14^{th} Material Subject: Simple Regression Linear

Undergraduate of Telecommunication Engineering

MUH1F3 - PROBABILITY AND STATISTICS

Telkom University
Center of eLearning & Open Education Telkom University
Jl. Telekomunikasi No.1, Bandung - Indonesia
http://www.telkomuniversity.ac.id

Lecturer: Nor Kumalasari Caecar Pratiwi, S.T., M.T. (caecarnkcp@telkomuniversity.ac.id)

السلام عليكم ورحمة الله وبركاته

WELCOME

TABLE OF CONTENTS:

LEARNING OBJECTIVES:

After careful study of this chapter, student should be able to do the following:

1. **Use simple linear regression for building empirical models to engineering and scientific data**
2. **Understand how the method of least squares is used to estimate the parameters in a linear regression model**

SIMPLE LINEAR REGRESSION

There are many variables **x** and **y** that would appear to be related to one another, but not in a deterministic fashion. A familiar example is given by variables **x** $=$ **high school grade point average (GPA)** and **y** $=$ **college GPA**. The value of **y** cannot be determined just from knowledge of **x**, and two different individuals could have the same **x** value but have very different **y** values. Yet there is a tendency for those who have high (low) high school GPAs also to have high (low) college GPAs. Knowledge of a students high school GPA should be quite helpful in enabling us to predict how that person will do in college. Other examples **x** $=$ **age of a child** and **y = size of that child's vocabulary**, **x=size of an engine**$(\mathbf{cm^3})$ and **y=fuel efficiency for an automobile equipped with that engine**, etc.

SIMPLE LINEAR REGRESSION

Regression analysis is the part of statistics that investigates the relationship between two or more variables related in a nondeterministic fashion. In this chapter, we generalize the deterministic linear relation $\mathbf{y} = \alpha + \beta\mathbf{x}$ to a linear probabilistic relationship, develop procedures for making various inferences based on the model, and obtain a quantitative measure (the correlation coefficient) of the extent to which the two variables are related. The simplest deterministic mathematical relationship between two variables **x** and **y** is a linear relationship:

$$\mathbf{y} = \alpha + \beta\mathbf{x} \tag{1}$$

The set of pairs $(\mathbf{x}, \mathbf{y})$ for which determines a straight line with slope β and **y**-intercept α. The objective of this section is to develop a linear probabilistic model. For the deterministic model, the actual observed value of **y** is a linear function of **x**. The appropriate generalization of this to a probabilistic model assumes that the expected value of **Y** is a linear function of **x**, but that for fixed **x** the variable **Y** differs from its expected value by a random amount.

SIMPLE LINEAR REGRESSION

There are parameters α, β, and $\sigma^{\mathbf{2}}$, such that for any fixed value of the independent variable **x**, the dependent variable is a random variable related to **x** through the model equation.

$$\mathbf{Y} = \alpha + \beta\mathbf{x} + \epsilon \tag{2}$$

The quantity ϵ in the model equation is a random variable, assumed to be normally distributed with $\mathbf{E}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{Var}(\mathbf{x}) = \sigma^{\mathbf{2}}$.

SIMPLE LINEAR REGRESSION

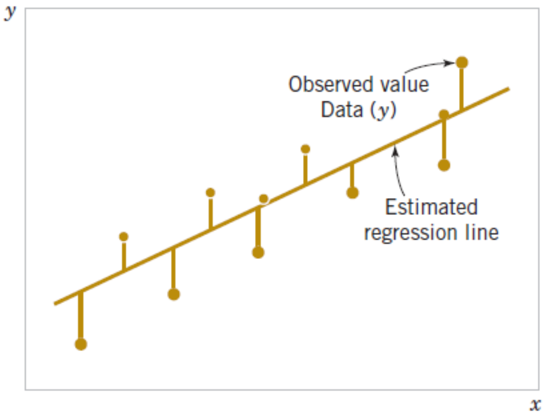

Figure 1: Deviations of the data from the estimated regression model

PROPERTIES OF THE LEAST SQUARES ESTIMATORS

The statistical properties of the least squares estimators $\overline{\alpha}$ and $\overline{\beta}$ may be easily described. Recall that we have assumed that the error term e in the model $\mathbf{Y} = \alpha + \beta\mathbf{x} + \epsilon$ is a random variable with $\mu = \mathbf{0}$ and $\mathbf{Var} = \sigma^2$.

$$\overline{\beta} = \frac{\left(n \cdot \sum_{i=1}^{n} X_i Y_i\right) - \left(\sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i\right)}{\left(n \cdot \sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{3}$$

While the intercept $\overline{\alpha}$

$$\overline{\alpha} = \overline{\mathbf{Y}} - \overline{\beta}\overline{\mathbf{X}} \tag{4}$$

Correlation Coefficients can be found by:

$$\rho = \frac{\left(n \cdot \sum_{i=1}^{n} X_i Y_i\right) - \left(\sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} Y_i^2\right) - \left(\sum_{i=1}^{n} Y_i\right)^2}} \tag{5}$$

PROPERTIES OF THE LEAST SQUARES ESTIMATORS

Average length of study / week (hours)	Mid Exam Score
2	35
3.2	39
6	55
8	80
10	94
3.9	52
8.2	88
6.5	65
9	89
2.6	30
8.5	75
7.5	86
4.5	51

PROPERTIES OF THE LEAST SQUARES ESTIMATORS

The data is the time spent studying and the Middle Exam score on Probability & Statistics for class A-01. Determine:

a. The Linear Regression Equation

b. If a student studies 1.5 hours / week, What is the estimated acquisition score of the Mid Exam

c. If a student studies for 11 Hours / Week, What is the estimated acquisition score of the Mid Exam d. The correlation coefficient between the variables **X** and **Y**

PROPERTIES OF THE LEAST SQUARES ESTIMATORS

Answer: Length of study time / week is the variable to be affect the acquisition score of Mid Exam, then the Length of Study is variable **X**. Whereas Mid Exam score is response, variables that are influenced by the Length of Study Time, the Value Mid Exam is the **Y** variable.

	X_i	Y_i	$X_i \cdot Y_i$	X_i^2	Y_i^2
	2	35	70	4	1225
	3.2	39	124.8	10.24	1521
	6	55	330	36	3025
	8	80	640	64	6400
	10	94	940	100	8836
	3.9	52	202.8	15.21	2704
	8.2	88	721.6	67.24	7744
	6.5	65	422.5	42.25	4225
	9	89	801	81	7921
	2.6	30	78	6.76	900
	8.5	75	637.5	72.25	5625
	7.5	86	645	56.25	7396
	4.5	51	229.5	20.25	2601
$\sum$	**79.9**	**839**	**5842.7**	**575.5**	**60123**
Rata-Rata	**6.15**	**64.54**			

PROPERTIES OF THE LEAST SQUARES ESTIMATORS

a. The Linear Regression Equation

$$\beta = \frac{\left(n \cdot \sum_{i=1}^{n} X_i Y_i\right) - \left(\sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i\right)}{\left(n \cdot \sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$\beta = \frac{\left(13 \cdot 5842.7\right) - \left(79.9 \cdot 839\right)}{\left(13 \cdot 575.5\right) - \left(79.9\right)^2} = \frac{8919}{1096.84} = 8.13$$

$$\alpha = \overline{Y} - b\overline{X}$$

$$\alpha = 64.54 - \left(8.13 \cdot 6.15\right) = 14.56$$

Simple Linear Regression equation becomes:

$$\overline{\mathbf{Y}} = \alpha + \beta\overline{\mathbf{X}}$$

$$\overline{\mathbf{Y}} = \mathbf{14.56} + \mathbf{8.13}\ \overline{\mathbf{X}}$$

PROPERTIES OF THE LEAST SQUARES ESTIMATORS

b. If a student studies 1.5 hours / week, What is the estimated acquisition score of the Mid Exam

$$\overline{X} = 1.5\ Hours \big/ Week$$

$$\mathbf{\overline{Y} = 14.56 + 8.13\,\overline{X} = 14.56 + 8.13 \cdot 1.5 = 26.755}$$

c. If a student studies for 11 Hours / Week, What is the estimated acquisition score of the Mid Exam

$$\overline{X} = 11\ Hours \big/ Week$$

$$\mathbf{\overline{Y} = 14.56 + 8.13\,\overline{X} = 14.56 + 8.13 \cdot 11 = 103.99}$$

PROPERTIES OF THE LEAST SQUARES ESTIMATORS

d. The correlation coefficient between the variables **X** and **Y**

$$\rho = \frac{\left(n \cdot \sum_{i=1}^{n} X_i Y_i\right) - \left(\sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} Y_i^2\right) - \left(\sum_{i=1}^{n} Y_i\right)^2}}$$

$$\rho = \frac{(13 \cdot 5842.7) - (79.9 \cdot 839)}{\sqrt{13\,(575.45) - (79.9)^2}\sqrt{13\,(60123) - (839)^2}}$$

$$\rho = \frac{8919}{\sqrt{1096.84} \cdot \sqrt{77678}} = \frac{8919}{33.12 \cdot 278.7} = \frac{8919}{9230.54} = 0.97$$

Thank You