

Chapter Seven

ANALYSIS OF VARIANCE (ANOVA)

7.1. Using Statistics

7.2. Hypothesis Test of Analysis of Variance

7.3. Theory and the Computations of ANOVA

7.4. ANOVA Table and Examples

After studying this chapter, you should be able to:

- Explain the purpose of ANOVA.
- Describe the model and computations behind ANOVA.
- Explain the test statistic F.
- Conduct a one-way ANOVA.
- Report ANOVA results in an ANOVA table.

7.1. Using Statistics

Analysis of variance is the first of several advanced statistical techniques to be discussed in this chapter. Along with regression analysis, described in the next chapter, ANOVA is the most commonly quoted advanced research method in the professional business and economic literature. What is analysis of variance? The name of the technique may seem misleading.

ANOVA is a statistical method for determining the existence of differences among several population means. While the aim of ANOVA is to detect differences among several population *means*, the technique requires the analysis of different forms of *variance* associated with the random samples under study- hence the name ***analysis of variance***. The original ideas of analysis of variance were developed by the English statistician **Sir Ronald A. Fisher** during the first part of the 20th century. ANOVA is well known and mostly use the area dealt with agricultural experiments where crops were given different “treatments,” such as being grown using different kinds of fertilizers. The researchers wanted to determine whether all treatments under study were equally effective or whether some treatments were better than others. **Better** is referred to those treatments that would produce crops of greater average weight. This question is answerable by the analysis of variance. Since the original work involved different *treatments*, the term remained, and we use it interchangeably with *populations* even when no actual treatment is administered. Thus, for example, if we compare the mean income in four different communities, we may refer to the four populations as four different *treatments*. In the next section, we will develop the simplest form of analysis of variance - the one-factor, fixed-effects and completely randomized design model.

7.2. Hypothesis Test of Analysis of Variance

The hypothesis test of analysis of variance is as follows:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_r$$

$$H_A: \text{Not all } \mu_i \text{ (} i = 1, 2, \dots, r \text{) are equal}$$

There are k populations, or treatments, under study. We draw an independent random sample from each of the k populations. The size of the sample from population i ($i = 1, 2, \dots, r$) is n_i , and the total sample size is $n = n_1 + n_2 + \cdots + n_r$.

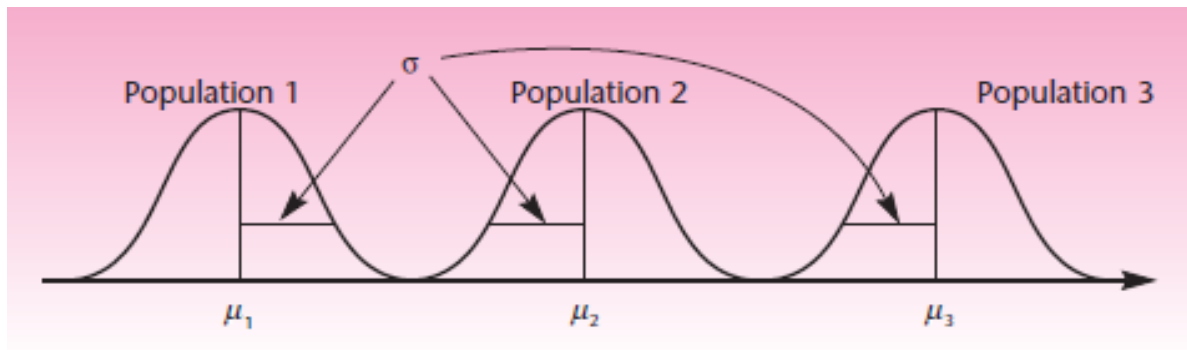
From the r samples we compute several different quantities, and these lead to a computed value of a test statistic that follows a known F distribution when the null hypothesis is true and some assumptions hold. From the value of the test statistic and the critical value for a given level of significance, we are able to make a determination of whether we believe that the r population means are equal.

Usually, the number of compared means r is greater than 2. Why greater than 2? If r is equal to 2, then the test for equality of two population means; although we could use ANOVA to conduct such a test, we have seen relatively simple tests of such hypotheses: the two-sample t tests of independency.

In this chapter, we are interested in investigating whether *several* population means may be considered equal. This is a test of a joint hypothesis about the equality of several population parameters. But why can we not use the two-sample t tests repeatedly? Suppose we are comparing $r = 5$ treatments. Why can we not conduct all possible pairwise comparisons of means using the two-sample t test? There are 10 such possible comparisons (10 choices of five items taken two at a time, found by using a combinatorial formula ${}^nC_r = {}^5C_2 = 10$). It should be possible to make all 10 comparisons. However, if we use, say, $\alpha = 0.05$ for each test, then this means that the probability of committing a type I error in any particular test (deciding that the two population means are not equal when indeed they are equal) is 0.05. If each of the 10 tests has a 0.05 probability of a type I error, what is the probability of a type I error if we state, "Not all the means are equal" (i.e., rejecting H_0).

If we need to compare more than two populations' means and we want to remain in control of the probability of committing a type I error, we need to conduct a joint test. Analysis of variance provides such a joint test of the hypotheses. The reason for ANOVA's widespread applicability is that in many situations we need to compare more than two populations simultaneously. Even in cases in which we need to compare only two treatments, say, test the relative effectiveness of two different prescription drugs, our actual test may require the use of a third treatment: a control treatment.

Figure 7.1: Three normally distributed populations with different means but with equal variance



The required assumptions of ANOVA:

- 1) We assume independent random sampling from each of the r populations.
- 2) We assume that the r populations under study are normally distributed, with means μ_i that may or may not be equal, but with equal variances (δ^2).

Suppose, for example, that we are comparing three populations and want to determine whether the three populations mean μ_1 , μ_2 and μ_3 are equal. We draw separate random samples from each of the three populations under study, and we assume that the three populations are distributed as shown in figure 7.1.

These model assumptions are necessary for the test statistic used in analysis of variance to possess an F distribution when the null hypothesis is true. If the populations are not exactly normally distributed, but have distributions that are close to a normal distribution, the method still yields good results. If, however, the distributions are highly skewed or otherwise different from normality, or if the population variances are not approximately equal, then ANOVA should not be used, and instead we must use a nonparametric technique called the Kruskal-Wallis test.

The Test Statistic

As mentioned earlier, when the null hypothesis is true, the test statistic of analysis of variance follows an F distribution. F distribution has two kinds of degrees of freedom: degrees of freedom for the numerator and degrees of freedom for the denominator.

In the analysis of variance, the numerator degrees of freedom are $r - 1$, and the denominator degrees of freedom are $n - r$. Analysis of variance is an involved technique, and it is difficult and time-consuming to carry out the required computations by hand. Consequently, computers are indispensable in most situations involving analysis of variance, and we will make extensive use of the computer in this chapter. For now, let us assume that a computer is available to us and that it provides us with the value of the test statistic.

ANOVA test statistic = $F(r - 1, n - r)$

Figure 7.2: The ANOVA test statistic for $r = 4$ populations and a total sample size $n = 54$

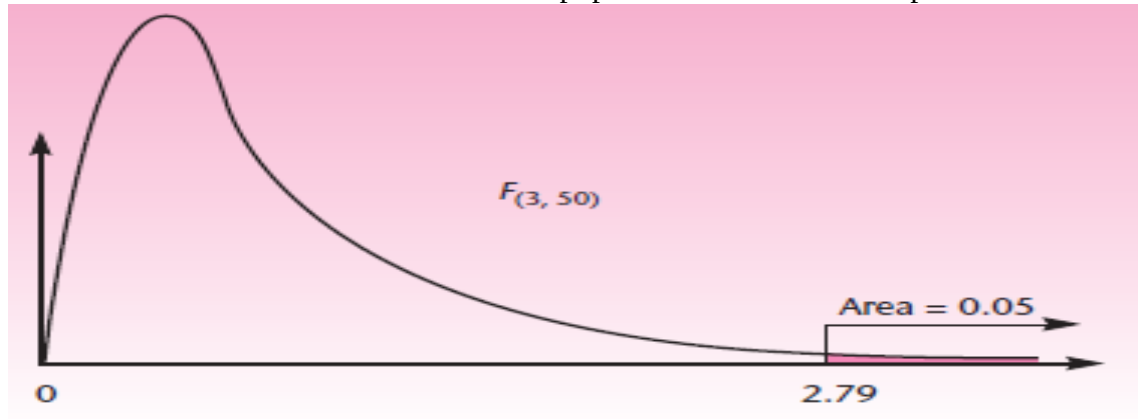


Figure 7.2 shows the F distribution with 3 and 50 degrees of freedom, which would be appropriate for a test of the equality of four population means using a total sample size of 54. Also shown is the critical point for $\alpha = 0.05$, found in F table. The critical point is 2.79. For reasons explained in the next section, the test is carried out as a right-tailed test.

We now have the basic elements of a statistical hypothesis test within the context of ANOVA: the null and alternative hypotheses, the required assumptions, and a distribution of the test statistic when the null hypothesis is true.

7.3. Theory and the Computations of ANOVA

Recall that the purpose of analysis of variance is to detect differences among several population means based on evidence provided by random samples from these populations. How can this be done? We want to compare r population means. We can use r random samples, one from each population. Each random sample has its own mean. The mean of the sample from population i will be denoted by $\bar{x}_i$. We may also compute the mean of all data points in the study, regardless of which population they come from. The mean of all the data points (when all data points are considered a single set) is called the *grand mean* and is denoted by $\bar{\bar{X}}$. These means are given by the following equations.

The mean of sample i ($i = 1, 2, \dots, r$) is:-

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

The **grand mean**, the mean of all the data points, is

$$\bar{\bar{X}} = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}}{n}$$

Where, X_{ij} the particular data is point in position j within the sample from population i . The subscript i denote the population, or treatment, and runs from 1 to r . The subscript j denotes the data point within the sample from population i ; thus, j runs from 1 to n_i .

In example 7.1, $r = 3, n_1 = 21, n_2 = 20, n_3 = 22$, and $n = n_1 + n_2 + n_3 = 63$.

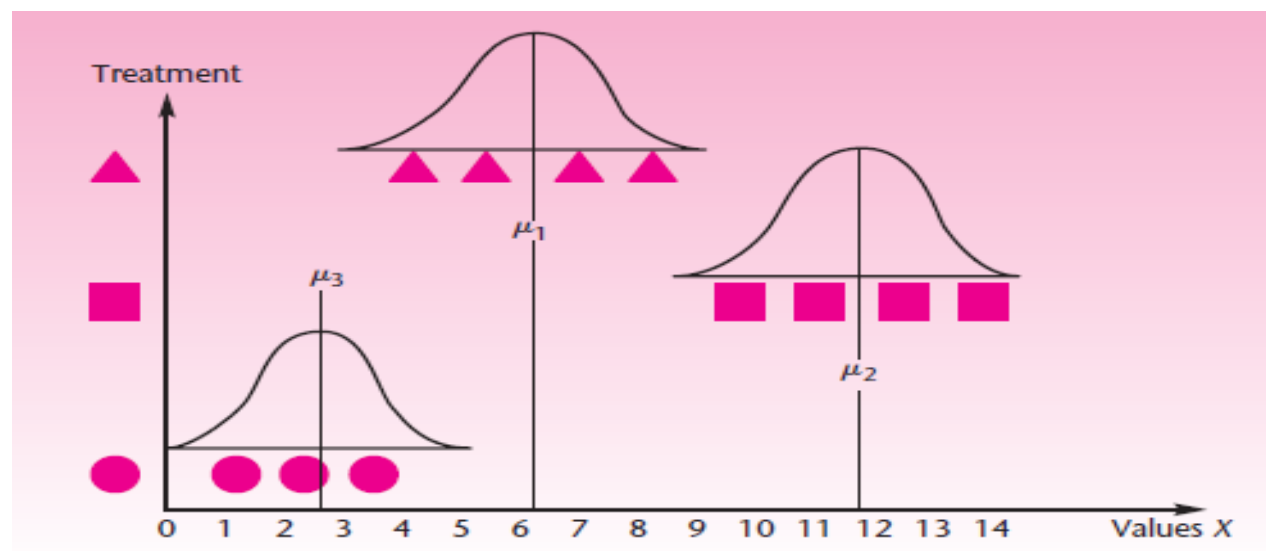
The third data point (person) in the group of 21 people who consumed Brazilian coffee is denoted by $X_{13} =$ (that is, $i = 1$ denotes treatment 1 and $j = 3$ denotes the third point in that sample). We will now define the main principle behind the analysis of variance.

If the r population means are different (i.e., at least two of the population means are *not* equal), then the variation of the data points about their respective sample means $\bar{X}_i$ is likely to be *small* when compared with the variation of the r sample means about the grand mean $\bar{\bar{X}}$.

Table 7.1 Data and the Various Sample Means for Triangles, Squares, and Circles

Treatment i	Sample Point j	Value x_{ij}
$i = 1$ Triangle	1	4
Triangle	2	5
Triangle	3	7
Triangle	4	8
Mean of triangles		6
$i = 2$ Square	1	10
Square	2	11
Square	3	12
Square	4	13
Mean of squares		11.5
$i = 3$ Circle	1	1
Circle	2	2
Circle	3	3
Mean of circles		2
Grand mean of all data points		6.909

Figure 7.3: Samples of Triangles, Squares, and Circles and their respective populations (the three populations are normal with equal variance but with different means)



The Sum-of-Squares Principle

When the population means are different, the error deviations in the data are small when compared with the treatment deviations. We made general statements about the average error being small when compared with the average treatment deviation. The error deviations measure how close the data within each group are to their respective group means. The treatment

deviations measure the distances between the various groups. It therefore seems intuitively plausible that when these two kinds of deviations are of about equal magnitude, the population means are about equal. Why? Because when the average error is about equal to the average treatment deviation, the treatment deviation may itself be viewed as just another error. That is, the treatment deviation in this case is due to pure chance rather than to any real differences among the population means. In other words, when the average **t** is of the same magnitude as the average **e**, both are estimates of the internal variation **within** the data and carry no information about a difference between any two groups-about a difference in population means.

We define the **total deviation** of a data point X_{ij} (denoted by Tot_{ij}) as the deviation of the data point from the grand mean: $Tot_{ij} = X_{ij} - \bar{\bar{X}}$

For any data point X_{ij} ,

$$Tot = t + e \quad \text{Or} \quad \text{Total deviation} = \text{Treatment deviation} + \text{Error deviation}$$

$$Tot_{ij}^2 = (x_{ij} - \bar{\bar{x}})^2$$

$$\text{Sum square total (SST)} = \sum_{j=1}^k \sum_{i=1}^{n_i} Tot_{ij}^2 = \sum_{j=1}^r \sum_{i=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2$$

$$\text{Sum square treatment (SSTR)} = \sum_{i=1}^k n_i Tr_{ij}^2 = \sum_{j=1}^{n_i} n_i (\bar{X}_i - \bar{\bar{X}})^2$$

$$\text{Sum square error (SSE)} = \sum_{j=1}^k \sum_{i=1}^{n_i} e_{ij}^2 = \sum_{j=1}^r \sum_{i=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

The Sum-of-Squares Principle

The sum-of-squares total (SST) is the sum of the two terms: the sum of squares for treatment (SSTR) and the sum of squares for error (SSE).

$$SST = SSTR + SSE$$

$$\sum_{j=1}^r \sum_{i=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2 = \sum_{j=1}^{n_i} n_i (\bar{X}_i - \bar{\bar{X}})^2 + \sum_{j=1}^r \sum_{i=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

Consider the total sum of squares, SST. In computing this sum of squares, we use the entire data set and information about *one* quantity computed from the data: the grand mean (because, by definition, SST is the sum of the squared deviations of all data points from the grand mean). Since we have a total of n data points and one restriction,

The number of degrees of freedom associated with SST is $n - 1$.

The sum of squares for treatment SSTR is computed from the deviations of r sample means from the grand mean. The r sample means are considered r independent data points, and the grand mean (which can be considered as having been computed from the r sample means) thus reduces the degrees of freedom by 1.

The number of degrees of freedom associated with SSTR is $r - 1$.

The sum of squares for error SSE is computed from the deviations of a total of n data points ($n = n_1 + n_2 + \dots + n_r$) from r different sample means. Since each of the sample means acts as a restriction on the data set, the degrees of freedom for error are $n - r$. This can be seen another way: There are r groups with n_i data points in group i . Thus, each group, with its own sample mean acting as a restriction, has degrees of freedom equal to $n_i - 1$. The total number of degrees of freedom for error is the sum of the degrees of freedom in the r groups:

$$df = (n_1 - 1) + (n_2 - 1) + \dots + (n_r - 1) = n - r.$$

The number of degrees of freedom associated with SSE is $n - r$.

An important principle in analysis of variance is that the degrees of freedom of the three components are *additive* in the same way that the sums of squares are additive.

$$df(\text{total}) = df(\text{treatment}) + df(\text{error})$$

This can easily be verified by noting the following: $n - 1 = (r - 1) + (n - r)$ the r drops out. We are now ready to compute the average squared deviation due to treatment and the average squared deviation due to error.

Mean Squares (MS)

In finding the average squared deviations due to treatment and to error, we divide each sum of squares by its degrees of freedom. We call the two resulting averages **mean square treatment (MSTR)** and **mean square error (MSE)**, respectively.

$$MSTR = \frac{SSTR}{r-1} \quad \quad \quad MSE = \frac{SSE}{n-r}$$

The Expected Values of the Statistics MSTR and MSE under the Null Hypothesis

When the null hypothesis of ANOVA is true, all r population means are equal, and in this case there are no treatment effects. In such a case, the average squared deviation due to “treatment” is just another realization of an average squared error. In terms of the expected values of the two mean squares, we have and

$$E(MSE) = \sigma^2 \quad \quad \quad E(MSTR) = \sigma^2 + \frac{\sum_{j=1}^{n_i} n_i (\mu_i - \mu)^2}{r-1}$$

Where μ_i is the mean of population i and μ is the combined mean of all r populations.

When the null hypothesis of ANOVA is true and all r population means are equal, MSTR and MSE are two independent, unbiased estimators of the common population variance σ^2 .

If, on the other hand, the null hypothesis is not true and differences do exist among the r population means, then MSTR will tend to be larger than MSE. This happens because, when not all population means are equal.

The F Statistic

The preceding discussion suggests that the ratio of MSTR to MSE is a good indicator of whether the r population means are equal. If the r population means are equal, then MSTR/MSE would tend to be close to 1. Remember that both MSTR and MSE are sample statistics derived from our data. As such, MSTR and MSE will have some randomness associated with them, and they are not likely to exactly equal their expected values. Thus, when the null hypothesis is true, MSTR/MSE will vary around to 1. When not all the r population means are equal, the ratio MSTR/MSE will tend to be greater than 1 because the expected value of MSTR will be larger than the expected value of MSE. How is large “large enough” for to reject the null hypothesis?

This is where statistical inferences want to determine whether the difference between our observed value of MSTR/MSE and the number 1 is due just to chance variation, or whether MSTR/MSE is significantly greater than 1 implying that not all the population means are equal. We will make the determination with the aid of the F distribution.

Under the assumptions of ANOVA, the ratio MSTR/MSE possesses an F distribution with $r - 1$ degrees of freedom for the numerator and $n - r$ degrees of freedom for the denominator when the null hypothesis is true.

F statistic computations required for arriving at the value of the test statistic. The test statistic in

analysis of variance is
$$F_{(r-1, n-r)} = \frac{\text{MSTR}}{\text{MSE}} = \frac{\frac{\text{SSTR}}{r-1}}{\frac{\text{SSE}}{n-r}}$$

7.4. ANOVA Table and Example

Test the hypothesis of three population mean ANOVA hypothesis whether the means of the three populations are equal or not at $\alpha = 0.01$ (for a right-tailed test).

Table 7.2: Computations for Triangles, Squares and Circles

Treatment i	j	Value x_{ij}	$x_{ij} - \bar{x}_i$	$(x_{ij} - \bar{x}_i)^2$
Triangle	1	4	$4 - 6 = -2$	$(-2)^2 = 4$
Triangle	2	5	$5 - 6 = -1$	$(-1)^2 = 1$
Triangle	3	7	$7 - 6 = 1$	$(1)^2 = 1$
Triangle	4	8	$8 - 6 = 2$	$(2)^2 = 4$
Square	1	10	$10 - 11.5 = -1.5$	$(-1.5)^2 = 2.25$
Square	2	11	$11 - 11.5 = -0.5$	$(-0.5)^2 = 0.25$
Square	3	12	$12 - 11.5 = 0.5$	$(0.5)^2 = 0.25$
Square	4	13	$13 - 11.5 = 1.5$	$(1.5)^2 = 2.25$
Circle	1	1	$1 - 2 = -1$	$(-1)^2 = 1$
Circle	2	2	$2 - 2 = 0$	$(0)^2 = 0$
Circle	3	3	$3 - 2 = 1$	$(1)^2 = 1$
			Sum = 0	Sum = 17

$$SSE = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = 17.00$$

$$SSTR = \sum_{i=1}^r n_i (\bar{x}_i - \bar{\bar{x}})^2 = 4(6 - 6.909)^2 + 4(11.5 - 6.909)^2 + 3(2 - 6.909)^2 = 159.9$$

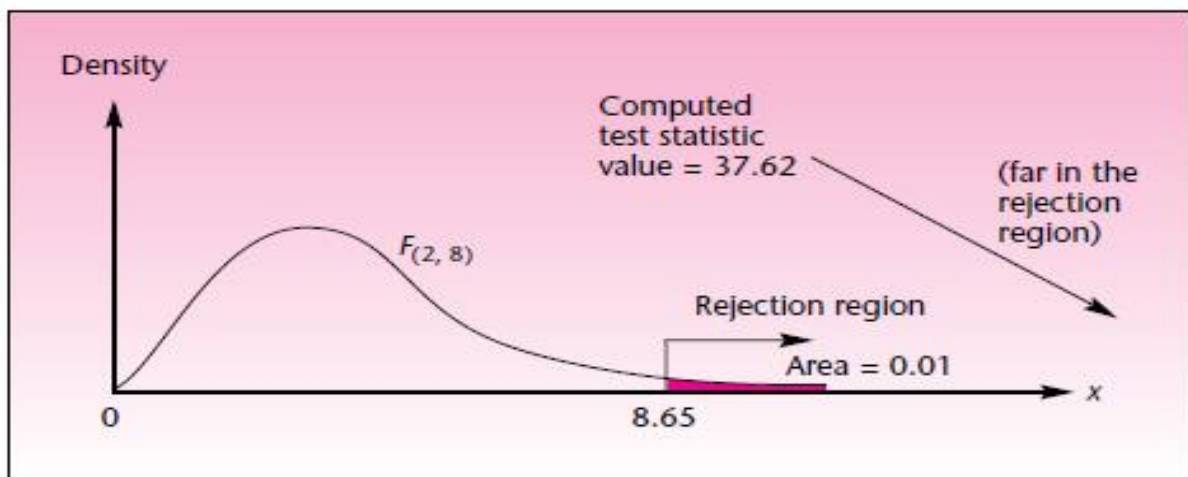
$$MSTR = \frac{SSTR}{r - 1} = \frac{159.9}{2} = 79.95$$

$$MSE = \frac{SSE}{n - r} = \frac{17}{8} = 2.125$$

$$F_{(2, 8)} = \frac{MSTR}{MSE} = \frac{79.95}{2.125} = 37.62$$

$$F_{(3-1, 11-3)} = F(2, 8) = 8.65$$

Figure 7.4: Rejecting the Null Hypothesis in the Triangles, Squares, and Circles Example



The critical value is 8.65. We can therefore reject the null hypothesis. Since 37.62 is much greater than 8.65, the p-value is much smaller than 0.01. This is shown in Figure 7.4.

An essential tool for reporting the results of an analysis of variance is the ANOVA table. An ANOVA table lists the sources of variation: treatment, error, and total. (In the two-factor ANOVA, which we will see in later sections, there will be more sources of variation.) The ANOVA table lists the sums of squares, the degrees of freedom, the mean squares, and the F ratio. The table format simplifies the analysis and the interpretation of the results. The structure of the ANOVA table is based on the fact that both the sums of squares and the degrees of freedom are additive. We will now present an ANOVA table for the triangles, squares, and circles example. Table 7.2 shows the results computed above.

Table 7.3: ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Treatment	$SSTR = 159.9$	$r - 1 = 2$	$MSTR = \frac{SSTR}{r - 1}$ $= 79.95$	$F = \frac{MSTR}{MSE}$ $= 37.62$
Error	$SSE = 17.0$	$n - r = 8$	$MSE = \frac{SSE}{n - r}$ $= 2.125$	
Total	$SST = 176.9$	$n - 1 = 10$		

The last entry in the table is the main objective of our analysis: the F ratio, which is computed as the ratio of the two entries in the previous column. No other entries appear in the last column.

EXAMPLE 2

Club Med has more than 30 major resorts worldwide, from Tahiti to Switzerland. Many of the beach resorts are in the Caribbean, and at one point the club wanted to test whether the resorts on Guadeloupe, Martinique, Eleuthera, Paradise Island, and St. Lucia were all equally well liked by vacationing club members. The analysis was to be based on a survey questionnaire filled out by a random sample of 40 respondents in each of the resorts. From every returned questionnaire, a general satisfaction score, on a scale of 0 to 100, was computed. Analysis of the survey results yielded the statistics given in Table 7.4.

The results were computed from the responses by using a computer program that calculated the sums of squared deviations from the sample means and from the grand mean. Given the values of SST and SSE, construct an ANOVA table and conduct the hypothesis test.

Solution:

Let us first construct an ANOVA table and fill in the information we have: $SST = 112,564$, $SSE = 98,356$, $n = 200$ and $r = 5$. This has been done in Table 7.5. We now compute $SSTR$ as the difference between SST and SSE and enter it in the appropriate place in the table. We then divide $SSTR$ and SSE by their respective degrees of freedom to give us $MSTR$ and MSE . Finally, we divide $MSTR$ by MSE to give us the F ratio. All these quantities are entered in the ANOVA table. The result is the complete ANOVA table for the study, Table 7.4.

Table 7.4: Club Med Survey Results

Resort i	Mean Response $\bar{X}_i$
1. Guadeloupe	89
2. Martinique	75
3. Eleuthera	73
4. Paradise Island	91
5. St. Lucia	85
$SST = 112,564$	$SSE = 98,356$

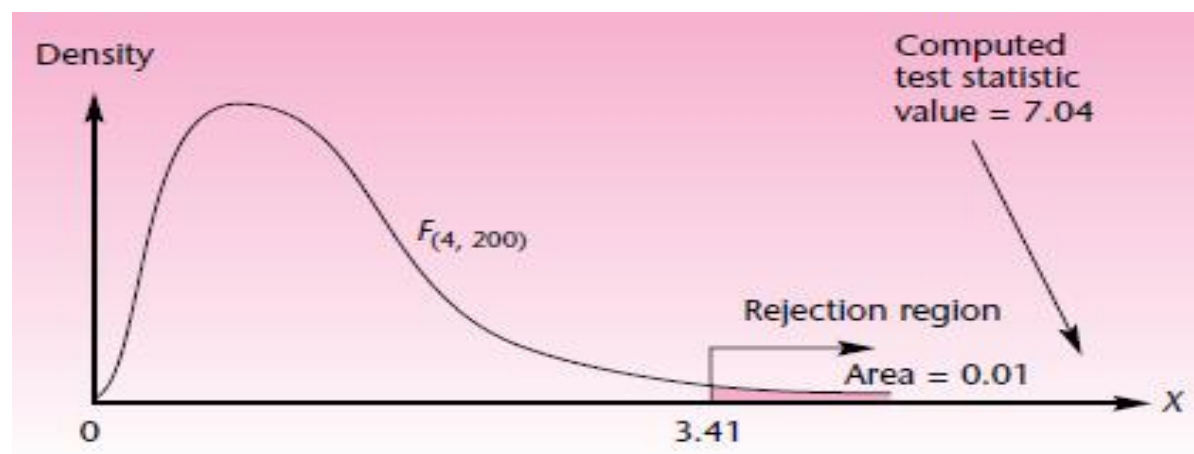
Table 7.5: Preliminary ANOVA Table for Club Med Example

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Treatment	$SSTR =$	$r - 1 = 4$	$MSTR =$	$F =$
Error	$SSE = 98,356$	$n - r = 195$	$MSE =$	
Total	$SST = 112,564$	$n - 1 = 199$		

Table 7.6: ANOVA Table for Club Med Example

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Treatment	$SSTR = 14,208$	$r - 1 = 4$	$MSTR = 3,552$	$F = 7.04$
Error	$SSE = 98,356$	$n - r = 195$	$MSE = 504.4$	
Total	$SST = 112,564$	$n - 1 = 199$		

Figure 7.5: Club Med Test



As shown in Table 7.6, the test statistic value is $F(4, 195) = 7.04$. As often happens, the exact number of degrees of freedom we use the nearest entry, which is the critical point for F with 4 degrees of freedom for the numerator and 200 degrees of freedom for the denominator. The critical point for $\alpha = 0.01$ is 3.41. The test is illustrated in figure 7.5. Since the computed test statistic value falls in the rejection region for $\alpha = 0.01$, we reject the null hypothesis and note that the p-value is smaller than 0.01. We may conclude that, based on the survey results and our assumptions, it is likely that the five resorts studied are not equal in terms of average vacationer satisfaction.

Exercise

Gulfstream Aerospace Company produced three different prototypes as candidates for mass production as the company's newest large-cabin business jet, the *Gulfstream IV*. Each of the three prototypes has slightly different features, which may bring about differences in performance. Therefore, as part of the decision-making process concerning which model to produce, company engineers are interested in determining whether the three proposed models

have about the same average flight range. Each of the models is assigned a random choice of 10 flight routes and departure times, and the flight range on a full standard fuel tank is measured (the planes carry additional fuel on the test flights, to allow them to land safely at certain destination points). Range data for the three prototypes, in nautical miles (measured to the nearest 10 miles), are as follows

Prototype A	Prototype B	Prototype C
4,420	4,230	4,110
4,540	4,220	4,090
4,380	4,100	4,070
4,550	4,300	4,160
4,210	4,420	4,230
4,330	4,110	4,120
4,400	4,230	4,000
4,340	4,280	4,200
4,390	4,090	4,150
4,510	4,320	4,220

Do all three prototypes have the same average range? Construct an ANOVA table, and carry out the test. Explain your results.

Chapter Eight

Regression and Correlation Analysis

Introduction

Linear regression and correlation is studying and measuring the linear relationship among two or more variables. When only two variables are involved, the analysis is referred to as simple correlation and simple linear regression analysis, and when there are more than two variables the term multiple regression and partial correlation is used. **Regression Analysis:** is a statistical technique that can be used to develop a mathematical equation showing how variables are related. **Correlation Analysis:** deals with the measurement of the closeness of the relationship which are described in the regression equation.

We say there is correlation when the two series of items vary together directly or inversely

Correlation Analysis

The measure of the degree of relationship between two continuous variables is known as correlation coefficient. The population correlation coefficient is represented by ρ and its estimator by r . The correlation coefficient r is also called Pearson's correlation coefficient since it was developed by Karl Pearson. r is given as the ratio of the covariance of the variables x and y to the product of the standard deviations of x and y . Symbolically,

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

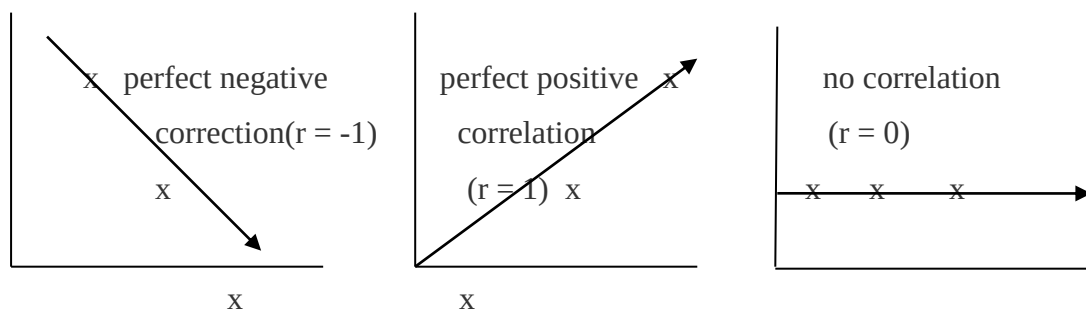
$$r = \frac{\sum \frac{(x - \bar{x})(y - \bar{y})}{n-1}}{\sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}} = \frac{Cor(x, y)}{sd(x).sd(Y)}$$

The numerator is termed as the sum of products of x and y, SP_{xy} . In the denominator, the first term is called the sum of squares of x, SS_x , and the second term is called the sum of squares of y, SS_y . Thus,

$$r = \frac{SP_{xy}}{\sqrt{SS_x} \sqrt{SS_y}}$$

The correlation coefficient is always between -1 and $+1$, i.e., $-1 \leq r \leq +1$.

- $r = -1$ implies perfect negative linear correlation between the variables under consideration.
- $r = +1$ implies perfect positive linear correlation between the variables under consideration.
- $r = 0$ implies there is no linear relationship between the two variables: but there could be a non-linear relationship between them. In other words, when two variables are uncorrelated, $r = 0$, but when $r = 0$, it is not necessarily true that the variables are uncorrelated.



Spearman rank correlation coefficient

The Pearson coefficient of correlation requires precise numerical values (i.e., continuous data) for the variables. Moreover, it is applicable only under the condition that the variables are normally distributed. However, in many instances such numerical measurements may not be possible (for instance, job performance, taste, intelligence, etc.). Moreover, the variables may not come from normally distributed

populations. In such cases, we can compute a non- parametric measure of association that is based on ranks, called the **Spearman rank correlation coefficient**.

The steps involved in computing Spearman rank correlation coefficient are as follows:

- Rank the X's among themselves giving rank 1 to the largest (or smallest) observation, rank 2 to the second largest (or second smallest) observation, and so on.
- Rank the Y's similarly.
- Find $d = \text{rank of } x - \text{rank of } y$ for each pair of observations.
- Find $\sum d^2$ (the sum of squares of the differences between each pair of ranks)

The Spearman rank correlation coefficient (r_s) is given by:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Here n is the number of paired observations

	Advert (X)	Rank(X)	Sales (Y)	Rank(Y)	d	d sq
1	4.69	16	12.23	19	-3	9
2	6.41	24	11.84	16	8	64
3	5.47	22	12.25	20	2	4
4	3.43	11	11.10	13	-2	4
5	4.39	13	10.97	11	2	4
6	2.15	6	8.75	6	0	0
7	1.54	2	7.75	3	-1	1
8	2.67	8	10.50	10	-2	4
9	1.24	1	6.71	1	0	0
10	1.77	4	7.60	2	2	4
11	4.46	14	12.46	22	-8	64
12	1.83	5	8.47	4	1	1
13	5.15	20	12.27	21	-1	1
14	5.25	21	12.57	23	-2	4
15	1.72	3	8.87	7	-4	16
16	3.04	9	11.15	14	-5	25
17	4.92	19	11.86	17	2	4
18	4.85	17	11.07	12	5	25
19	3.13	10	10.38	9	1	1
20	2.29	7	8.71	5	2	4
21	4.90	18	12.07	18	0	0
22	5.75	23	12.74	24	-1	1
23	3.61	12	9.82	8	4	16
24	4.62	15	11.51	15	0	0
				TOTAL	0	256

The Spearman rank correlation coefficient is calculated as:

$$\begin{aligned}
 r_s &= 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \\
 &= 1 - \frac{6(256)}{24[(24)^2 - 1]} = 1 - 0.111304 = 0.888696
 \end{aligned}$$

Simple Linear Regression Analysis

Regression is concerned with bringing out the nature of relationship and using it to know the best approximate value of one variable corresponding to a known value of other variable.

Simple linear regression deals with method of fitting a straight line (regression line) on a sample of data of two variables in terms of equation so that if the value of one variable is given we can predict the value of the other variable.

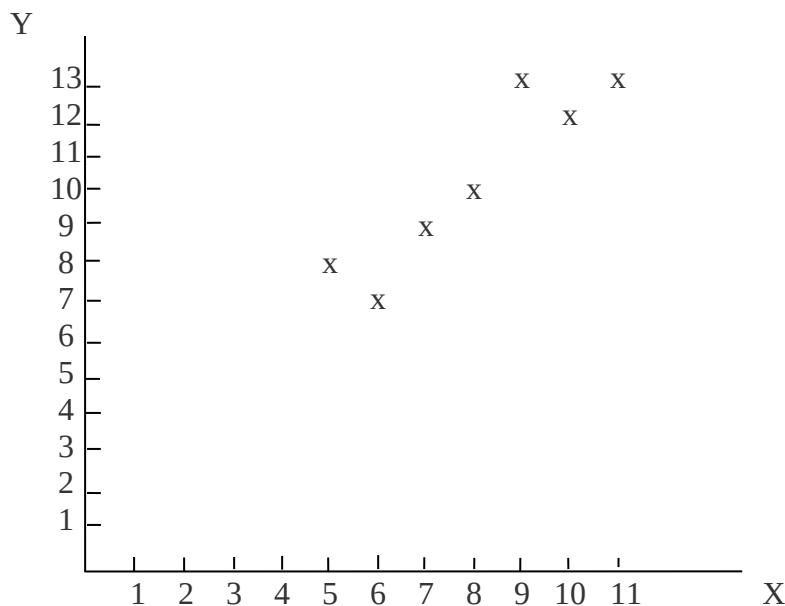
In other words if we have two variables under study one may represent the cause and the other may represent the effect. The variable representing the cause is known as independent (predictor or repressor) variable and it is usually denoted by X. The variable representing the effect is known as dependent (predicted) variable and is usually denoted by Y. Then, if the relationship between the two variables is a straight line, it is known as simple linear regression.

When there are more than two variables and one of them is assumed to be dependent up on the others, the functional relationship between the variables is known as multiple linear regressions.

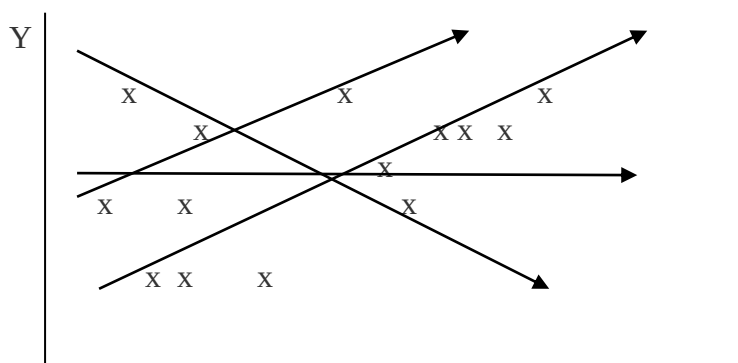
Scatter diagram: is a plot of all ordered pairs (x, y) on the coordinate plane which is necessary to discover whether the relationship b/n two variables indeed best explained by straight line.

Example:

Advertizing budget (X)	5	6	7	8	9	10	11
Profit(Y)	8	7	9	10	13	12	13



So if we draw a line, the regression line is one that passes through almost all or closest to all points in the scatter diagram.



The simple linear regression of Y on X in the population is given by:

$$Y = \alpha + \beta X + \varepsilon$$

Where, α = y-intercept, β = slope of the line or regression coefficient, ε = is the error term

The y-intercept α and the regression coefficient β are the population parameters. We obtain the estimates of α and β from the sample. The estimators of α and β are denoted by a and b, respectively. The fitted regression line is thus,

$$Y_e = a + bx$$

The above algebraic equation is known as a regression line. The method of finding such a relationship is known as fitting regression line. For each observed value of the variable X, we can find out the value of Y. The computed values of Y are known as the expected values of Y and are denoted by Y_e .

The observed values of Y are denoted by Y. The difference between the observed and the expected values $Y - Y_e$, is known as error or residual, and is denoted by e. The residual can be positive, negative or zero.

A best fitting line is one for which the sum of squares of the residuals, $\sum e_i^2$, is minimum. For this purpose the principle called the method of least squares is used.

According to the principle of least squares, one would select a and b such that

$$\sum e_i^2 = \sum (Y - Y_e)^2 \text{ is minimum where } Y_e = a + bx.$$

The simple linear regression of Y on X is specified concisely by the equation $Y \approx \alpha + \beta x$, where α is any constant, and β is a non-zero real number.

This is called the approximated **population regression line**, where α and β are the parameters. To be more specific, α is the mean of the population of Y (say, μ_y) that corresponds to the population mean of X (say, μ_x). β is the slope of the regression line, or the change in Y per unit change in X. Both α and β are called regression coefficients. The relationship in Definition 8.1 assumes that Y mostly depends on X; then, we say that most of the variation in Y is explained due to this relationship with X. Thus, we have to introduce a random component that may account for the unexplained variation in Y, denoted by ε , and rewrite it as

$$Y = \alpha + \beta x + \varepsilon.$$

This is called the **mathematical model** in linear regression. The added term ε (epsilon) is called an error term which accounts for the above approximation. The parameters α and β are unknown and, hence, they must be estimated. Let “a” be an estimate of α and “b” be that of β . Since a and b are obtained by the least squares method (to be explained below), they are called the least square estimates.

For the pair of the i^{th} member of the sample given by a set of paired data:

(x_i, y_i) , $i = 1, 2, \dots, n$, we have, the sample linear regression model,

$$y_i = a + bx_i + e_i, \quad i = 1, 2, \dots, n.$$

The error introduced by each pair (x_i, y_i) is, thus, given by $e_i = y_i - a - bx_i$.

The least square estimates of the regression coefficients, a and b, are obtained by minimizing the sum of the squares of the error terms. That is, minimizing

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2 = f(a, b), \text{ say.}$$

Differentiating $f(a, b)$ with respect to a and b and equating these partial derivatives to zero, we obtain

$$* \quad \frac{\partial f(a, b)}{\partial a} = 2 \sum_{i=1}^n (y_i - a - bx_i)(-1) = 0. \quad \text{Or} \quad \sum_{i=1}^n (y_i - a - bx_i) = 0.$$

From which, using properties of summation and noting that a and b are constants, we get

$$\sum_{i=1}^n y_i = an + b \sum_{i=1}^n x_i.$$

$$* \quad \frac{\partial f(a, b)}{\partial b} = 2 \sum_{i=1}^n (y_i - a - bx_i)(-x_i) = 0. \quad \text{Or} \quad \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2.$$

Then, we have the following system of two linear equations called normal equations:

Result of the **normal equations** in regression are given by

$$\sum_{i=1}^n y_i = an + b \sum_{i=1}^n x_i \quad \text{(a)}$$

$$\sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 \quad \text{(b)}$$

This summary result provides two equations with two unknown variables (a, & b), hence we can solve them simultaneously to get unique values for a and b, which is the objective of the least squares method. However, to make further analysis, these normal equations can be solved using determinants or the method of elimination (the limits of summation are omitted for simplicity). To use the elimination method, solving the first of the two normal equations (a) gives as

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

Now, replace this in the second equation (2.1b), to get

$$\begin{aligned}\sum xy &= \sum x(\bar{y} - b\bar{x}) + b\sum x^2 = \bar{y}\sum x - b\bar{x}\sum x + b\sum x^2 \\ &= n\bar{x}\bar{y} + b\left(\sum x^2 - n\bar{x}^2\right), \text{ since } \sum x = n\bar{x}.\end{aligned}$$

Or, $b\left(\sum x^2 - n\bar{x}^2\right) = \sum xy - n\bar{x}\bar{y}$. Solving this for b yields:

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

Solving these normal equations simultaneously we can get the values of **a** and **b** as follows:

$$\begin{aligned}b &= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \\ a &= \bar{y} - b\bar{x}\end{aligned}$$

Regression analysis is useful in predicting the value of one variable from the given values of another variable.

To further simplify the formula for b, recall the sample variance of X:

$$S_x^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{\sum x^2 - n\bar{x}^2}{n-1}, \text{ (unit 2), from which,}$$

$$\sum (x - \bar{x})^2 = (n-1)S_x^2 = \sum x^2 - n\bar{x}^2$$

Now, let us define the sample covariance:

Definition: - The sample covariance is defined by

$$S_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1} = \frac{\sum xy - n\bar{x}\bar{y}}{n-1}.$$

So that the numerator is

$$\sum xy - n\bar{x}\bar{y} = (n-1)S_{xy} = \sum (x - \bar{x})(y - \bar{y}).$$

Thus, 8.3 is further condensed as: $b = \frac{S_{xy}}{S_x^2}$.

Result given the sample data $(x_i, y_i), i = 1, 2, \dots, n$, the coefficients of the least square line,

$$y = a + bx, \text{ are } b = \frac{S_{xy}}{S_x^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

Where S_x^2 = the sample variance of X, and S_{xy} = the sample covariance.

Having obtained a and b, the least square line, $y = a + bx$, is referred to as the “best line” in the sense that it provides the best approximation to the paired data. To verify that a and b minimize $\sum e^2$, we can use the second derivative tests for both:

$$\frac{\partial^2 f(a,b)}{\partial^2 a} = 2n > 0, \text{ and } \frac{\partial^2 f(a,b)}{\partial^2 b} = 2\sum x^2 > 0$$

The results obtained allow us to predict (estimate) the value of Y whenever a possible value of X is given, using the **prediction equation**: - $y_{\text{est}} = a + bx$

Here, Y_{est} stands for the predicted value of Y for the given value of X.

Notation: To indicate that $y = a + bx$ is the regression of Y on X, we rewrite it as:

$$y = a_{yx} + b_{yx}x$$

Example: A researcher wants to find out if there is any relationship b/n height of the son and his father. He took random sample 6 fathers and their sons. The height in inch is given in the table bellow (i) Find the regression line of Y on X

(ii) What would be the height of the son if his father's height is 70 inch?

Height of father (X)	63	65	66	67	67	68
Height of the son (Y)	66	88	65	67	69	70

Solution: $\sum X = 396$, $\sum Y = 425$, $\sum X^2 = 26152$, $\sum XY = 26740$, $\sum Y^2 = 27355$

$$(i) \quad b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{6(26740) - (396)(425)}{6(26152) - (396)^2} = 0.625$$

$$a = \bar{y} - b\bar{x} = \frac{\sum Y - b\sum X}{n} = \frac{425 - (0.625)(396)}{6} = 29.58$$

$$\Rightarrow Y = 29.58 + 0.625X$$

(ii) If $X=70$, then

$$Y = 29.58 + 0.625(70) = 73.33, \text{ thus the height of the son is 70 inch}$$

Standard Error of estimates: measures the average amount by which the estimated Y_e values depart from the corresponding observed Y values (dispersion of observed values around the line of regression Y on X)

$$S_{x,y} = \sqrt{\sum \frac{(y_i - y_{ei})^2}{n-2}}, \text{ where } Y_e = \alpha + \beta X + \varepsilon \text{ and}$$

Y_i is observed (actual) value of y

Example: given the observation (2, 2), (4, 5), (6, 4) and (8, 7), we can get the regression line $Y_e = 1 + 0.7X$. Find the standard error of the estimates of the regression line.

Solution:

$$Y_e = 1 + 0.7X_i, i = 1, 2, 3, 4$$

$$\text{Then } Y_{e1} = 1 + 0.7(2) \quad Y_{e3} = 1 + 0.7(6) = 5.2$$

$$= 1 + 0.7(2) = 2.4 \quad Y_{e4} = 1 + 0.7(8) = 6.6$$

$$Y_{e2} = 1 + 0.7(4) = 3.8$$

$$\Rightarrow S_{x,y} = \sqrt{\sum \frac{(y_i - y_{ei})^2}{n-2}} = \sqrt{\frac{1}{2}[(2-2.4)^2 + \dots + (7-6.6)^2]} = 1.26$$

Exercise: - The following table shows the number of hours (X) a learner spent studying and the marks (Y) each learner received in an examination:

x	8	5	11	13	10	6	18	15	2	9
y	65	44	79	72	70	54	90	85	33	56

Assuming simple linear relationship between X and Y,

a) Find the estimated regression equation of Y on X.

b) Give the predicted value of Y for X = 12.

Answer

a) $b = 3.596$, and $a = 29.92$. Hence, the equation is $y_{\text{est}} = 29.92 + 3.596x$.

b) When $X = 12$, $\hat{y}_{12} = 29.92 + 3.596(12) = 73.1$.