bar{X})\bar{Y}}{S_{xx}} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_{xx}}\right) Y_i$

$$= \sum_{i=1}^n \frac{X_i - \bar{X}}{S_{xx}} y_i = \sum_{i=1}^n k_i Y_i \tag{8}$$

- (6)→(9)

$\because \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$
$= \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}$ (A1)

$$= \frac{\sum_1^n X_i Y_i - n\bar{X}\bar{Y}}{S_{xx}} \tag{9}$$

where (8) suggests that b_1 is a linear combination of Y_i (assume constant X) and hence is a linear estimator.

$$k_i = \frac{X_i - \bar{X}}{S_{xx}} = \frac{X_i - \bar{X}}{\sum_1^n (X_i - \bar{X})^2}$$

→ $\sum k_i = 0$ since (B)

Proof (7,8,9): .....

Equivalent formula for b_0

Show b_0 is a linear comb. of Y_i

$$b_0 = \bar{Y} - b_1\bar{X} = \sum_1^n \frac{1}{n} Y_i - \bar{X}\sum_1^n k_i Y_i$$

$$= \sum_{i=1}^n (\frac{1}{n} - k_i\bar{X}) Y_i$$

$$= \sum_1^n w_i Y_i, \quad w_i = \frac{1}{n} - k_i\bar{X}$$

which suggests that b_0 is also a linear combination of Y_i and hence is a linear estimator.

Exercise (in below,first show (1) and use it to prove (3) and (4))

(*)

1. $\sum_{i=1}^n k_i = 0, \sum_{i=1}^n k_i X_i = 1$
2. $\sum_{i=1}^n k_i^2 = \frac{1}{S_{xx}}$
3. $\sum_{i=1}^n w_i = 1$
4. $\sum_{i=1}^n w_i X_i = 0$

LS estimators are BLUE

Gauss-Markov Theorem

Under the Gauss-Markov assumptions, the Ordinary Least Square (OLS) estimators, $\hat{\beta}_0, \hat{\beta}_1$ are the Best Linear Unbiased Estimator (BLUE), that is

1. Unbiased: $E(b_0) = \beta_0$, and $E(b_1) = \beta_1$
2. Linear: $b_1 = \sum_{i=1}^{n} k_i Y_i$, $b_0 = \sum_{i=1}^{n} w_i Y_i$. ✓
3. Best: b_0, b_1 have the smallest variance among the class of all linear unbiased estimators.

- prove it using linear algebra (*).
- prove it using calculus.

Q: T or F: OLS estimators b_0, b_1 have the smallest variance among all the unbiased estimators.

A: False. b_0, b_1 have smallest variance among LUE.

Show the unbiasedness of b_0, b_1

A1 : Q1-b: show $\sum_i^n (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$

Note that $b_1 = \sum k_i Y_i$ and we have

$$\sum_1^n k_i = \sum_1^n \frac{X_i - \bar{X}}{S_{xx}} = \frac{1}{S_{xx}} \sum_1^n (X_i - \bar{X}) = 0$$

$$S_{xx} = \sum_1^n (X_i - \bar{X})^2 = \sum_1^n X_i^2 - n\bar{X}^2$$

From previous slide, we have

$$E(b_1) = E(\sum_1^n k_i Y_i) = \sum_1^n k_i E(\beta_0 + \beta_1 X_i + \epsilon_i)$$

$$= \sum_1^n k_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_1^n k_i + \beta_1 \sum_1^n X_i k_i$$

$$= 0 + \beta_1 \frac{\sum_1^n X_i^2 - n\bar{X}^2}{S_{xx}} = \beta_1$$

$$E(b_0) = E(\bar{Y} - b_1\bar{X}) = (\beta_0 + \beta_1\bar{X}) - \beta_1\bar{X} = \beta_0$$

$E(\bar{Y})$ $E(b_1)$

$X_i k_i = X_i \frac{(X_i - \bar{X})}{S_{xx}} = \frac{X_i^2 - X_i\bar{X}}{S_{xx}}$

$\sum k_i X_i$

A1 Q1-b $= \frac{1}{S_{xx}} \sum_{i=1}^n (X_i^2 - X_i\bar{X}) = \frac{1}{S_{xx}} (\sum X_i^2 - n\bar{X}^2) = 1$

Proof that b_0 is the best

A detailed proof is posted

- $b_0 = \sum_1^n w_i Y_i$, so $V(b_0) = \sum_1^n w_i^2 \sigma^2$.
- Let $\tilde{b}_0$ be another unbiased linear estimator of β_0, $\tilde{b}_0 = \sum_1^n a_i y_i$. If $a_i = w_i$, then $\tilde{b}_0 = b_0$.

$$\tilde{b}_0 = \sum_1^n a_i y_i = \sum_1^n a_i(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 \sum_1^n a_i + \beta_1 \sum_1^n a_i X_i + \sum_1^N a_i \epsilon_i$$

Unbiasedness of $\tilde{b}_0$, $E(\tilde{b}_0) = \beta_0$, occurs iff $\sum a_i = 1, \sum a_i x_i = 0$

- $V(\tilde{b}_0) = E(\tilde{b}_0 - E(\tilde{b}_0))^2 = E((\sum a_i \epsilon_i)^2)$, this gives

$$V(\tilde{b}_0) = \sum a_i^2 E\epsilon_i^2 + \sum_{i \neq j} 2 a_i a_j E(\epsilon_i \epsilon_j) = \sigma^2 \sum a_i^2$$

- Since a_i is arbitrary, we let $a_i = w_i + d_i$, want to show $\sum a_i^2 \geq \sum w_i^2$.
- Can show $\sum w_i = 1, \sum w_i X_i = 0$, so $\sum a_i = 1 = \sum w_i + \sum d_i$ and $\sum a_i X_i = 0 = \sum w_i X_i + \sum d_i X_i$ implies $\sum d_i = 0$ and $\sum d_i X_i = 0$
- Using $\sum d_i = 0$ and $\sum d_i X_i = 0$, we can show $\sum w_i d_i = \sum (\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{xx}}) d_i = 0$
- $\sum a_i^2 = \sum w_i^2 + 2 \sum w_i d_i + \sum d_i^2 \geq \sum w_i^2$

Proof that b_1 is the best (PP43-44)

Detailed proof is posted.

- $b_1 = \sum k_i Y_i$, $k_i = \frac{X_i - \bar{X}}{S_{xx}}$ and we also have $\sum k_i = 0$, $\sum k_i x_i = 1$
- let $\tilde{b}_1$ be another linear unbiased estimator of β_1, $\tilde{b}_1 = \sum c_i Y_i$.
- $E(\tilde{b}_1) = E(\sum c_i(\beta_0 + \beta_1 X_i + \epsilon_i)) = \beta_1$ is equivalent to $\sum c_i = 0, \sum c_i X_i = 1$.
- since c_i is arbitrary, we let $c_i = k_i + d_i$,

$$\sum c_i = 0 \rightarrow \sum k_i + \sum d_i = \sum d_i = 0,$$

$$\sum c_i x_i = 1 = \sum k_i x_i + \sum d_i x_i \rightarrow \sum d_i x_i = 0$$

- $V(b_1) = V(\sum k_i Y_i) = \sigma^2 \sum k_i^2$
- Now show that $V(\tilde{b}_1) \geq V(b_1)$

$$V(\tilde{b}_1) = \sigma^2 \sum c_i^2 = \sigma^2 \sum (k_i + d_i)^2 = \sigma^2 (\sum k_i^2 + 2 \sum k_i d_i + \sum d_i^2)$$

since $\sum k_i d_i = 0$, so we have

$$V(\tilde{b}_1) = \sigma^2 (\sum k_i^2 + \sum d_i^2) \geq V(b_1)$$

Estimation of error terms variance σ^2

- Error sum of squares (SSE) or residual sum of square (RSS)

$$SSE = \sum_{1}^{n} e_i^2 = \sum_{1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- SSE has n-2 degrees of freedom associated with it. Two degrees of freedom are lost because both β_0 and β_1 had to be estimated in obtaining estimated means $\hat{Y}_i$
- In LS method, the error term variance $\sigma^2 = V(\epsilon_i)$ for all i, is estimated by the error mean square (MSE)

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_1^n e_i^2}{n-2} = \frac{\sum_1^n (Y_i - \hat{Y}_i)^2}{n-2}$$

Show $E(MSE) = \sigma^2$

This is equivalent to show

$$E(SSE) = E\{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\} = (n-2)\sigma^2$$

$$E\{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\} = \sum_{i=1}^{n} E(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} Var(Y_i - \hat{Y}_i) + \{E(Y_i - \hat{Y}_i)\}^2$$

$$= \sum_{i=1}^{n} Var\{(Y_i - \bar{Y} - b_1(X_i - \bar{X}))^2\}$$

$$= \sum_{i=1}^{n}\{Var(Y_i - \bar{Y}) - 2Cov(Y_i - \bar{Y}, b_1(X_i - \bar{X})) + Var(b_1)(X_i - \bar{X})^2\}$$

$$= \sum_{i=1}^{n}\{Var(Y_i - \bar{Y}) - 2Cov((Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)(X_i - \bar{X})^2\}$$

$$= \sum_{i=1}^{n}\{Var(\varepsilon_i - \bar{\varepsilon}) - 2Cov((Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)(X_i - \bar{X})^2\}$$

$$= (n-1)\sigma^2 - 2Cov(\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= (n-1)\sigma^2 - 2Cov(b_1\sum_{i=1}^{n}(X_i - \bar{X})^2, b_1) + Var(b_1)\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= (n-1)\sigma^2 - Var(b_1)\sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-2)\sigma^2.$$

Detailed proof is posted.

$V(X) = E(X^2) - (EX)^2$

$\hookrightarrow E(X^2) = V(X) + (EX)^2$

$E(Y_i) = \beta_0 + \beta_1 X_i$

$E(\hat{Y}_i) = E(b_0) + E(b_1)X_i = \beta_0 + \beta_1 X_i$

- $Cov(aY_1, Y_2) = Cov(Y_1, aY_2)$ ↑ show it by defn.

$V(b_0)$, $V(b_1)$ and their estimates

$$w_i = \frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{xx}}$$

$$k_i = \frac{X_i - \bar{X}}{S_{xx}}$$

$$V(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)$$

$$V(b_1) = \frac{\sigma^2}{S_{xx}}$$

thus

$$V(b_0) = V(\sum w_i Y_i) = \sum w_i^2 \sigma^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)$$

$$V(b_1) = V(\sum k_i Y_i) = \sum k_i^2 \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

Estimators of V(b_0) and V(b_1) are obtained by replacing σ^2 by its point estimator MSE

$$s^2(b_0) = MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right) = \hat{var}(b_0)$$

$$s^2(b_1) = \frac{MSE}{S_{xx}} = \hat{var}(b_1)$$

Example 2: SLR Estimation (by hand)

- Annual salary (Y) and years of service (X)

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
i=1	3	34	-5	-4	25	16	20
i=2	6	34	-2	-4	4	16	8
i=3	10	38	2	0	4	0	0
i=4	8	37	9	-1	0	1	0
i=5	13	47	5	9	25	81	45
Sum	40	190	0	0	58	114	73

Above calculation gives

- $\bar{X} = 40/5 = 8$
- $\bar{Y} = 190/5 = 38$.

$$b_1 = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_1^n (X_i - \bar{X})^2} = \frac{73}{58} = 1.258621$$

$$b_0 = \bar{Y} - b_1 \bar{X} = 38 - 1.258621 \times 8 = 27.931$$

Example 2: SLR Estimation (by hand)

- Find $\hat{Y}_i = 27.931 + 1.25862 X_i$ ← estimated / fitted reg. line
  - $\hat{Y}_i = c(31.70686, 35.48272, 40.51720, 37.99996, 44.29306)$ ← Predicted values
- Find $e_i = Y_i - \hat{Y}_i$
  - $e_i = c(2.29314, -1.48272, -2.51720, -0.99996, 2.70694)$ ← residuals
- Estimate σ^2 by MSE: $s^2 = \hat{\sigma}^2 = \sum e_i^2/(n-2) = 7.373563$
  - $\hat{\sigma} = \sqrt{7.373563} = 2.715431$

$$MSE = \frac{\Sigma e_i^2}{n-2} = \frac{SSE}{n-2}$$

Topics for next lecture

- Properties of fitted regression line
- Parameter estimation by MLE method
- Inferece of SLR
- ...

Practice problems after Week 01- Lecture I

Highly recommend you do #3 and #4 to develop skills you need for upcoming assignment, test and exam.

1. Reading chapter sections in textbook: 1.1,1.3,1.6.
2. Try exercise in textbook
   - 1.3, 1.5, 1.6, 1.7, 1.8, 1.11, 1.16, 1.18, 1.20(*), 1.21(*), 1.24(*), 1.29, 1.30, 1.33, 1.36, 1.39a, 1.40, 1.41a.
   - For questions marked (*), the SAS code & output is posted with the solutions.
   - You only need to interpret that output.
3. Follow the instruction in Week 00 slides, install R, R Studio and necessary R package.
   - Make sure you could produce test.PDF from test.Rmd successfully.
   - Make sure you know how to split the a whole PDF into sub PDF files.
4. Picking up R on data camp. Make sure you will complete all 3 courses I assign before June 23, 2017. Weight for these three course is 3% of this course.
5. Copy and paste the R code in R provided in the next 3 slides for Example 2. You should have the same output.
6. Try the exercises on slide 35.

Example 2: SLR Estimation (using R)

R code to find b_0, b_1

```
X=c(3,6,10,8,13)     # assign predictor observations to object X
Y=c(34,34,38,37,47) # assign response observations to object Y
lmfit = lm(Y~X)      # fitting data with a simple linear regression
lmfit$coef           # print the b0 and b1 estimates
```

```
## (Intercept)           X
##   27.931034    1.258621
```

Example 2: SLR Estimation (using R)

- Find $\hat{Y}_i = b_0 + b_1 X_i$
- Find $e_i = Y_i - \hat{Y}_i$
- Estimate σ^2 by MSE $s^2 = \hat{\sigma}^2 = \sum e_i^2/(n-2)$

R code:

```
b0=lmfit$coef[1]     # assign estimated intercept value to b0
b1=lmfit$coef[2]     # assign estimated slope value to b1
Yhat=b0+b1*X         # find fitted response value : Y_i= b0+b1*X_i
Yhat                 # have a look of the fitted value
```

```
## [1] 31.70690 35.48276 40.51724 38.00000 44.29310
```

```
e=Y-Yhat             # find error e_i= Y_i-fitted Y_i
e                    # have a look of the error observations
```

```
## [1]  2.293103 -1.482759 -2.517241 -1.000000  2.706897
```

```
mse=sum(e^2 )/(5-2) # find MSE=SSE/(n-2)
sqrt(mse)
```

```
## [1] 2.715431
```

Example 2: SLR Estimation

R code:

```r
summary(lmfit)    # summary information from the fitted SLR
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      1      2      3      4      5
##  2.293 -1.483 -2.517 -1.000  2.707
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.9310     3.1002    9.01  0.00289 **
## X             1.2586     0.3566    3.53  0.03864 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.715 on 3 degrees of freedom
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.7413
## F-statistic: 12.46 on 1 and 3 DF,  p-value: 0.03864
```