201110

Regression and Correlation

Assoc.Prof Phisanu Chiawkhun
Statistics department
Faculty of Science
Chiangmai University

Introduction

- This chapter will discuss the methods of displaying and describing the relationship between two quantitative variables.
- The data used to study the relationship between two variables are bivariate data

The Bivariate data

- The bivariate data obtained by measuring both variables on the same individual unit. (X,Y)

  For example :

  The data of midterm score (X) and final score for a sample of students (Y).

  These data will help us study the association between two variables.

What do we study in this unit ?

- 1. We study the linear regression model , which is a method for developing and equation of a line that predicts the value of one quantitative variable from another quantitative variable.
- 2. We also study the correlation which measures the strength and direction of the linear relationship between two quantitative variables.

The linear regression model.

- When we construct the regression model, we use the bivariate data that we obtained by definition :

  The response or <u>dependent variable </u>is the variable that we want to predict <u>denoted by “Y ” </u>and the <u>independent variable denoted by “X”</u>

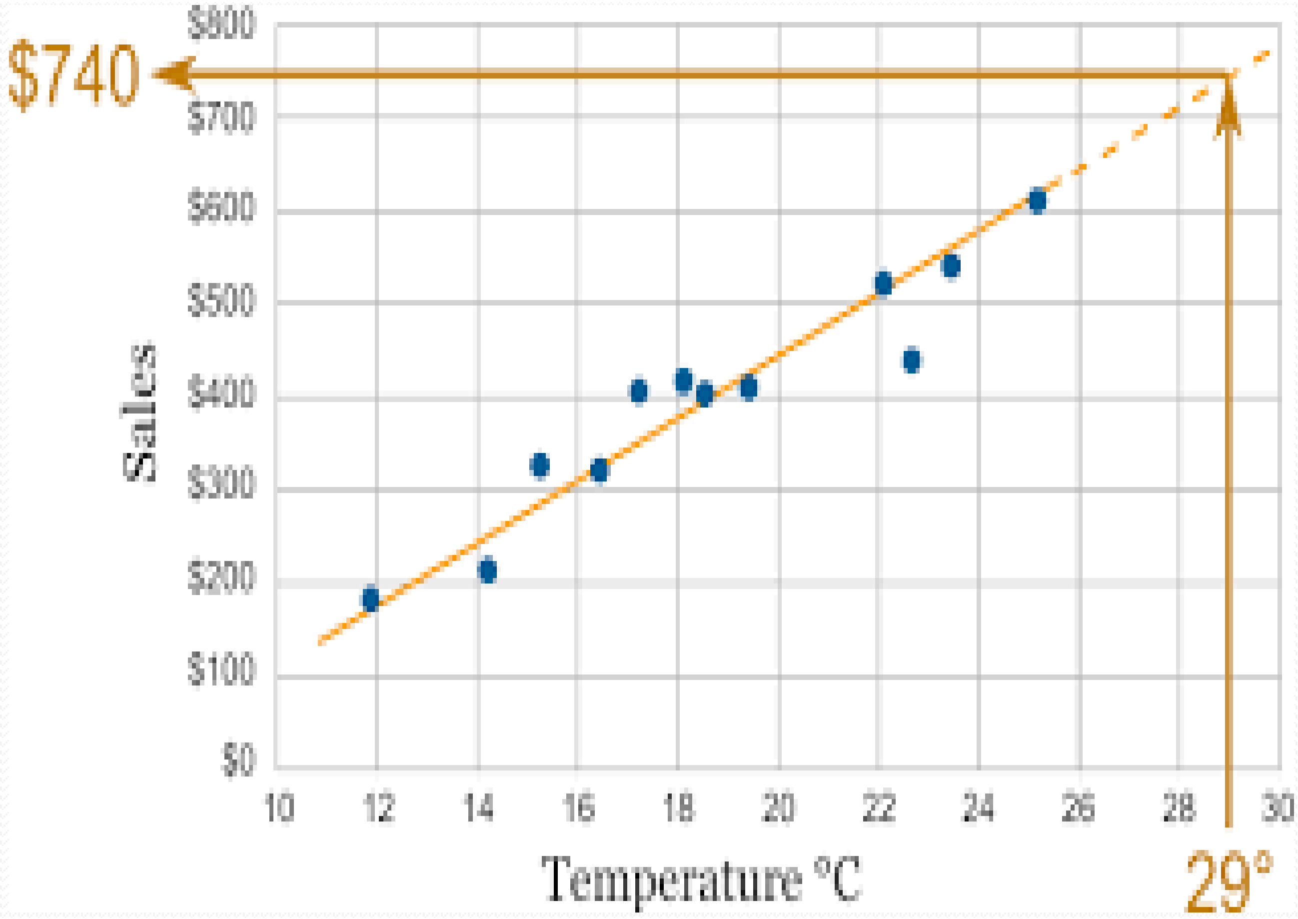
The linear regression model.

- The data that we collected in term of bivariate data (X,Y) .
- We construct the scatter plot to show the relationship between two quantitative variables.

  X variable are marked on the horizontal axis.

  Y variable are marked on the vertical axis.

The linear regression model.



Linear regression equation

- The method we will use for finding the regression line is called least squares regression the resulting is call the least squares regression equation

$$y = A + Bx + \varepsilon$$

Where y is response or dependent variable

x is response or dependent variable

B is slope of the line or regression coefficient

A is y-intercept

ε is random errors

- The slope (b) is amount of increase (or decrease) in Y for every 1-unit increase in X.
- Y-intercept is the value of Y when X is set equal to zero.
- We will construct the sample regression equation by

$$\hat{y} = a + bx$$

Calculating Linear regression equation

- We can calculate the linear regression equation by using the least squares method. The result expressions for a and b are given as

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$b = \frac{n\sum xy - (\sum x)(\sum (y))}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

Example 1

- A small set of data with n =5 observations on x is the midterm exam score and y is the final exam score , has been in the table shown below

X: midterm exam score	Y : final exam score
8	9
10	13
12	14
14	15
16	19

Construct the scatter plot and estimate the regression equation .

Solution

- For computing the a and b. we calculate :

$$\sum x = 60 \qquad \sum y = 70 \qquad \sum xy = 884$$

$$\sum x^2 = 760 \qquad \sum y^2 = 1{,}032 \qquad \bar{x} = 12 \qquad \bar{y} = 14$$

- $$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$b = \frac{884 - 5(12)(14)}{760 - 5(12)^2} = \frac{44}{40} = 1.1$$

- $$a = \bar{y} - b\bar{x}, \qquad a = 14 - (1.1)12 = 0.8$$

- The regression equation is ,

$$\hat{y} = 0.8 + 1.1x$$

- b =1.1 means that if X (midterm score) increase 1 unit y (final score) will increase 1.1 units
- a = 0.8 means that if x =0 y is 0.8
- If X = 10 we can estimate the value of Y by
- Y ^ = 0.8 + 1.1(10) = 11.8