Stat 332

Regression Analysis

Chapter 1

Linear Regression with One Predictor Variable

- Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that a response or outcome variable can be predicted from the other, or others.
- This methodology is widely used in business, the social and behavioral sciences, the biological sciences, and many other disciplines.

A few examples of applications are:

1- Sales of a product can be predicted by utilizing the relationship between sales and amount of advertising expenditures.

2- The performance of an employee on a job can be predicted by utilizing the relationship between performance and a battery of aptitude tests.

3- The size of the vocabulary of a child can be predicted by utilizing the relationship between size of vocabulary and age of the child and amount of education of the parents.

4- The length of hospital stay of a surgical patient can be predicted by utilizing the relationship between the time in the hospital and the severity of the operation.

➢ Relations between Variables

Functional Relation between Two Variables (mathematical relations)

Example (1-1)

$$Y=2X$$

X Units	*Y (Dollars)* Sales
75	150
25	50
130	260

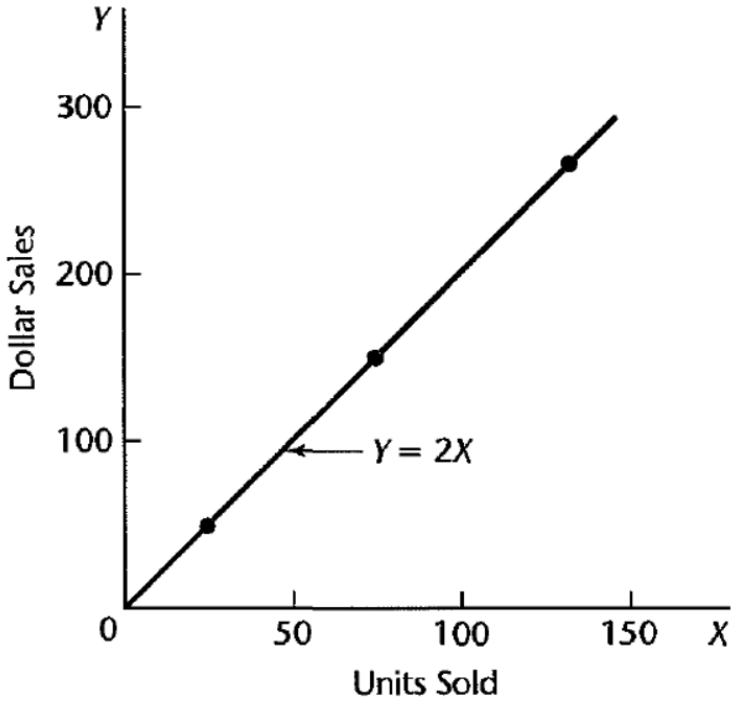


Fig.(1.1)

➢ Statistical Relation between Two Variables

Example (1-2)

Performance evaluations for 10 employees were obtained at midyear and at year-end. These data are plotted in Figure 1.2. Year-end evaluations are taken as the *dependent* or *response variable Y*, and midyear evaluations as the *independent, explanatory*, or *predictor*

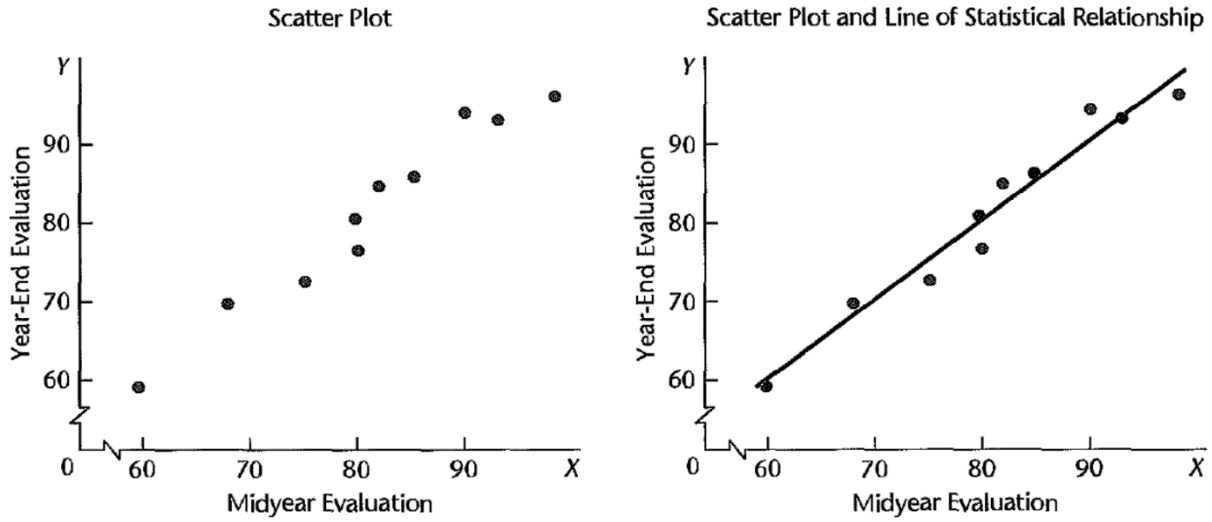
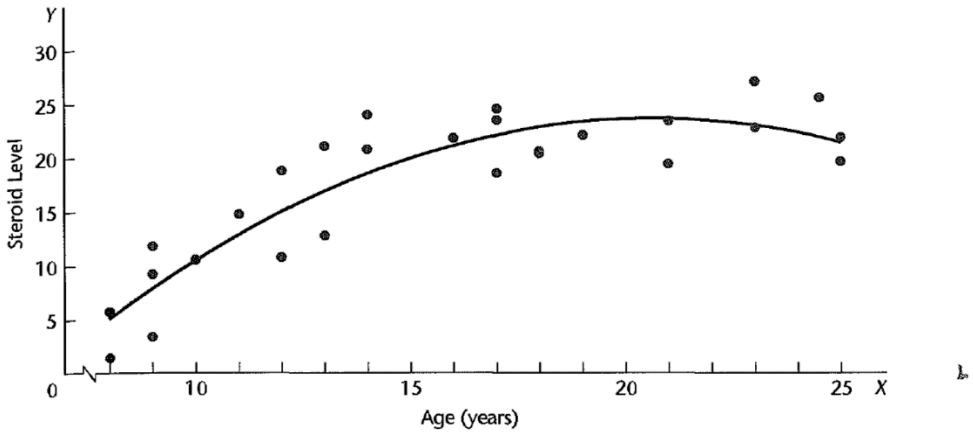


Figure (1.2)

Example (1-3)

Figure 1.3 presents data on age and level of a steroid in plasma for 27 healthy females between 8 and 25 years old. The data strongly suggest that the statistical relationship is *curvilinear* (not linear). The curve of relationship has also been drawn in Figure 1.3. It implies that, as age increases, steroid level increases up to a point and then begins to level off. Note again the scattering of points around the curve of statistical relationship, typical of all statistical relations

FIGURE 1.3 Curvilinear Statistical Relation between Age and Steroid Level in Healthy Females Aged 8 to 25.



What is the functional relations to describe Fig 1.2 and Fig 1.3?

It is a regression relation or regression model !

➢ Uses of Regression Analysis

Regression analysis serves three major purposes:

(1) description
(2) control
(3) prediction

➢ Regression and Causality

The existence of a statistical relation between the **response variable *Y*** and the **explanatory or predictor variable X** does not imply in any way that *Y* depends causally on X.

No matter how strong is the statistical relation between X and *Y*, no cause-and-effect pattern is necessarily implied by the regression model.

For example, data on size of vocabulary *(X)* and writing speed *(Y)* for a sample of young children aged 5-10 will show a positive regression relation. This relation does not imply, however, that an increase in vocabulary causes a faster writing speed.

Here, other explanatory variables, such as age of the child and amount of education, affect both the vocabulary *(X)* and the writing speed *(Y)*. Older children have a larger vocabulary and a faster writing speed.

Simple Linear Regression Model with Distribution of Error Terms Unspecified

Let

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1.1)$$

where:

Y_i is the value of the response variable in the ith trial
β_0 and β_1 are parameters
X_i is a known constant, namely, the value of the predictor variable in the ith trial

X_i is a known constant, namely, the value of the predictor variable in the ith trial
ε_i is a random error term with mean $E\{\varepsilon_i\} = 0$ and variance $\sigma^2\{\varepsilon_i\} = \sigma^2$; ε_i and ε_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for all $i, j; i \neq j$)
$i = 1, \ldots, n$

Remark: we will use the symbol $Var(\varepsilon_i) = \sigma^2\{\varepsilon_i\} = \sigma^2$

Important Features of the Model are:

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

$$Var(Y_i) = \sigma^2\{Y_i\} = \sigma^2$$

Example (2-4)

A consultant for an electrical distributor is studying the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week and the time required to prepare the bids. Suppose that regression model (1.1) is applicable and is as follows:

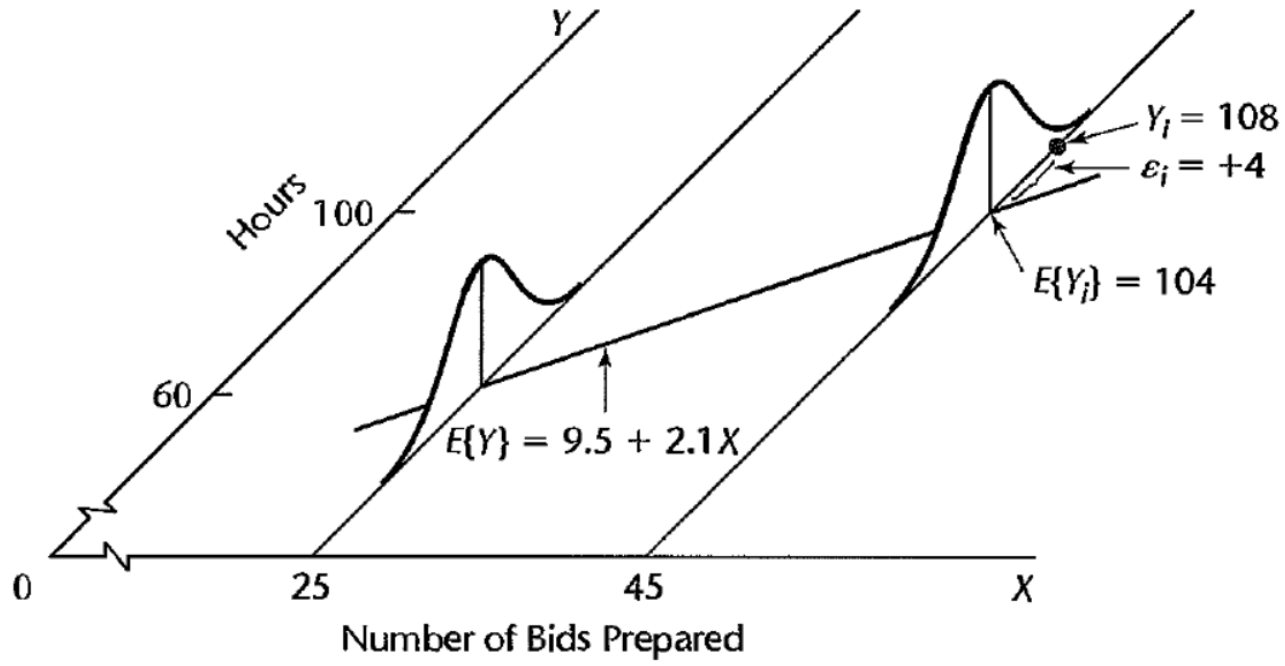
$$Y_i = 9.5 + 2.1X_i + \varepsilon_i$$

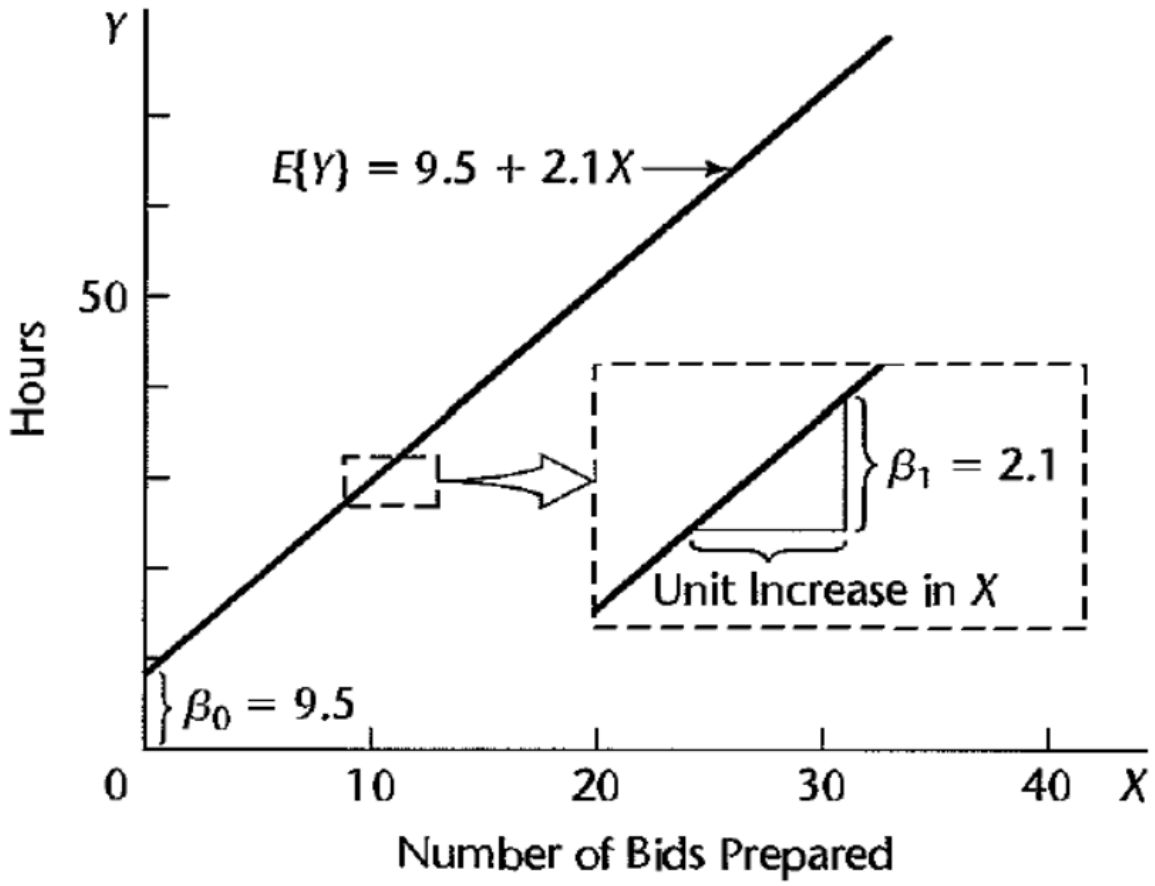
where X is the number of bids prepared in a week and Y is the number of hours required to prepare the bids. Figure 1.6 contains a presentation of the regression function:

$$E\{Y\} = 9.5 + 2.1X$$

Suppose that in the ith week, $X_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$. In that case, the error term value is $\varepsilon_i = 4$, for we have

$$E\{Y_i\} = 9.5 + 2.1(45) = 104$$




Y
E{Y} = 9.5 + 2.1X
50
Hours
β1 = 2.1
Unit Increase in X
β0 = 9.5
0
10
20
30
40
X
Number of Bids Prepared