STA302/1001 - Methods of Data Analysis I

(Week 01 lecture notes)

Wei (Becky) Lin

Sept. 12, 2016

UNIVERSITY OF TORONTO

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**
  - Sections L0101/L2001
    - Tuesday 10:10-12:00 in **MS2158** (no lecture on Tuesday, November 8 (Fall Break))
    - Thursday 10:10-11:00 in **OI-G162**. (class on Thursday, October 27 will take place in ES1050)
  - Section L5101: Thursday 17:10-20:00 in **PB-B150**.
- **Office Hours**
  - Me: Tue. 2-3pm, Thu. 12:30-1:30pm in SS6026 or SS6007(starts from 2nd week)
  - TAs: TBA
- **Textbook(s)**
  - *Applied Linear Regression Models*, 4th 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
- Portal contains a **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 it often.

Notes about syllabus

- **Computing**: R and R studio software are used for assignments and you need to be able to interpret R output for 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 install R and Rstudio on your computer after this class.
  - 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	25%	Oct. 18 (L0101/L2001), Oct. 20 (L5101)	10:00-12:00 (L0101/L2001), 18:00-20:00 (L5101)	TBA
Make-up Midterm		TBA	TBA	
Assignment 1	10%	Thursday, Oct. 13rd	L0101/2001: due 10:10 L5101: due 17:10	OI-G162 PB-B150
Assignment 2	10%	Thursday, Nov. 17th	L0101/2001: due 10:10 L5101: due 17:10	OI-G162 PB-B150
Assignment 3	10%	Thursday, Dec. 1st	L0101/2001: due 10:10 L5101: due 17:10	OI-G162 PB-B150
Final Exam	45%	Posted by A&C on Oct. 21st	TBA	TBA

Important dates

- **Midterm (25%):** Tue. Oct. 18 (L0101/2001), Thu. Oct. 20(L5101). Make-up midterm date: TBA.
- **Assignments (30%)**
  - A1: due Oct. 13.
  - A2: due Nov. 17.
  - A3: due Dec. 01.
- **Final exame (45%)**: timetable for F section code courses is available and posted by Art&Sci.

Do and Do Not

{Do}

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

{Do Not}

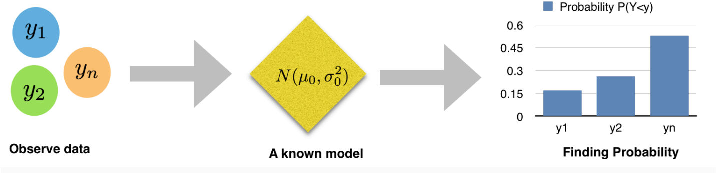
- Don't copy, and don't let anyone copy from you.
- It is academic dishonesty to present someone else's work as your own, or to allow your work to be copied for this purpose.
- The person who allows her/his work to be copied is equally guilty, and subject to disciplinary action by the university.
- ....

Course Objective

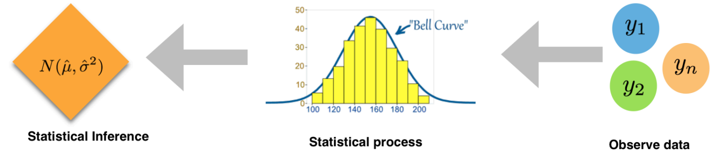
- Course covers a large part of the theory and gain practical skills of developing linear regression models for inference, prediction and interpreting the results.
  - Least squares / MLE estimation.
  - Inference for regression parameters.
  - Model diagnostics and remedial procedure.
  - Multiple linear 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)

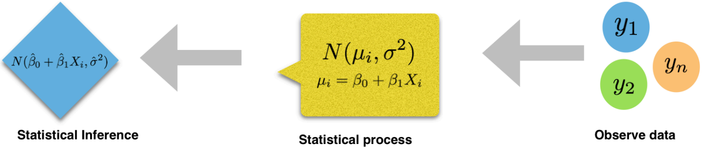


- Introduction to statistical inference (eg, STA261: know how to estimate model parameter θ, CI,hypothesis testing, etc)



Connection to pre-requisite course

- 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

Week 01- Learning objectives & Outcomes

- 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.

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 sutdy 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.

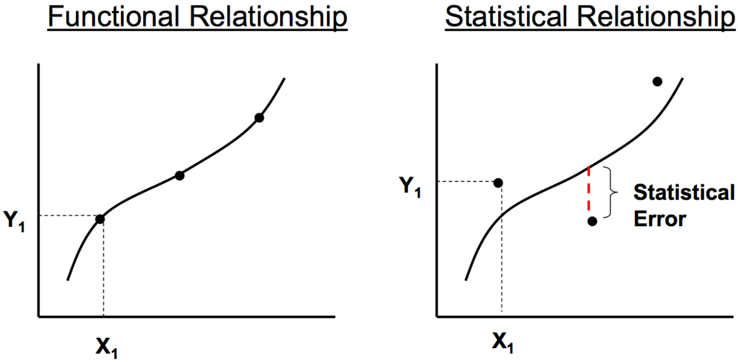
What a data looks like?

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)

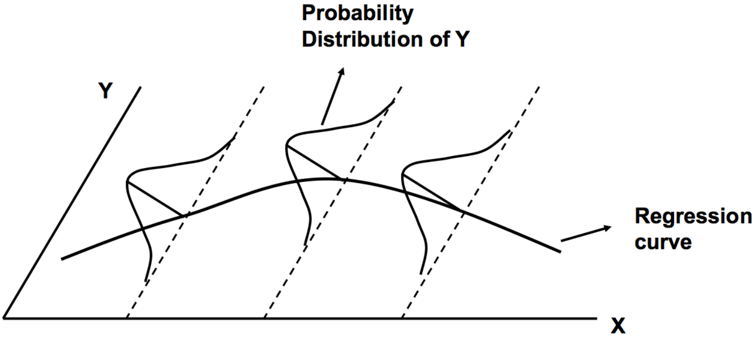


- For each of these functional relationships, the equation, $Y = f(X)$, exactly describes the relationship between the two variables. We are not interested in the functional relationship in this course.
- Instead, we are interested in **statistical relationships**, in which the relationships between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

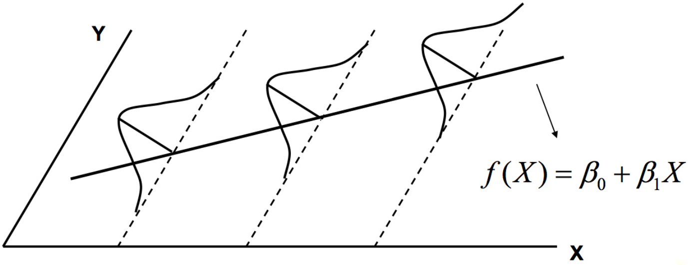
Regression model (non-linear)



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.

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.

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

Parameter interpretation

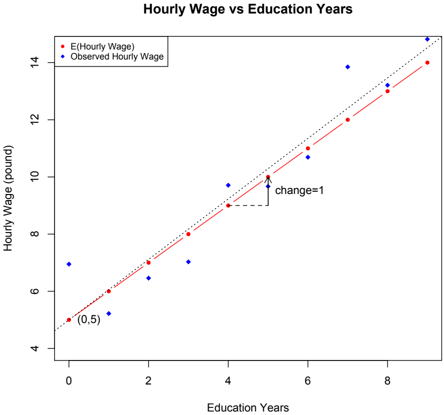
- β_0: Y-intercept, it gives the starting salary
- β_1: slope, it gives hourly wage raise

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

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

- 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)



The observed Y goes up and down around the true Y. In real world, we don't observed the true Y, 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?

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

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

This implies that

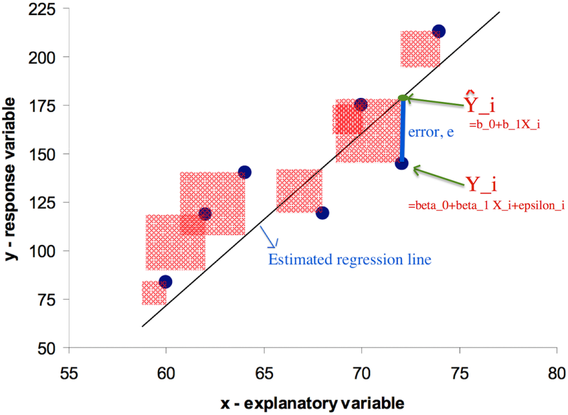
$$Y_i = \hat{Y}_i + e_i = (b_0 + b_1 X_i) + e_i$$

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 observed ϵ_i, but $

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

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



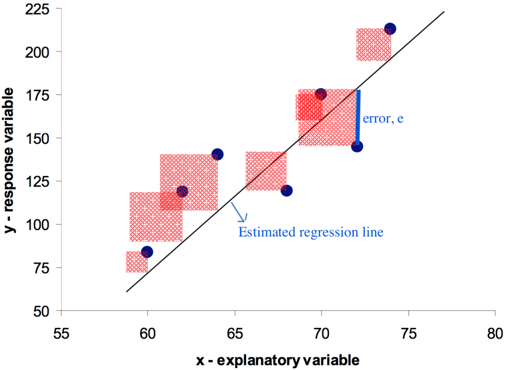
Estimation by Least Squares method

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.
- **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.

Least Square Method



- 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.

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

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

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

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

where

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

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

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

this also implies

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

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

Proof

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

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

These lead to the **Normal equations:**

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

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

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

Review on Positive Definite matrix

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

- In general, a symmetric matix 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 $a > 0$.

Equivalent formula for b_1

$$
\begin{aligned}
b_1 &= \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_1^n (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}} && (6) \\
&= \frac{\sum_1^n (X_i - \bar{X}) Y_i}{S_{xx}} && (7) \\
&= \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{xx}} y_i = \sum_{i=1}^{n} k_i Y_i && (8) \\
&= \frac{\sum_1^n X_i Y_i - n\bar{X}\bar{Y}}{S_{xx}} && (9)
\end{aligned}
$$

where (9) suggests that b_1 is a linear combination of Y_i (assume constant X) and hence is a linear estimator.

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

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

Equivalent formula for b_0

$$
\begin{aligned}
b_0 = \bar{Y} - b_1\bar{X} &= \sum_1^n \frac{1}{n} Y_i - \bar{X} \sum_1^n k_i Y_i \\
&= \sum_{i=1}^n (\frac{1}{n} - k_i\bar{X}) Y_i \\
&= \sum_1^n w_i Y_i, \quad w_i = \frac{1}{n} - k_i\bar{X}
\end{aligned}
$$

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 (2) and use them to prove (3) and (4))

1. $\sum_{i=1}^n k_i = 0$
2. $\sum_{i=1}^n k_i X_i = 1$
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.

Show the unbiasedness of b_0, b_1

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

$$\begin{aligned} E(b_1) &= E(\sum_1^n k_i Y_i) = \sum_1^n k_i E(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= \sum_1^n k_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_1^n k_i + \beta_1 \sum_1^n X_i k_i \\ &= 0 + \beta_1 \frac{\sum_1^n X_i^2 - n\bar{X}^2}{S_{xx}} = \beta_1 \end{aligned}$$

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

Proof that b_0 is the best

Proof that b_1 is the best (PP43-44)

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)

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

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

This is equivalent to show

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

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

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

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

thus

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

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

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

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

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

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

Topics for next week

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

Practice problems after Week 1 lectures

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. Install R and R Studio.
   - how to install R and R Studio on window `https://www.youtube.com/watch?v=MFfRQuQKGYg`
   - how to install R and R Studio on window `https://www.youtube.com/watch?v=Ywj6yNfc5nM`
4. Copy and paste the R code in R provided in the next 3 slides for Example 2. You should have the same output.
5. Try the exercises on slide 35.

Example 2: SLR Estimation (using R)

R code to find b_0, b_1

```r
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:

```
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
```