Chapter 1: Linear Regression with One Predictor Variable

Regression analysis is a statistical methodology that utilizes the relation between two or more _____ variables so that a _____ or _____ 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.

Examples of Regression applications:

- Sales of a product can be predicted by utilizing the relationship between sales and amount of _____.
- The performance of an employee on a job can be predicted by utilizing the relationship between performance and a _____.
- The size of the vocabulary of a child can be predicted by utilizing the relationship between size of vocabulary and _____ of the child and _____.
- 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.
-

Note 1. In the first few chapters, we take up regression analysis when a _____ predictor variable is used for _____ the response or outcome variable of interest.

1.1 Relations between Variables

We look at the difference between a *functional relation* and a *statistical relation.*

Functional Relation between Two Variables

A functional relation between two variables is expressed by a mathematical formula. If X denotes the independent variable and Y the dependent variable, a functional relation is of the form:

Given a particular value of X, the function f indicates the corresponding value of Y.

Example 1. Consider the relation between dollar sales (Y) of a product sold at a fixed price and number of units sold (X). If the selling price is \$2 per unit, the relation is expressed by the equation:

$$Y = 2X$$

This functional relation is shown in Table 1. Number of units sold and dollar sales during three recent periods (while the unit price remained constant at \$2) were as follows:

Period	Units Sold	Sales
1	75	
2	25	
3	130	

Note 2. Note that all fall directly _____ of functional relationship. **This is characteristic of all functional relations.**

Statistical Relation between Two Variables

A statistical relation, unlike a functional relation, is not a _____ one. In general, the observations for a statistical relation _____ fall directly on-the curve of relationship.

Example 2. Performance evaluations for 10 employees were obtained at midyear and at year-end. These data are plotted in Figure 1. Year-end evaluations are taken as the dependent or response variable Y, and midyear evaluations as the independent, explanatory, or predictor variable X.

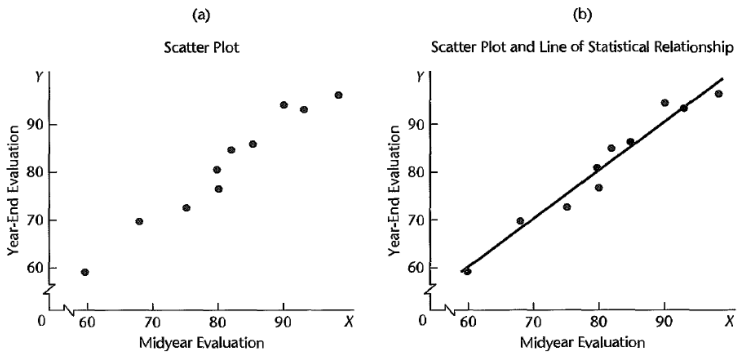


Fig. 1. Statistical Relation

Note 3. Because of the scattering of points in a statistical relation, Figure 1 is called a scatter diagram or scatter plot. In statistical terminology, each point in the scatter diagram represents a trial or a case.

Example 3. Figure 2 presents data on _____ and _____ in plasma for _____ healthy females between 8 and 25 years old. The data strongly suggest that the statistical relationship is _____ (not linear). The curve of relationship has also been drawn in Figure 2. It implies that, as _____ 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.

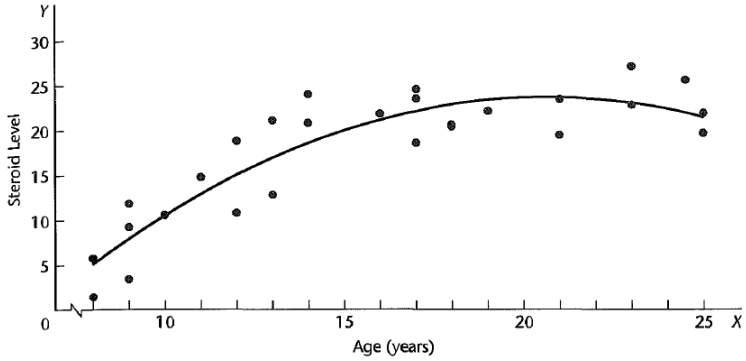


Fig. 2. Curvilinear Statistical Relation

1.3 Simple Linear Regression Model with Distribution of Error Terms Unspecified

Formal Statement of Model

We consider a basic regression model where there is only one predictor variable and the regression function is linear. The model can be stated as follows:

- This regression model is said to be _____, _____ in the _____, and _____ in the _____ variable.
- It is "simple" in that there is _____,
- "linear in the parameters," because no parameter appears as an _____ or is _____ by another parameter,
- "linear in the predictor variable," because this variable appears only in the _____.
- A model that is linear in the parameters and in the predictor variable is also called a _____ model.

Important Features of Model

1. The response Y_i in the i^{th} trial is the sum of two components:

   (a)

   (b)

2. Since $E\{\epsilon_i\} = 0$,

   Thus, the response Y_i, when the level of X in the i^{th} trial is X_i, comes from a probability distribution whose mean is:

We therefore know that the regression function is:

since the regression function relates the means of the probability distributions of Y for given X to the level of X.

3. The response Y_i in the i^{th} trial exceeds or falls short of the value of the regression function by the error term amount _____.
4. The error terms ϵ_i are assumed to have constant variance σ_2 It therefore follows that the responses Y_i have the same constant variance:

Thus, regression model assumes that the probability distributions of Y have the same variance σ_2, regardless of the level of the predictor variable X.

5. The error terms are assumed to be uncorrelated. Since the error terms ϵ_i and ϵ_j are uncorrelated, so are the responses Y_i and Y_j.

Example 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 our regression model is applicable and is as follows:

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

where X is the number of bids prepared in a week and Y is the number of hours required to prepare the bids.

Figure 4 contains a presentation of the regression function:

Suppose that in the i^{th} week, $X_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$. In that case, what is the error term value? $\epsilon_i =$

Figure 4 displays the probability distribution of Y when $X = 45$ and indicates from where in this distribution the observation $Y_i = 108$ came. Note again that the error term ϵ_i is simply the _____ of _____ from its _____ value _____.

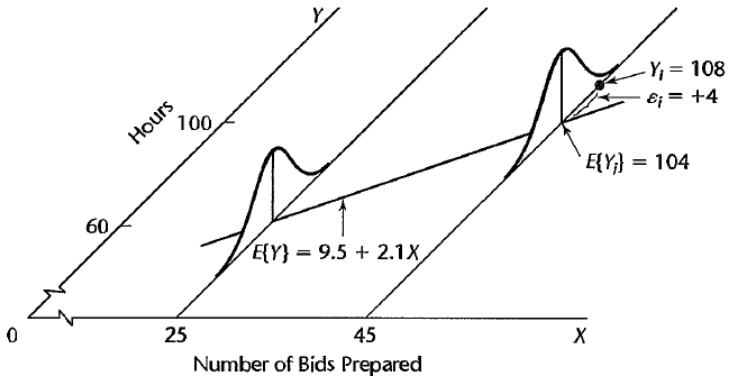


Fig. 3. Statistical Relation

Meaning of Regression Parameters

- The parameters _____ and _____, in the regression model are called _____.
- _____ is the _____ of the regression line
- It indicates,

- _____ is the _____ of the regression line
- When the _____ of the model includes $X = 0$,

Example 5. For the electric distribution example above, interpret the regression coefficients.

1.4 Data for Regression Analysis

Data for regression analysis may be obtained from nonexperimental or experimental studies.

Observational Data

- Observational data are data obtained from _____ studies.
- These studies do not control the _____ of interest.
- Example: company officials wished to study the relation between age of employee ___ and number of days of illness last year ___. The needed data for use in the regression analysis were obtained from personnel records. Such data are observational data since the explanatory variable, age, is not controlled.
- Regression analyses are frequently based on observational data, since often it is not feasible to conduct controlled experimentation. In the company personnel example just mentioned, for instance, it would not be possible to control age by assigning ages to persons.
- A major limitation of observational data is that they often do not provide adequate information about _____. For example, a positive relation between age of employee and number of days of illness in the company personnel example may not imply that number of days of illness is the direct result of age. It might be that younger employees of the company primarily work indoors while older employees usually work outdoors, and that work location is more directly responsible for the number of days of illness than age.

Experimental Data

- It is possible to conduct a controlled experiment to provide data from which the regression parameters can be estimated.
- Example: an insurance company that wishes to study the relation between productivity of its analysts in processing claims ___ and length of training ___. Nine analysts are to be used in the study. Three of them will be selected at random and trained for two weeks, three for three weeks, and three for five weeks.
- The productivity of the analysts during the next 10 weeks will then be observed. The data so obtained will be experimental data because control is exercised over the explanatory variable, length of training.

When control over the explanatory variable(s) is exercised through random assignments, the resulting experimental data provide much stronger information about **cause-and-effect relationships** than do observational data.

– The reason is that randomization tends to balance out the effects of any other variables that might affect the response variable, such as the effect of aptitude of the employee on productivity.

Completely Randomized Design

Read page no.13 in the textbook and explain what "completely randomized design" means.

1.5 Overview of Steps in Regression Analysis

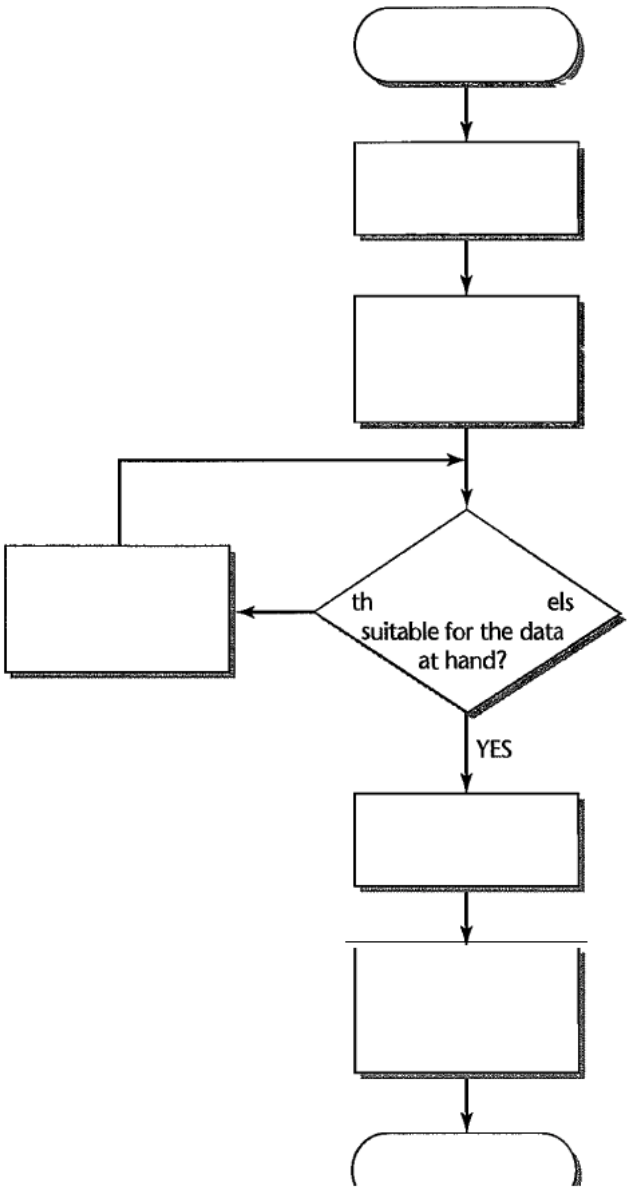


Fig. 4. Typical Strategy for Regression Analysis

1.6 Estimation of Regression Function

Example 6. In a small-scale study of persistence, an experimenter gave three subjects a very difficult task. Data on the age of the subject (X) and on the number of attempts to accomplish the task before giving up (Y) follow:

Subject i	1	2	3
Age X_i	20	55	30
number of attempts Y_i	5	12	10

Method of least-Squares

To find "good" estimators of the regression parameters _____ and _____, we use the method of least squares. For the observations _____ for each case, the method of least squares considers the deviation of _____ from its expected value:

In particular, the method of least squares requires that we consider the _____. This criterion is denoted by Q:

According to the method of least squares, the estimators of _____ and _____ are those values b_____ and _____ respectively, that minimize the criterion Q for the given sample observations _____.

Example 7. Figure below shows the Persistence study data with the regression line:

$$\hat{Y} = 9 + 0(X)$$

The vertical deviation for the first case is:

The vertical deviation for the second case is:

The vertical deviation for the third case is:

Then, the criterion Q is:

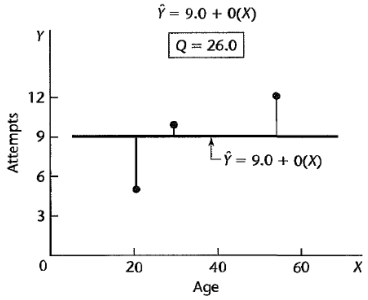


Figure below shows the Persistence study data with the regression line:

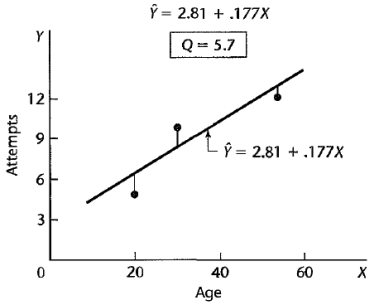
$$\hat{Y} = 2.81 + 0.177X$$

The vertical deviation for the first case is:

The vertical deviation for the second case is:

The vertical deviation for the third case is:

Then, the criterion Q is:



Analytical procedure to find Least Squares(LS) Estimators (proof)

For given sample observations (X_i, Y_i), the quantity Q is a function of _____ and _____.

Find the values of _____ and _____, that minimize Q by differentiating Q with respect to _____ and _____:

Example 8. The Toluca Company manufactures refrigeration equipment as well as many replacement parts. In the past, one of the replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum lot size for producing this part. The production of this part involves setting up the production process (which must be done no matter what is the lot size) and machining and assembly operations. One key input for the model to ascertain the optimum lot size was the relationship between lot size and labor hours required to produce the lot To determine this relationship, data on lot size and work hours for 25 recent production runs were utilized. The production conditions were stable during the six-month period in which the 25 runs were made and were expected to continue to be the same during the next three years, the planning period for which the cost improvement program was being conducted. Use the table below to calculate the least squares estimates b_o and b_l

Run i	Lot size X_i	Work hours Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	80	399	10	86.72	867.2	100	7,520.4
2	30	121	-40	-191.28	7,651.2	1,600	36,588.0
3	50	221	-20	-91.28	1,825.6	400	8332.0
.	.	.	.	.	.	.	.
23	40	244	-30	-68.28	2,048.4	900	4,662.2
24	80	342	10	29.72	297.2	100	883.3
25	70	323	0	10.72	0.0	0	114.9
Total	1,750	7,807	0	0	70,690	19,800	307,203

Finding LS estimates in *R*: Refer the R codes provided

Point Estimation of Mean Response

Given sample estimators b_o and b_l of the parameters in the regression function:

we estimate the regression function as follows:

where _____ (read _____) is the value of the estimated regression function at the level X of the predictor variable.

- We call a value of the response variable a response and $E\{Y\}$ the _____.
- The mean response stands for the mean of the probability distribution of Y corresponding to the level X of the predictor variable.
- _____ then is a point estimator of the mean response when the level of the predictor variable is X.
- For the cases in the study, we will call _____, the _____

Example 9. For the Toluca Company example, we found that the least squares estimates of the regression coefficients are: $b_o = 62.37$ $b_1 = 3.5702$.

1. What is the estimated regression function?
2. Estimate that the mean number of work hours required for production runs of $X = 65$ units.

3. Interpret this value

Residuals

The _____ residual is the difference between the observed value _____ and the corresponding fitted value _____. This residual is denoted by _____ and is defined in general as follows:

For our regression model, the residual e_i becomes:

Example 10. Calculate the residuals for the first two cases of the Toluca Company example. Use the LS equation from the previous example.

The residuals for the first two cases of the Toluca Company example are illustrated graphically in Figure 5

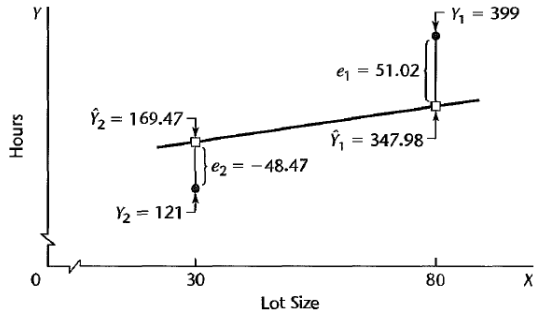


Fig. 5. Residuals of first two cases of the Toluca Company example

Note 4. **Model error term and the residual**

- The model error term value: _____ is the vertical deviation of _____ from the _____ true regression line and hence is _____.
- The residual: _____ is the vertical deviation of _____ from the fitted value _____on the estimated regression line, and it is known.

Properties of Fitted Regression line

1. The sum of the residuals is zero:

2. The sum of the squared residuals, _____, is a minimum.
3. The sum of the observed values _____; equals the sum of the fitted values _____:

4. The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial:

5. From the properties 1) and 4), the sum of the weighted residuals is zero when the residual in the ith trial is weighted by the fitted value of the response variable for the ith trial:

6. The regression line always goes through the point _____.

1.7 Estimation of Error Terms Variance σ^2

Recall that the variance of each observation Y_i for regression our model _____ is _____, the same as that of each error term _____.

An estimator of the standard deviation a is simply _____, the positive square root of _____.

Example 11. Use the *R* output from the Toluca Company example to answer the following questions

1. What is s (the Residual standard error or estimator of the standard deviation σ)
2. What is MSE (mean square error or residual mean square)

3. What is SSE (error sum of squares or residual sum of squares)

1.8 Normal Error Regression Model

The normal error regression model is as follows:

where:

_____ is the observed response in the ith trial

_____ is a known constant, the level of the predictor variable in the ith trial

_____ and _____, are parameters

_____ are independent _____