Supervised Learning: Simple Regression

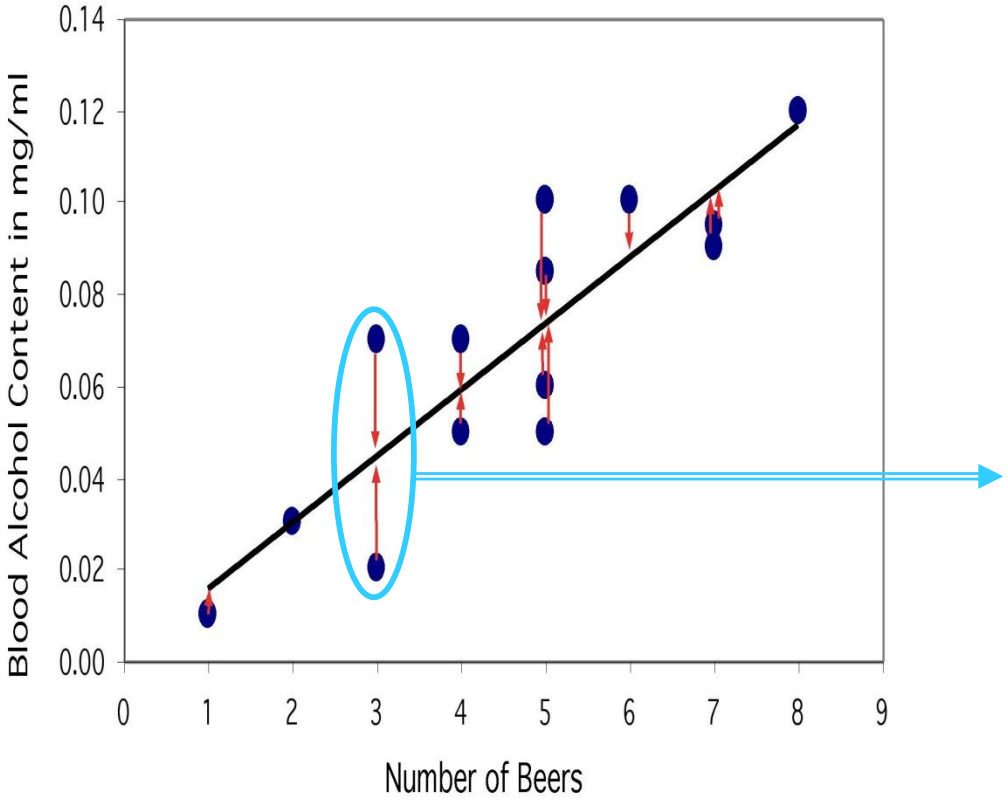
Dr. Yuzana Win (Nagasaki University, Japan)
Lecturer
Department of Computer Engineering and Information Technology

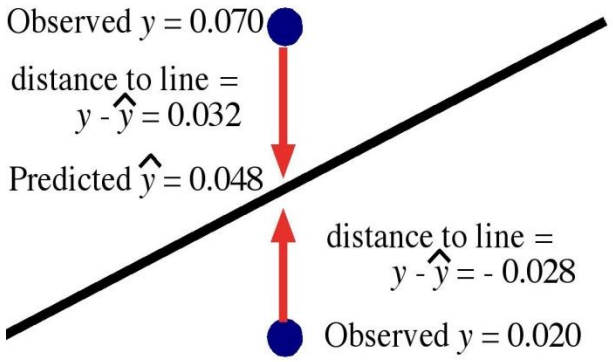
Lecture Objectives

- To introduce
  - What is Regression?
  - Regression Techniques or Methods
  - How does the Regression works and why we need it?

What is Regression?

- Regression is used to map a data item to a real valued prediction variable





What is Regression?

- The specific nature of the analysis target is represented by a single number
- Regression modeling represents a powerful and elegant method for estimating the value of continuous target variable
  - Choose a linear function so that f(x) represents that number
  - $f(\boldsymbol{x}) \approx t$ (Minimize the difference between the function output and the target value)

When would you need Regression?

- Regression is useful when we want to forecast a response using a new set of predictors.
  - For example, predict electricity consumption of a household for the next hour given the outdoor temperature, time of day, and the number of residents in that household

- Regression is used in many different fields:
  - Economy
  - Computer Science
  - Social sciences and so on

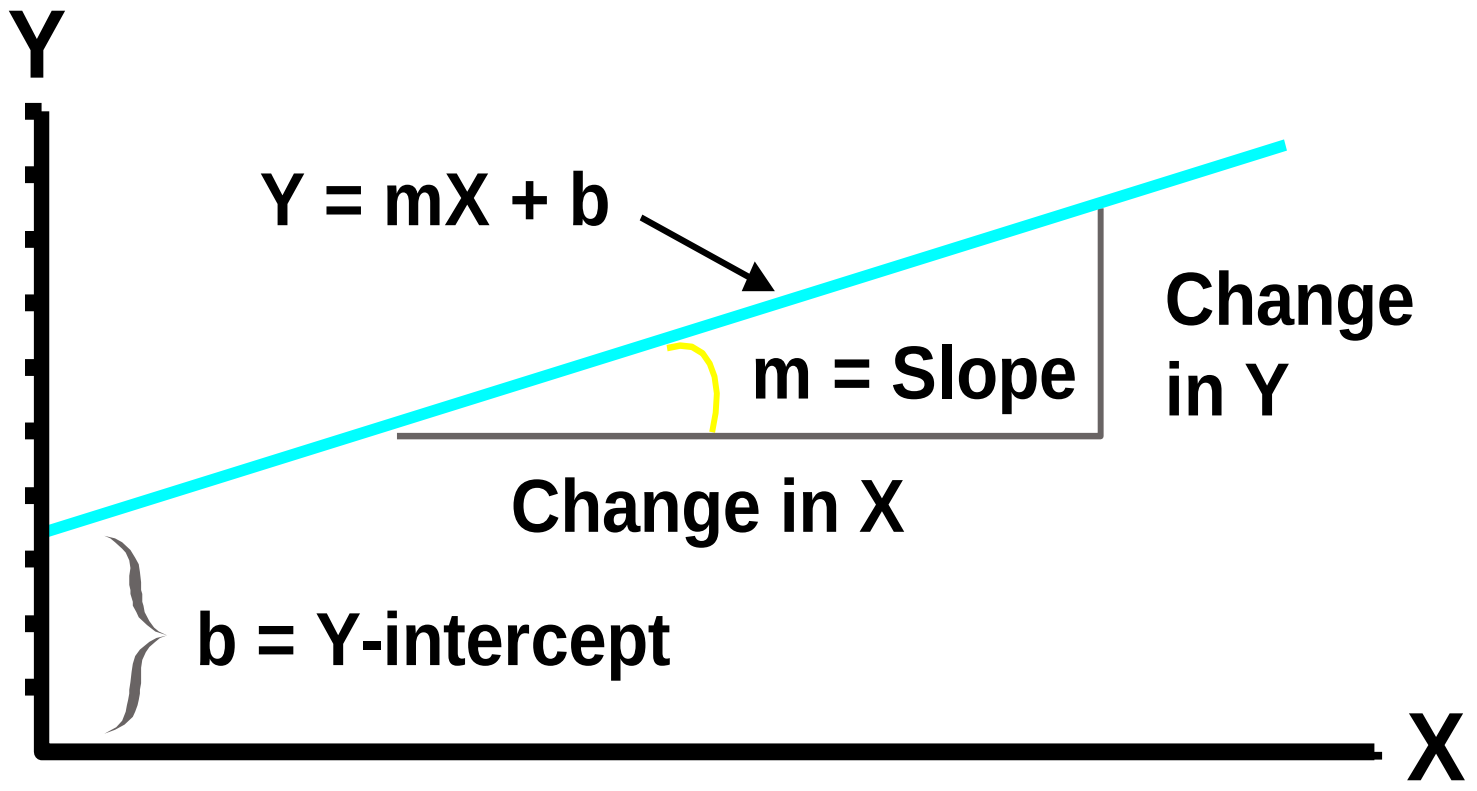
Regression Techniques or Methods

- Linear Regression
- Multiple Regression
- Logistic Regression
- LASSO

Linear Equation

- Linear Equation is

$$y = mx + b \quad \text{——————} \quad (1)$$



Simple Linear Regression

- examine the relationship between quantitative variables x and y via a mathematical equation 1
- Simple linear regression, where a straight line is used to approximate the relationship between x and y
- The motivation for using the technique:
  - Forecast the value of a dependent variable (y) from the value of independent variables (x_1, x_2,…x_k.)
  - Analyze the specific relationships between the independent variables and the dependent variable.

Simple Linear Regression

- The regression model is

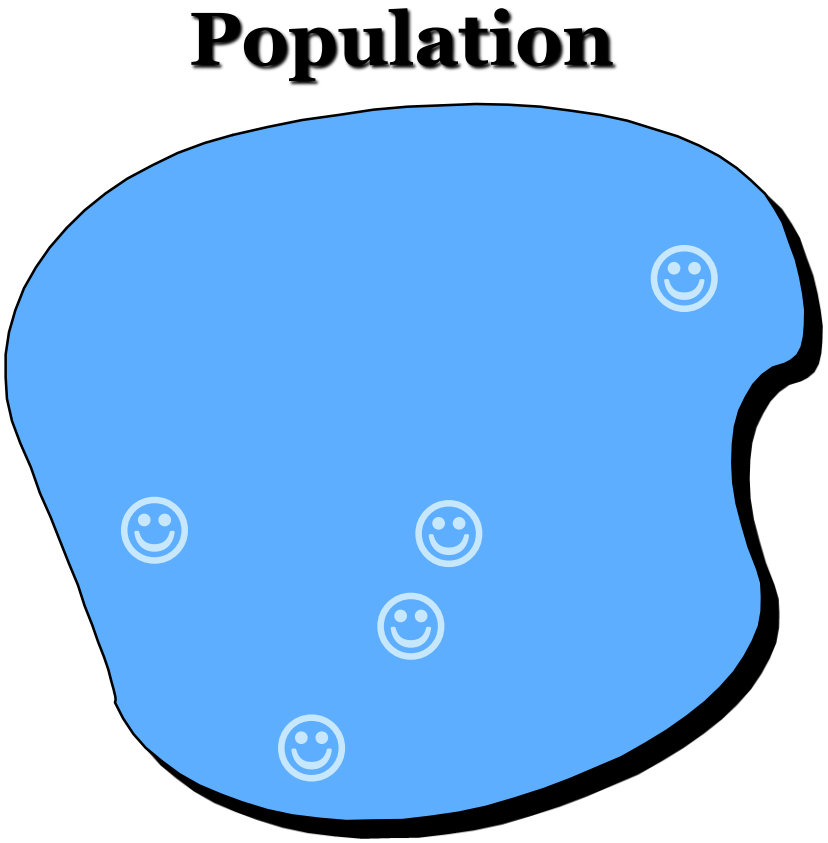
$$y = \beta_0 + \beta_1 x + \varepsilon$$

x is the independent variable
y is the dependent variable

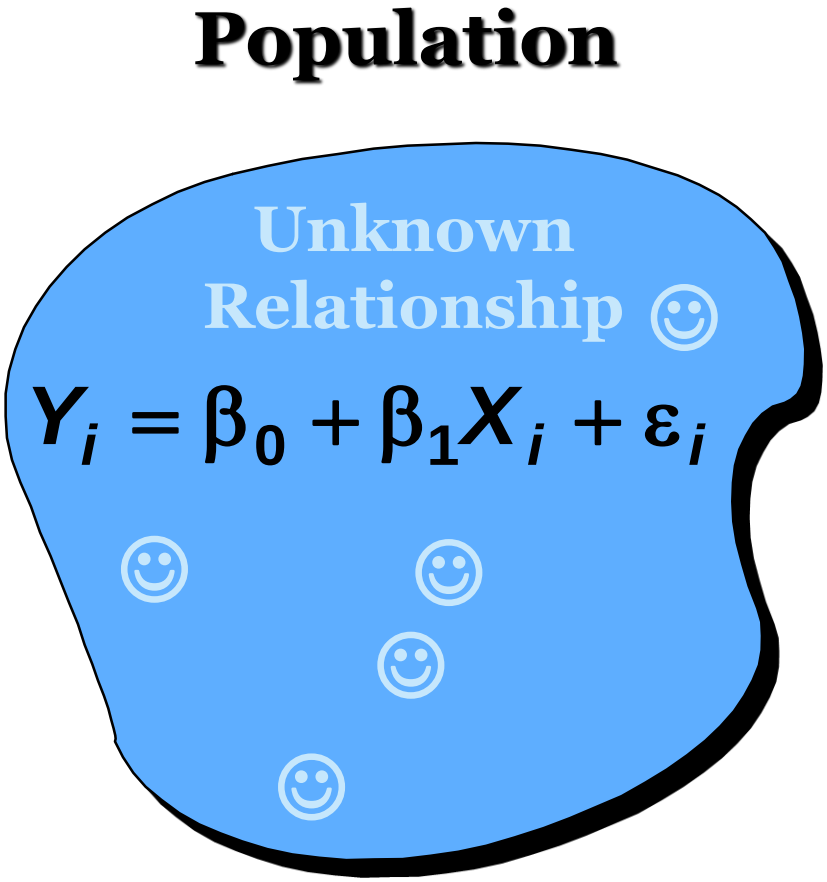
- The relationship between x and y is a linear or straight line relationship.

- Two parameters to estimate:
  - the slope of the line β_1 and the y-intercept β_0 (where the line crosses the vertical axis).
  - ε is random error component.

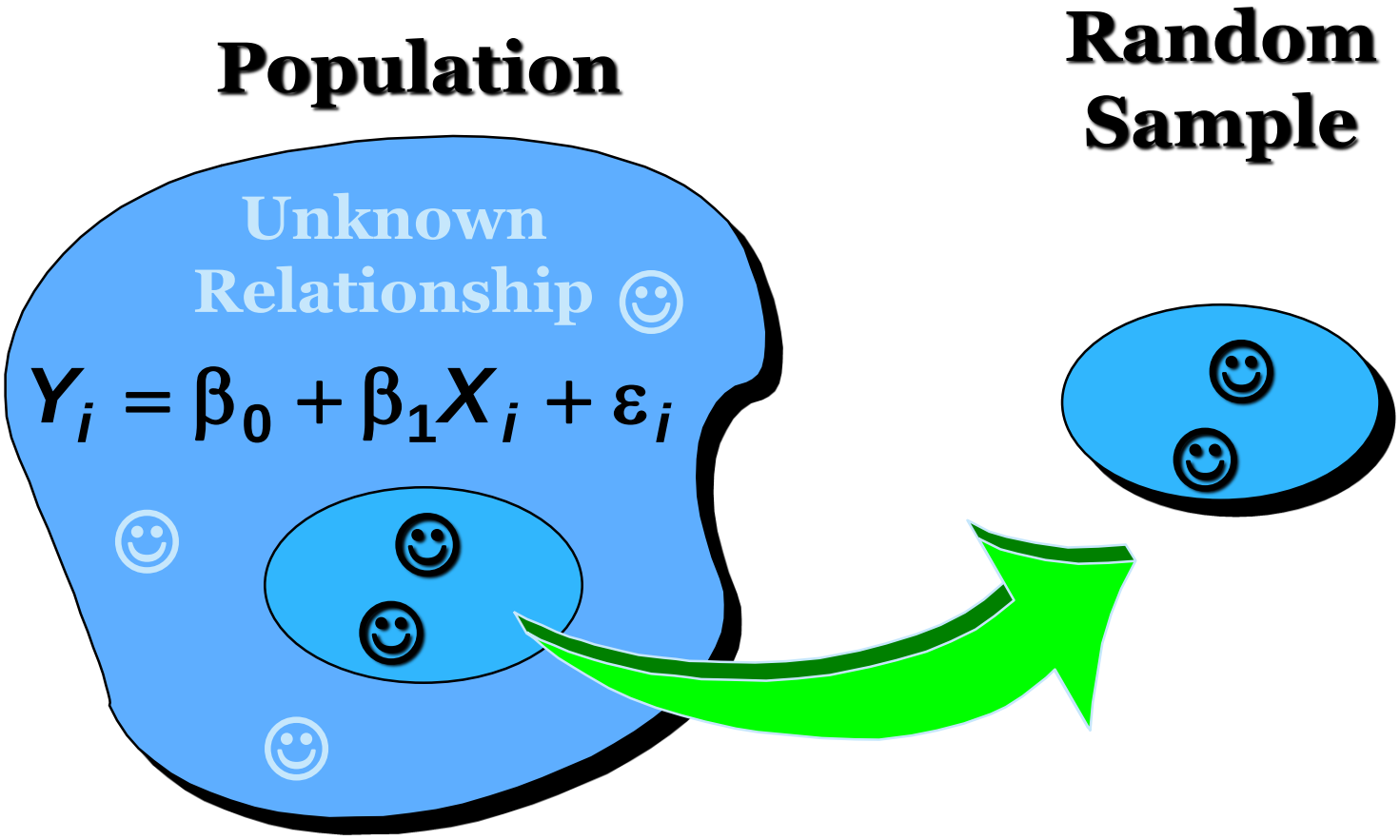
Population & Sample Regression



Population & Sample Regression

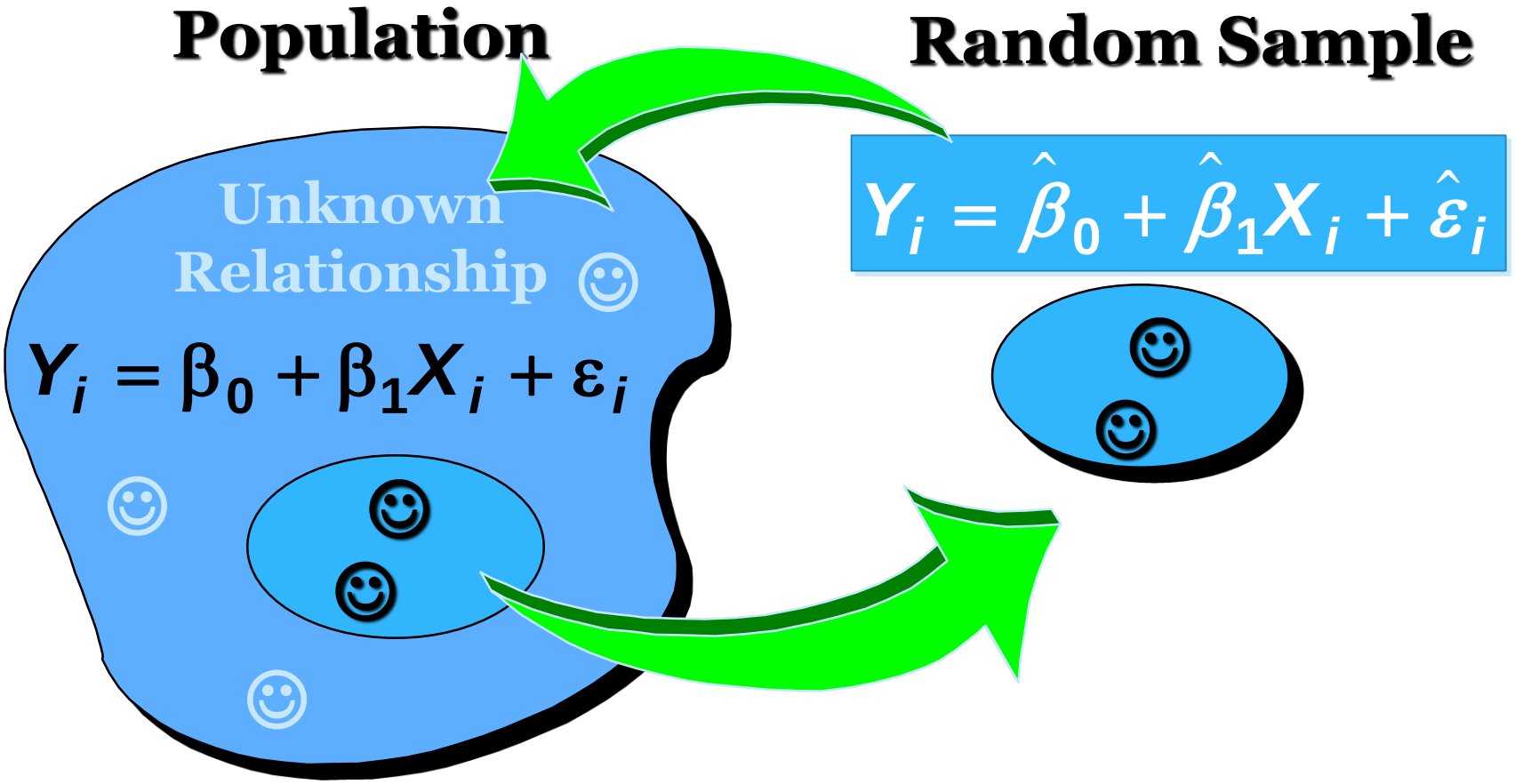


Population & Sample Regression



CEIT
Think smart, be different.

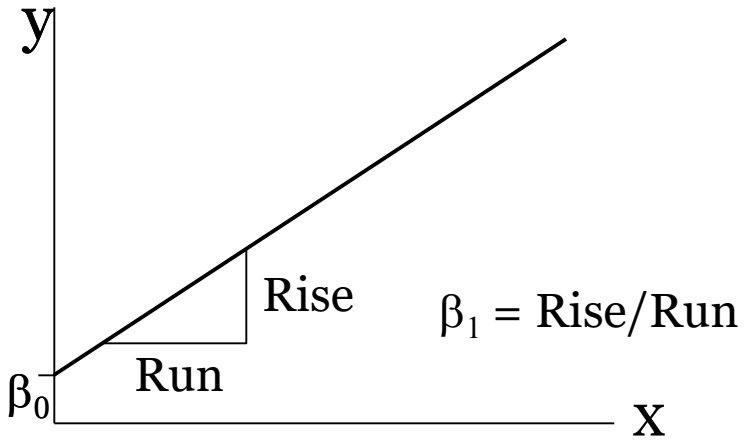
Population & Sample Regression



- From the sample of values of x and y, estimates b_0 of β_0 and b_1 of β_1 are obtained using the least squares or another method.
- The resulting estimate of the model is $\hat{y} = b_0 + b_1 x$

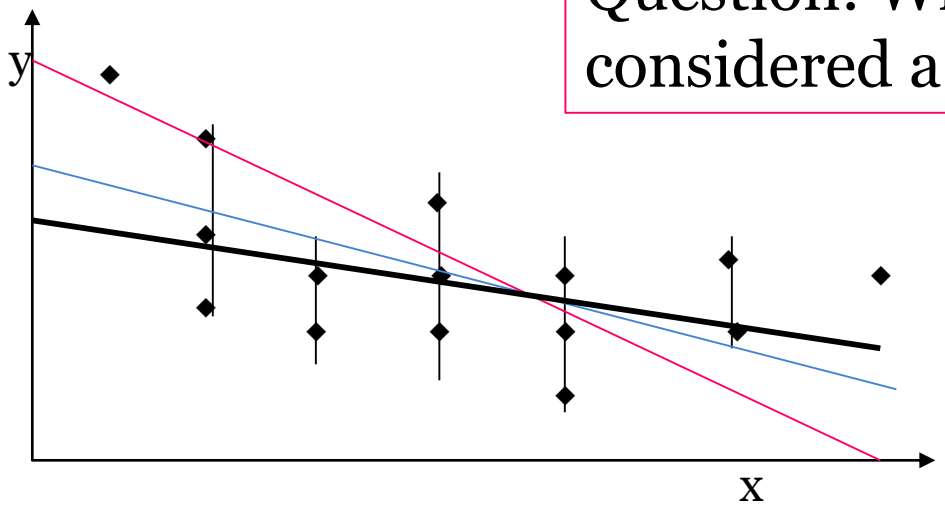
- The symbol $\hat{y}$ is termed "y hat" and refers to the predicted values of the dependent variable y that are associated with values of x, given the linear model.

β_0 and β_1 are unknown population parameters, therefore are estimated from the data.



Estimating the Coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data



Question: What should be considered a good line?

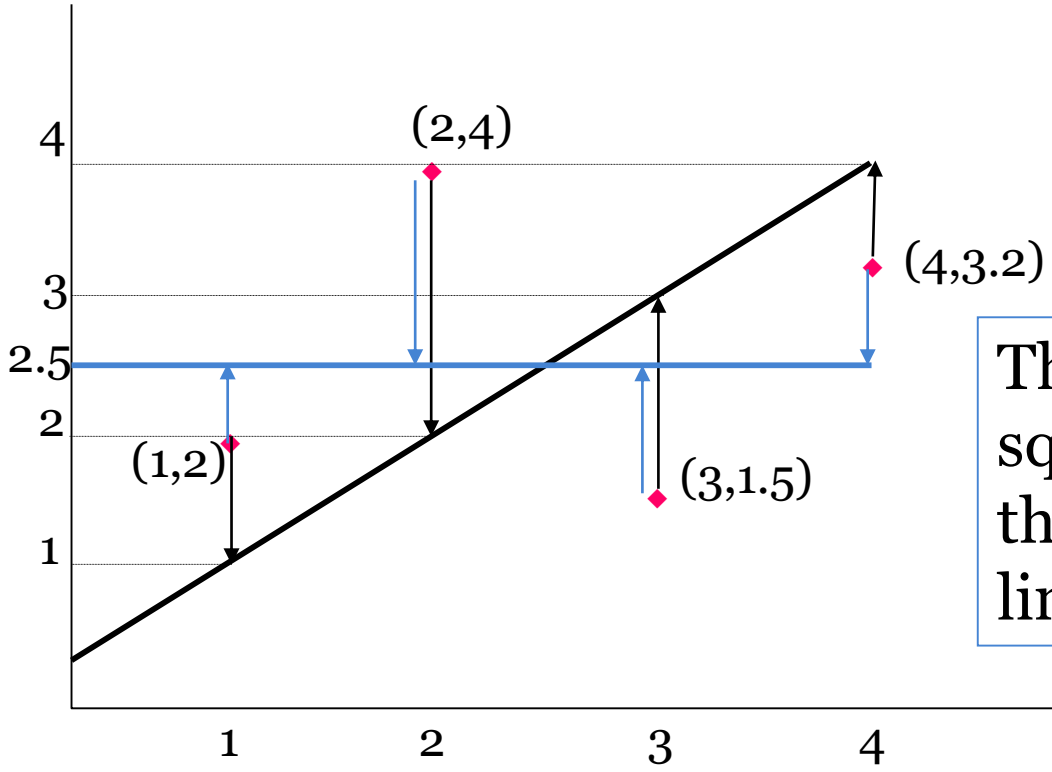
The Least Squares (Regression) Line

A good line is one that minimizes the sum of squared differences between the points and the line

The Least Squares (Regression) Line

Let us compare two lines

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$



The smaller the sum of squared differences, the better the fit of the line to the data.

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2$

$= 3.99$

Derivation of the formulas for estimating the y-intercept and slope of the estimated regression line

- we have $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \ldots, n$
- The least-squares line is that line which minimizes the population sum of squared errors,
  $SSE_p = \sum^n_{i=1} \varepsilon_i^2 \quad = \sum^n_{i=1} (y_i - \beta_0 - \beta_1 x_i)^2$
- find the value of β_0 and β_1 that minimize $\sum^n_{i=1} \varepsilon_i^2$ by differentiating equation with respect to β_0 and β_1 and setting the results equal to zero

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

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

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

Distributing the summation gives us

$$\sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

which is reexpressed as

$$b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- Solving equation for b_1 and b_0,

$$b_1 = \frac{\sum x_i y_i - [(\sum x_i)(\sum y_i)]/n}{\sum x_i^2 - (\sum x_i)^2/n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where n is the total number of observations, $\bar{x}$ is the mean value for the predictor variable, $\bar{y}$ is the mean value for the response variable, and the summations are i = 1 to n...

CEIT
Think smart, be different.

Example: Simple Linear Regression Model for *cereals* data set

- The *cereals* data set (Data and Story Library) contains nutritional information for *77 breakfast cereals* and includes the following variables:
  - Cereal name
  - Cereal manufacturer
  - Type (hot or cold)
  - Calories per serving
  - Grams of protein
  - Grams of fat
  - Milligrams of sodium
  - Grams of fiber
  - Grams of carbohydrates
  - Grams of sugar
  - Milligrams of potassium
  - Percentage of recommended daily allowance of vitamins (0%, 25%, or 100%)
  - Weight of one serving
  - Number of cups per serving
  - Shelf location (1, bottom; 2, middle; 3, top)
  - Nutritional rating, as calculated by *Consumer Reports*

Example: Simple Linear Regression Model for *cereals* data set

TABLE 2.1 Excerpt from the *Cereals* Data Set: Eight Fields, First 16 Cereals

Cereal Name	Manuf.	Sugars	Calories	Protein	Fat	Sodium	Rating
100% Bran	N	6	70	4	1	130	68.4030
100% Natural Bran	Q	8	120	3	5	15	33.9837
All-Bran	K	5	70	4	1	260	59.4255
All-Bran Extra Fiber	K	0	50	4	0	140	93.7049
Almond Delight	R	8	110	2	2	200	34.3848
Apple Cinnamon Cheerios	G	10	110	2	2	180	29.5095
Apple Jacks	K	14	110	2	0	125	33.1741
Basic 4	G	8	130	3	2	210	37.0386
Bran Chex	R	6	90	2	1	200	49.1203
Bran Flakes	P	5	90	3	0	210	53.3138
Cap'n crunch	Q	12	120	1	2	220	18.0429
Cheerios	G	1	110	6	2	290	50.7650
Cinnamon Toast Crunch	G	9	120	1	3	210	19.8236
Clusters	G	7	110	3	2	140	40.4002
Cocoa Puffs	G	13	110	1	1	180	22.7364

Example: Simple Linear Regression Model for cereals data set

TABLE 2.2 Summary Statistics for Finding b_0 and b_1

Cereal Name	Sugars, x	Rating, y	xy	x^2
100% Bran	6	68.4030	410.418	36
100% Natural Bran	8	33.9837	271.870	64
All-Bran	5	59.4255	297.128	25
All-Bran Extra Fiber	0	93.7049	0.000	0
Almond Delight	8	34.3848	275.078	64
Apple Cinnamon Cheerios	10	29.5095	295.095	100
Apple Jacks	14	33.1741	464.437	196
Basic 4	8	37.0386	296.309	64
Bran Chex	6	49.1203	294.722	36
Bran Flakes	5	53.3138	266.569	25
Cap'n Crunch	12	18.0429	216.515	144
Cheerios	1	50.7650	50.765	1
Cinnamon Toast Crunch	9	19.8236	178.412	81
Clusters	7	40.4002	282.801	49
Cocoa Puffs	13	22.7364	295.573	169
⋮	⋮	⋮		
Wheaties Honey Gold	8	36.1876	289.501	64
	$\sum x_i = 534$ $\bar{x} = 534/77$ $= 6.935$	$\sum y_i = 3285.26$ $\bar{y} = 3285.26/77$ $= 42.6657$	$\sum x_i y_i$ $= 19,186.7$	$\sum x_i^2$ $= 5190$

find that

$$b_1 = \frac{\sum x_i y_i - \left[\left(\sum x_i\right)\left(\sum y_i\right)\right]/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} = \frac{19,186.7 - (534)(3285.26)/77}{5190 - (534)^2/77}$$

$$= \frac{-3596.791429}{1486.675325} = -2.42$$

$$b_0 = \bar{y} - b_1\bar{x} = 42.6657 - 2.42(6.935) = 59.4$$

Example: Simple Linear Regression Model for *cereals* data set

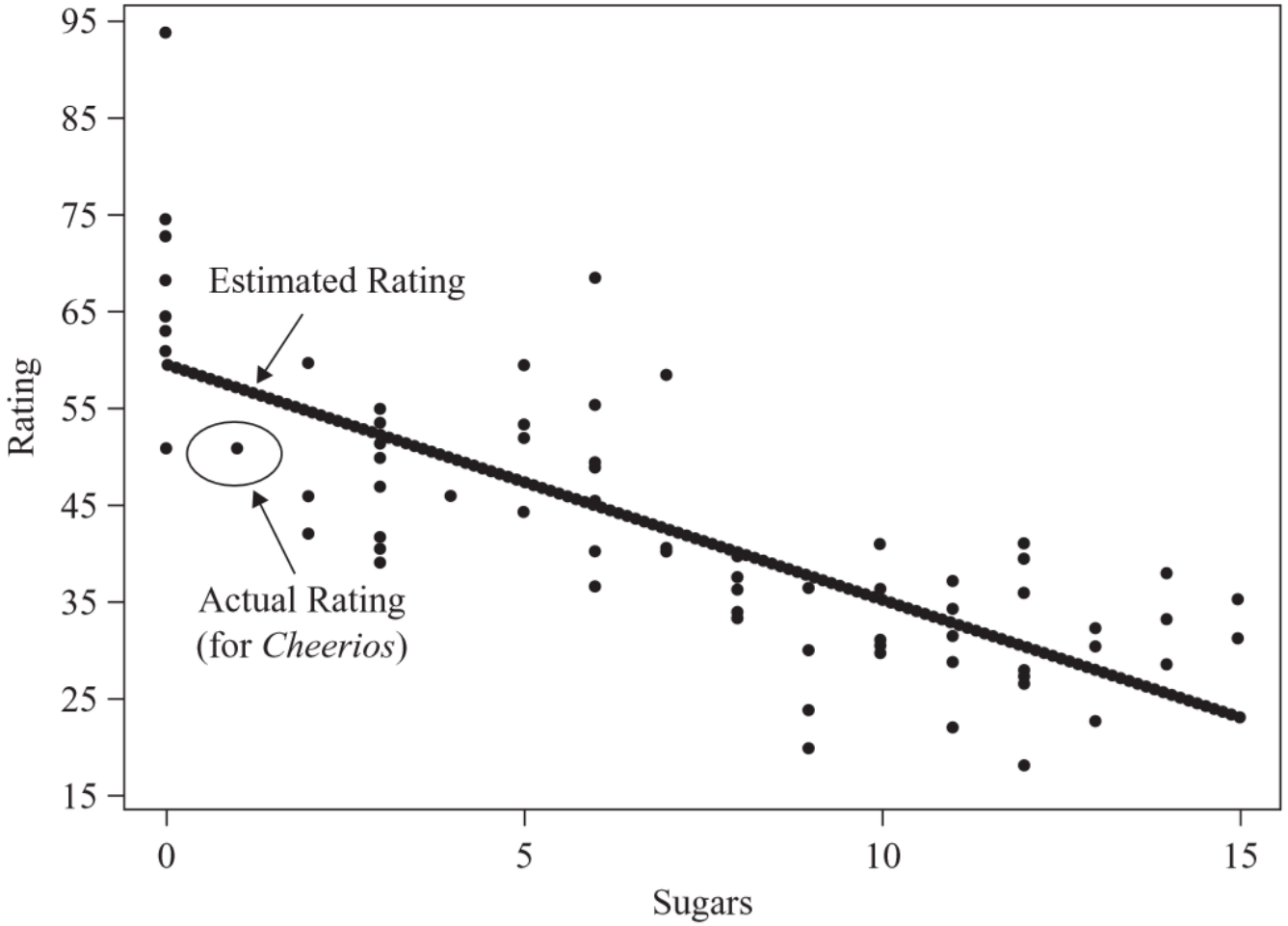


Figure 2.1 Scatter plot of nutritional rating versus sugar content for 77 cereals.

Example: Simple Linear Regression Model for *cereals* data set

- Figure 2.1 presents a scatter plot of the nutritional rating versus the sugar content for the 77 cereals, along with the least-squares regression line.
- The regression equation or estimated regression equation (ERE) is

$$\hat{y} = b_0 + b_1 x$$

- In this case, the ERE is given as

$$\hat{y} = 59.4 - 2.42(sugars).$$

So that, $b_0 = 59.4$ $b_1 = -2.42$ are obtained

Note: The slope of the regression line indicates the estimated change in y per unit increase in x.

Example: Simple Linear Regression Model for *Orienteering*

TABLE 2.3 SSE for the Orienteering Example

Subject	Time, x (hours)	Distance, y (km)	Score Predicted, $\hat{y} = 6 + 2x$	Error in Prediction, $y - \hat{y}$	(Error in Prediction)^2, $(y - \hat{y})^2$
1	2	10	10	0	0
2	2	11	10	1	1
3	3	12	12	0	0
4	4	13	14	−1	1
5	4	14	14	0	0
6	5	15	16	−1	1
7	6	20	18	2	4
8	7	18	20	−2	4
9	8	22	22	0	0
10	9	25	24	1	1
					$\text{SSE} = \sum (y - \hat{y})^2 = 12$

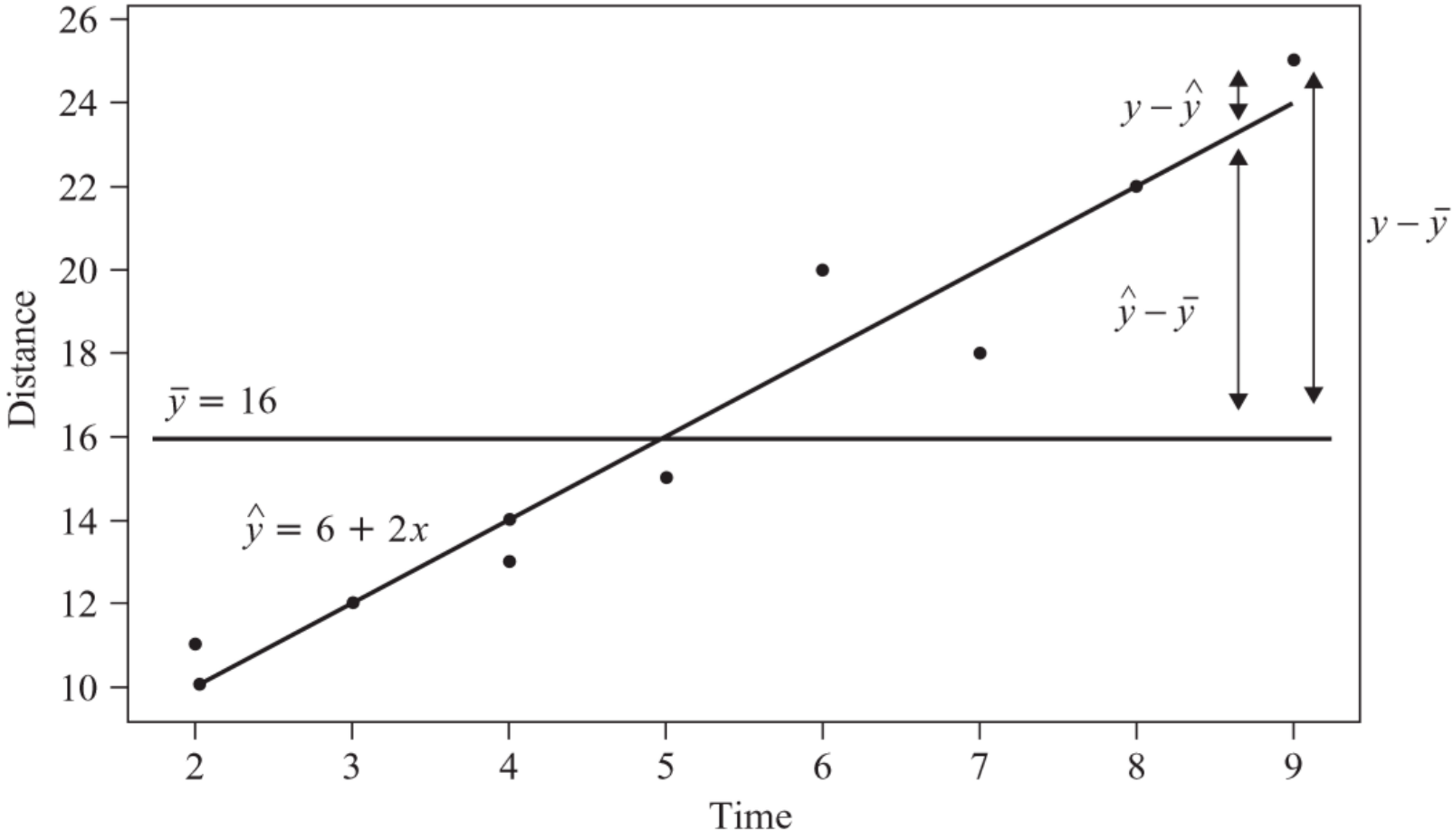


Figure 2.2 The regression line has a smaller prediction error than the sample mean.

Example: Simple Linear Regression Model for *Orienteering*

Subject	Time (x)	Distance (y)	xy	x^2
1	2	10	20	4
2	2	11	22	4
3	3	12	36	9
4	4	13	52	16
5	4	14	56	16
6	5	15	75	25
7	6	20	120	36
8	7	18	126	49
9	8	22	176	64
10	9	25	225	81
∑	50	160	908	304
Mean	5	16		

$$b_1 = \frac{\sum x_i y_i - \left[\left(\sum x_i\right)\left(\sum y_i\right)\right]/n}{\sum x_i^2 - \left(\sum x_i\right)^2/n} = \frac{908 - \left\{\frac{(50 * 160)}{10}\right\}}{304 - \left\{\frac{(50)2}{10}\right\}}$$

$$= \frac{908 - 800}{304 - 250} = \frac{108}{54} = 2$$

$$b_0 = \bar{y} - b_1\bar{x} : = 16 - (2 * 5) = 16 - 10 = 6$$

References

- Jiawei Han and Micheline Kamber. *Data Mining - Concepts and Techniques*. MorganKaufmann Publishers, 2001
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar:*Introduction to Data Mining*, Addison-Wesley

Next Week Lecture

- Unsupervised Learning: Association Analysis with Apriori Algorithm

Thank You