## 11.7 MULTIPLE LINEAR REGRESSION

In the majority of applications, the response of an experiment can be predicted more adequately not on the basis of a single independent input variable but on a collection of such variables. A typical situation is one in which there are a set of, say, *k* input variables and the response *Y* is related to them by the relation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e \qquad (11.70)$$

where $x_j$, $j = 1, 2, \ldots, k$ is the level of the *j*th input variable and *e* is a *random error* that we shall assume is normally distributed with mean 0 and constant variance $\sigma^2$. The parameters $\beta_0, \beta_1, \ldots, \beta_k$ and $\sigma^2$ are assumed to be unknown and must be estimated from the data, which are the values of $Y_1, Y_2, \ldots, Y_n$ where $Y_i$ is the response level corresponding to the *k* input levels $x_{i1}, x_{i2}, \ldots, x_{ik}$. The $Y_i$ are related to these input levels through

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \qquad (11.71)$$

If we let $B_0, B_1, \ldots, B_k$ denote estimators of $\beta_0, \beta_1, \ldots, \beta_k$, then the *sum of the squared* differences between the $Y_i$ and their estimated expected values is

$$\sum_{i=1}^{n} \left( Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \ldots - B_k x_{ik} \right)^2 \qquad (11.72)$$

The *least squares estimators* are those values of $B_0, B_1, \ldots, B_k$ that minimize the foregoing. To determine the least squares estimators, we repeatedly take partial derivatives of the preceding sum of squares first with respect to $B_0$, then to $B_1$, . . . , then to $B_k$. On equating these $k+1$ equations to 0, we obtain the following set of equations:

$$\sum_{i=1}^{n} \left( Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \ldots - B_k x_{ik} \right) = 0 \qquad (11.73)$$

$$\sum_{i=1}^{n} x_{i1} \left( Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \ldots - B_k x_{ik} \right) = 0$$

$$\sum_{i=1}^{n} x_{i2} \left( Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \ldots - B_k x_{ik} \right) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik} \left( Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \ldots - B_k x_{ik} \right) = 0$$

Rewriting these equations yields that the least squares estimators $B_0, B_1, \ldots, B_k$ satisfy the following set of linear equations, called the *normal equations*:

$$
\begin{aligned}
\sum_{i=1}^{n} Y_i &= nB_0 + B_1 \sum_{i=1}^{n} x_{i1} + B_2 \sum_{i=1}^{n} x_{i2} + \ldots + B_k \sum_{i=1}^{n} x_{ik} \qquad (11.74)\\
\sum_{i=1}^{n} x_{i1} Y_i &= B_0 \sum_{i=1}^{n} x_{i1} + B_1 \sum_{i=1}^{n} x_{i1}^2 + B_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots + B_k \sum_{i=1}^{n} x_{i1} x_{ik}\\
\sum_{i=1}^{n} x_{i2} Y_i &= B_0 \sum_{i=1}^{n} x_{i2} + B_1 \sum_{i=1}^{n} x_{i2} x_{i1} + B_2 \sum_{i=1}^{n} x_{i2}^2 + \ldots + B_k \sum_{i=1}^{n} x_{i2} x_{ik}\\
&\vdots\\
\sum_{i=1}^{n} x_{ik} Y_i &= B_0 \sum_{i=1}^{n} x_{ik} + B_1 \sum_{i=1}^{n} x_{ik} x_{i1} + B_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \ldots + B_k \sum_{i=1}^{n} x_{ik}^2
\end{aligned}
$$

Before solving the normal equations, it is convenient to introduce *matrix notation*. If we let

$$
\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad (11.75)
$$

$$
\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}
$$

Then $\mathbf{Y}$ is an $n \times 1$, $\mathbf{X}$ an $n \times (k+1)$, $\boldsymbol{\beta}$ an $(k+1) \times 1$ and $\mathbf{e}$ is an $n \times 1$ matrix.

*The multiple regression model* can now be written as

$$
\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad (11.76)
$$

In addition, if we let

$$
\mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix} \qquad (11.77)
$$

be the matrix of least squares estimators, then the normal Equations (11.74) can be written as

$$
\mathbf{X}^T \mathbf{X} \mathbf{B} = \mathbf{X}^T \mathbf{Y} \qquad (11.78)
$$

where $\mathbf{X}^T$ is the transpose of $\mathbf{X}$.

To see that Equation (11.78) is equivalent to the normal Equations (11.74), note that

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}$$

and

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_{i1}Y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}Y_i \end{bmatrix}$$

It is now easy to see that the matrix equation

$$\mathbf{X}^T\mathbf{X}\mathbf{B} = \mathbf{X}^T\mathbf{Y} \qquad (11.79)$$

is equivalent to the set of normal Equations (11.74). Assuming that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exists, which is usually the case, we obtain, upon multiplying it by both sides of the foregoing, that the least squares estimators are given by

$$\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \qquad (11.80)$$

Program R computes the least squares estimates.

**Example 1** The data in Table relate the suicide rate to the population size and the divorce rate at eight different locations.

| Location | Population in Thousands | Divorce Rate per 100000 | Suicide Rate per 100000 |
|---|---|---|---|
| Akron | 679 | 30.4 | 11.6 |
| Anaheim | 1420 | 34.1 | 16.1 |
| Buffalo | 1349 | 17.2 | 9.3 |
| Austin | 296 | 26.8 | 9.1 |
| Chicago | 6975 | 29.1 | 8.4 |
| Columbia | 323 | 18.7 | 7.7 |
| Detroit | 4200 | 32.6 | 11.3 |
| Gary | 633 | 32.5 | 8.4 |

Fit a multiple linear regression model to these data. That is, fit a model of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

where $Y$ is the suicide rate, $x_1$ is the population, and $x_2$ is the divorce rate.

We run R, and results are shown.

```
> Y<-c(11.6,16.1,9.3,9.1,8.4,7.7,11.3,8.4)
> x1<-c(679,1410,1349,296,6975,323,4200,633)
> x2<-c(30.4,34.1,17.2,26.8,29.1,18.7,32.6,32.5)
> model1<-lm(Y~x1+x2)
> summary(model1)

Call:
lm(formula = Y ~ x1 + x2)

Residuals:
      1       2       3       4       5       6       7       8 
 0.3269  4.0431  1.6386 -1.3288 -0.9678 -0.6081  0.3285 -3.4326 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.5079307  4.3566394   0.805    0.457
x1          -0.0002487  0.0004287  -0.580    0.587
x2           0.2609870  0.1587762   1.644    0.161

Residual standard error: 2.612 on 5 degrees of freedom
Multiple R-squared:  0.3531,	Adjusted R-squared:  0.09441 
F-statistic: 1.365 on 2 and 5 DF,  p-value: 0.3365
```

Thus the estimated regression line is

$$Y = 3.507 - 0.0002 x_1 + 0.2609 x_2$$

The value of $\beta_1$ indicates that the population does not play a major role in predicting the suicide rate (at least when the divorce rate is also given). Perhaps the population density, rather than the actual population, would have been more useful.

```
> model2<-lm(Y~x2)
> summary(model2)

Call:
lm(formula = Y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9835 -1.2412 -0.2565  0.9241  4.3364 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6642     4.1009   0.894    0.406
x2            0.2375     0.1448   1.640    0.152

Residual standard error: 2.463 on 6 degrees of freedom
Multiple R-squared:  0.3096,	Adjusted R-squared:  0.1945 
F-statistic: 2.691 on 1 and 6 DF,  p-value: 0.1521
```

$$Y = 3.6642 + 0.2375x_2$$

**EXAMPLE 2** A recently completed study attempted to relate job satisfaction to income (in 1,000S) and seniority for a random sample of 9 municipal workers. The job satisfaction value given for each worker is his or her own assessment of such, with a score of 1 being the lowest and 10 being the highest. The following data resulted.

| Yearly Income | Years on the Job | Job Satisfaction |
|---|---|---|
| 52 | 8 | 5.6 |
| 47 | 4 | 6.3 |
| 59 | 12 | 6.8 |
| 53 | 9 | 6.7 |
| 61 | 16 | 7.0 |
| 64 | 14 | 7.7 |
| 58 | 10 | 7.0 |
| 67 | 15 | 8.0 |
| 71 | 22 | 7.8 |

(a) Estimate the regression parameters.

(b) What qualitative conclusions can you draw about how job satisfaction changes when income remains fixed and the number of years of service increases?

(c) Predict the job satisfaction of an employee who has spent 5 years on the job and earns a yearly salary of $56,000.

```
> JobS<-c(5.6,6.3,6.8,6.7,7.0,7.7,7.0,8.0,7.8)
> YIncome<-c(52,47,59,53,61,64,58,67,71)
> yJob<-c(8,4,12,9,16,14,10,15,22)
> model<-lm(JobS~YIncome+yJob)
> summary(model)

Call:
lm(formula = JobS ~ YIncome + yJob)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.71365 -0.05968  0.04694  0.13145  0.34478 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1.20499    2.35109  -0.513    0.627  
YIncome      0.16195    0.05512   2.938    0.026 *
yJob        -0.11283    0.08001  -1.410    0.208  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3692 on 6 degrees of freedom
Multiple R-squared:  0.8263,	Adjusted R-squared:  0.7684 
F-statistic: 14.27 on 2 and 6 DF,  p-value: 0.005239
```

$$\text{JobS} = -1.2049 + 0.1619 * \text{YIncome} - 0.1128 * \text{yJob}$$

```
> mmodel<-lm(JobS~YIncome)
> summary(mmodel)

Call:
lm(formula = JobS ~ YIncome)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7627 -0.1791  0.1090  0.2806  0.3775 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  1.78396    1.08695   1.641  0.14475   
YIncome      0.08805    0.01825   4.824  0.00191 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3944 on 7 degrees of freedom
Multiple R-squared:  0.7688,	Adjusted R-squared:  0.7357 
F-statistic: 23.27 on 1 and 7 DF,  p-value: 0.001913
```

$$\text{JobS} = 1.7839 + 0.088 * \text{YIncome}$$

**Chapter 12** ANALYSIS OF VARIANCE

Bibliography: Sheldon Ross

## 12.1 INTRODUCTION

A large company is considering purchasing one of four different computer packages designed to teach a new programming language. Some influential people within this company have claimed that these packages are basically interchangeable in that the one chosen will have little effect on the final competence of its user.

To test this hypothesis the company has decided to choose 160 of its engineers, and divide them into 4 groups of size 40. Each member in group *i* will then be given teaching package *i*, $i = 1, 2, 3, 4$, to learn the new language. When all the engineers complete their study, a comprehensive exam will be given. The company then wants to use the results of this examination to determine whether the computer teaching packages are really interchangeable or not.

We clearly desire to be able to conclude that the teaching packages are indeed interchangeable when the average test scores in all the groups are similar and to conclude that the packages are essentially different when there is a large variation among these average test scores.

The *method of division* of the 160 engineers into 4 groups is of vital importance. For example, suppose that the members of the first group score significantly higher than those of the other groups. What can we conclude from this? Specifically, is this result due to teaching package 1 being a superior teaching package, or is it due to the fact that the engineers in group 1 are just better learners? It is essential that we divide the 160 engineers into the 4 groups in such away to make it extremely unlikely that one of these groups is inherently superior. The time tested method for doing this is to divide the engineers into 4 groups in a *completely random fashion*. That is, we should do it in such a way so that all possible divisions are equally likely; for in this case, it would be very unlikely that any one group would be significantly superior to any other group. So let us suppose that the division of the engineers was indeed done “at random.” (Whereas it is not at all obvious how this can be accomplished, one efficient procedure is to start by arbitrarily numbering the 160 engineers. Then generate a random permutation of the integers $1, 2, \ldots, 160$ and put the engineers whose numbers are among the first 40 of the permutation into group 1, those whose numbers are among the 41st through the 80th of the permutation into group 2, and so on.)

It is reasonable to suppose that *the test score of a given individual should be approximately a normal random variable having parameters that depend on the package from which he was taught*. Also, it is probably reasonable to suppose that whereas the *average* test score of an engineer will depend on the teaching package she was exposed to, the *variability* in the test score will result from the inherent variation of 160 different people and not from the particular package used. Thus, if we let $X_{ij}$, $i=1,\dots,4$, $j=1,\dots,40$, denote the test score of the *j*th engineer in group *i*, a reasonable model might be to suppose that the $X_{ij}$ are independent random variables with $X_{ij}$ having a normal distribution with unknown mean $\mu_i$ and unknown variance $\sigma^2$. The hypothesis that the teaching packages are interchangeable is then equivalent to the hypothesis that $\mu_1=\mu_2=\mu_3=\mu_4$ .

In this course, we present a technique that can be used to test such a hypothesis. It is known as the *analysis of variance*. This technique we use when all the *explanatory variables are categorical*. The explanatory variables are call *factors*, and each factor has two or more *levels*.

## 12.2. An Overview

Tests concerning multiple population means will be considered. The difference between the populations results from different levels of a factor. In next Section, we suppose that we have been provided samples of size *n* from *m* distinct populations and that we want to use these data to test the hypothesis that the *m* population means are equal. Since *the mean of a random variable depends only on a single factor*, this scenario is said to constitute a *one way analysis of variance*.

In some cases there *are two factors that determine the mean value of a variable*. Such a model is called a *two-way analysis of variance*.

In all of the models considered in this course, we assume that the data are normally distributed with the same (although unknown) variance $\sigma^2$. The idea of analysis of variance (ANOVA) is *to compare two or more means by comparing variances*. The analysis of variance approach for testing a null hypothesis $H_0$ concerning the population means is based on *deriving two estimators of the common variance* $\sigma^2$. The *first* estimator is a valid estimator of $\sigma^2$ whether the null hypothesis is true or not, while the *second* one is a valid estimator only when $H_0$ is true. In addition, when $H_0$ is not true this latter estimator will tend to exceed $\sigma^2$. The test will be to compare the values of these two estimators, and to reject $H_0$ when the ratio of the second estimator to the first one is sufficiently large. In other words, since the two estimators should be close to each other when $H_0$ is true (because they both estimate $\sigma^2$ in this case) whereas the second estimator should tend to be larger than the first when $H_0$ is not true, it is natural to reject $H_0$ when the second estimator is significantly larger than the first.

We will obtain estimators of the variance $\sigma^2$ by making use of certain facts concerning *chi-square* random variables. Suppose that $X_1, X_2, \ldots, X_N$ are independent normal random variables having possibly different means but a common variance $\sigma^2$, and let $\mu_i = E[X_i]$, $i = 1, 2, \ldots, N$. Since the variables

$$Z_i = (X_i - \mu_i)/\sigma \ , \qquad i = 1, 2, \ldots, N \tag{12.1}$$

have standard normal distributions, it follows from the definition of a chi-square random variable that

$$\sum_{i=1}^{N} Z_i^2 = \sum_{i=1}^{N} \frac{(X_i - \mu_i)^2}{\sigma^2} \tag{12.2}$$

is a chi-square random variable with *N* degrees of freedom. Now, suppose that each of the values $\mu_i$, $i = 1, 2, \ldots, N$, can be expressed as a linear function of a fixed set of *k* unknown parameters. Suppose, further, that we can determine estimators of these *k* parameters, which thus gives us estimators of the mean values $\mu_i$. If we let $\hat{\mu}_i$ denote the resulting estimator of $\mu_i$, $i = 1, 2, \ldots, N$, then it can be shown that the quantity

$$\sum_{i=1}^{N} \frac{(X_i - \hat{\mu}_i)^2}{\sigma^2} \tag{12.3}$$

will have a chi-square distribution with $N - k$ degrees of freedom.

In other words, we start with

$$\sum_{i=1}^{N} \frac{(X_i - E[X_i])^2}{\sigma^2} \tag{12.4}$$

which is a chi-square random variable with *N* degrees of freedom. If we now write each $E[X_i]$ as a linear function of *k* parameters and then replace each of these parameters by its estimator, then the resulting expression remains chi-square but with a degree of freedom that is reduced by 1 for each parameter that is replaced by its estimator.

For an illustration of the preceding, consider the case where all the means are known to be equal; that is,

$$E\left[X_i\right] = \mu \ , \quad i = 1, 2, \ldots, N \tag{12.5}$$

Thus $k = 1$, because there is only one parameter that needs to be estimated. Substituting $\bar{X}$, the estimator of the common mean μ, for $\mu_i$ in Equation (12.2), results in the quantity

$$\sum_{i=1}^{N} \frac{\left(X_i - \bar{X}\right)^2}{\sigma^2} \tag{12.6}$$

and the conclusion is that this quantity is a chi-square random variable with $N - 1$ degrees of freedom. But in this case where all the means are equal, it follows that the data $X_1, X_2, \ldots, X_N$ constitute a sample from a normal population, and thus Equation (12.6) is equal to $\left(N-1\right)S^2 / \sigma^2$, where $S^2$ is the sample variance. In other words, the conclusion in this case is just the well-known result that $\left(N-1\right)S^2 / \sigma^2$ is a chi-square random variable with $N - 1$ degrees of freedom.

## 12.3 ONE-WAY ANALYSIS OF VARIANCE

Consider *m* independent samples, each of size *n*, where the members of the *i*-th sample — $X_{i1}, X_{i2}, \ldots X_{in}$ — are normal random variables with unknown mean $\mu_i$ and unknown variance $\sigma^2$. That is,

$$X_{ij} \sim N\left(\mu_i, \sigma^2\right), \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n \tag{12.7}$$

We will be interested in testing

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_m \tag{12.8}$$

versus

$$H_1 : \text{ not all the means are equal}$$

We will be testing the null hypothesis that all the population means are equal against the alternative that at least two of them differ.

One way of thinking about this is to imagine that we have *m* different treatments, where the result of applying treatment *i* on an item is a normal random variable with mean $\mu_i$ and variance $\sigma^2$. We are then interested in testing the hypothesis that all treatments have the same effect, by applying each treatment to a different sample of *n* items and then analyzing the result.

Since there are a total of $n \times m$ independent normal random variables $X_{ij}$, it follows that the sum of the squares of their standardized versions will be a chi-square random variable with $nm$ degrees of freedom. That is,

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - E\left[X_{ij}\right]\right)^2 / \sigma^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - \mu_i\right)^2 / \sigma^2 \sim \chi^2_{nm} \qquad (12.9)$$

To obtain estimators for the *m* unknown parameters $\mu_1, \mu_2, \ldots, \mu_m$, let $X_{i\bullet}$ denote the average of all the elements in sample *i*; that is,

$$X_{i\bullet} = \sum_{j=1}^{n} X_{ij} / n \qquad (12.10)$$

The variable $X_{i\bullet}$ is the sample mean of the *i*-th population, and as such is the estimator of the population mean $\mu_i$, for $i = 1, \ldots, m$. Hence, if in Equation (12.9) we substitute the estimators $X_{i\bullet}$ for the means $\mu_i$, for $i = 1, \ldots, m$, then the resulting variable

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - X_{i\bullet}\right)^2 / \sigma^2 \qquad (12.11)$$

will have a chi-square distribution with $nm - m$ degrees of freedom. (1 degree of freedom is lost for each parameter that is estimated.) Let

$$SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - X_{i\bullet}\right)^2 \qquad (12.12)$$

and so the variable in Equation (12.11) is $SS_W / \sigma^2$. Because the expected value of a chi square random variable is equal to its number of degrees of freedom, it follows upon taking the expectation of the variable in (12.11) that

$$E\left[SS_W\right] / \sigma^2 = nm - m \qquad (12.13)$$

or, equivalently,

$$E\left[SS_W / \left(nm - m\right)\right] = \sigma^2 \qquad (12.14)$$

We thus have our *first estimator* of $\sigma^2$, namely, $SS_W/(nm-m)$. Note that this estimator was obtained without assuming anything about the truth or falsity of the null hypothesis.

**Definition** The statistic

$$SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - X_{i\bullet}\right)^2 \qquad (12.15)$$

is called the *within samples sum of squares* because it is obtained by substituting the sample population means for the population means in expression (12.9). The statistic

$$SS_W/(nm-m) \qquad (12.16)$$

is an estimator of $\sigma^2$.

Our second estimator of $\sigma^2$ will only be a valid estimator when the null hypothesis is true. So let us assume that $H_0$ is true and so all the population means $\mu_i$ are equal, say, $\mu_i = \mu$ for all *i*. Under this condition it follows that the *m* sample means $X_{1\bullet}, X_{2\bullet}, \ldots, X_{m\bullet}$ will all be normally distributed with the same mean μ and the same variance $\sigma^2/n$. Hence, the sum of squares of the *m* standardized variables

$$\frac{X_{i\bullet}-\mu}{\sqrt{\sigma^2/n}} = \sqrt{n}\left(X_{i\bullet}-\mu\right)/\sigma \qquad (12.17)$$

will be a chi-square random variable with *m* degrees of freedom. That is, when $H_0$ is true,

$$n\sum_{i=1}^{m}\left(X_{i\bullet}-\mu\right)^2/\sigma^2 \sim \chi_m^2 \qquad (12.18)$$

Now, when all the population means are equal to μ, then the estimator of μ is the average of all the $nm$ data values. That is, the estimator of μ is $X_{\bullet\bullet}$, given by

$$X_{\bullet\bullet} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}}{nm} = \frac{\sum_{i=1}^{m}X_{i\bullet}}{m} \qquad (12.19)$$

If we now substitute $X_{\bullet\bullet}$ for the unknown parameter μ, in (12.18) it follows, when $H_0$ is true, that the resulting quantity

$$n\sum_{i=1}^{m}\left(X_{i\bullet}-X_{\bullet\bullet}\right)^2/\sigma^2 \tag{12.20}$$

will be a chi-square random variable with $m-1$ degrees of freedom. That is, if we define $SS_b$ by

$$SS_b = n\sum_{i=1}^{m}\left(X_{i\bullet}-X_{\bullet\bullet}\right)^2 \tag{12.21}$$

then it follows that, when $H_0$ is true, $SS_b/\sigma^2$ is chi-square with $m-1$ degrees of freedom.

From the above we obtain that when $H_0$ is true,

$$E\left[SS_b\right]/\sigma^2 = m-1 \tag{12.22}$$

or, equivalently,

$$E\left[SS_b/\left(m-1\right)\right] = \sigma^2 \tag{12.23}$$

So, when $H_0$ is true, $SS_b/\left(m-1\right)$ is also an estimator of $\sigma^2$.

**Definition** The statistic

$$SS_b = n\sum_{i=1}^{m}\left(X_{i\bullet}-X_{\bullet\bullet}\right)^2 \tag{12.24}$$

is called the *between samples sum of squares*. When $H_0$ is true, $SS_b/\left(m-1\right)$ is an estimator of $\sigma^2$.

Thus we have shown that

| | |
|---|---|
| $SS_W/\left(nm-m\right)$ | allways estimates $\sigma^2$ |
| $SS_b/\left(m-1\right)$ | estimates $\sigma^2$ when $H_0$ is true |

Because it can be shown that $SS_b/\left(m-1\right)$ will tend to exceed $\sigma^2$ when $H_0$ is not true, it is reasonable to let the *test statistic* be given by

$$TS = \frac{SS_b/\left(m-1\right)}{SS_W/\left(nm-m\right)} \tag{12.25}$$

and to reject $H_0$ when TS is sufficiently large.

To determine how large TS needs to be to justify rejecting $H_0$, we use the fact that it can be shown that if $H_0$ is true then $SS_b$ and $SS_W$ are independent. It follows from this that, when $H_0$ is true, TS has an *F-distribution* with $m-1$ numerator and $nm-m$ denominator degrees of freedom.

Let $F_{m-1,nm-m,\alpha}$ denote the $100(1-\alpha)$ percentile of this distribution — that is,

$$P\left(F_{m-1,nm-m} > F_{m-1,nm-m,\alpha}\right) = \alpha \qquad (12.26)$$

where we are using the notation $F_{r,s}$ to represent an *F*-random variable with *r* numerator and *s* denominator degrees of freedom.

The significance level α test of $H_0$ is as follows:

$$\begin{array}{lll} \text{Reject} & H_0 \quad \text{if} & \dfrac{SS_b/(m-1)}{SS_W/(nm-m)} > F_{m-1,nm-m,\alpha} \qquad (12.27) \\ \text{Do not reject} & H_0 & \text{otherwise} \end{array}$$

A table of values of $F_{r,s,0.05}$ for various values of *r* and *s* is presented in Tables in statistic books. Part of these tables is presented in Table 1. For instance, from Table 1 we see that there is a 5 percent chance that an *F*-random variable having 3 numerator and 10 denominator degrees of freedom will exceed 3.71.

| s=Degrees of freedom for the Denominator | r= Degrees of freedom for the Numerator | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 |

Table 1

Another way of doing the computations for the hypothesis test that all the population means are equal is by computing the *p*-value. If the value of the test statistic is $TS = v$ , then the *p*-value will be given by

$$p-\text{value}=\Pr\left(F_{m-1,nm-m}\geq v\right) \qquad (12.28)$$

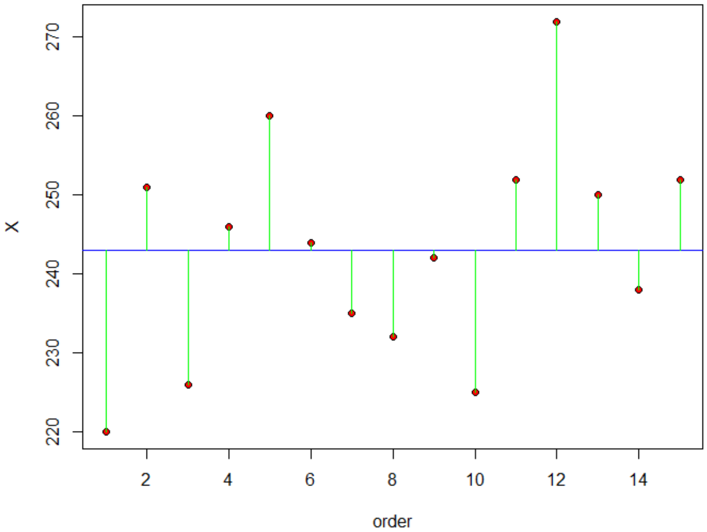
**Example** An auto rental firm is using 15 identical motors that are adjusted to run at a fixed speed to test 3 different brands of gasoline. Each brand of gasoline is assigned to exactly 5 of the motors. Each motor runs on 10 gallons of gasoline until it is out of fuel.
The following represents the total mileages obtained by the different motors:

| | | | | | |
|---|---|---|---|---|---|
| Gas 1: | 220 | 251 | 226 | 246 | 260 |
| Gas 2: | 244 | 235 | 232 | 242 | 225 |
| Gas 3: | 252 | 272 | 250 | 238 | 256 |

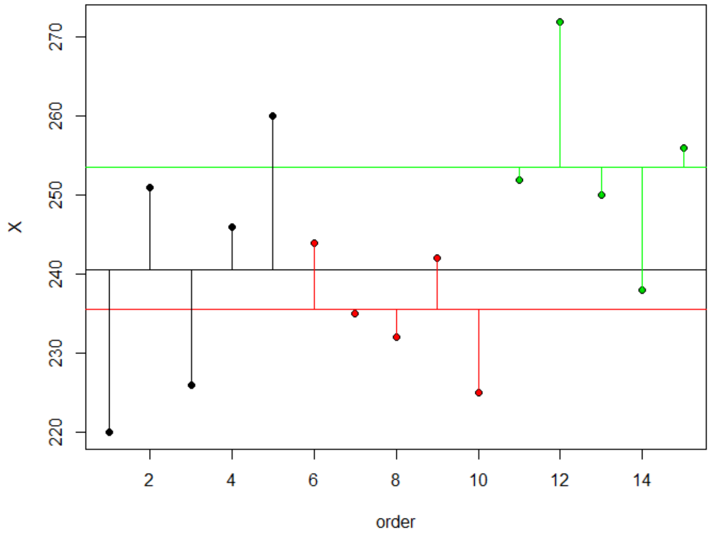
Test the hypothesis that the average mileage obtained is not affected by the type of gas used. Use the 5 percent level of significance.

The variance in *X*, the mileage, is large.

The total sum of squares $\sum_{i=1}^{15}(x-\bar{x})^2$



Instead of fitting the overall mean value thorough the data, let us fit the individual treatment means:



When the *means are significantly different*, then the sum of squares computed from the individual treatment means will be smaller than the sum of squares computed from the overall mean. We judge the significance of the difference between the two sums of squares using analysis of variance.

$$X_{1\bullet} = 240.6 \qquad X_{2\bullet} = 235.6 \qquad X_{3\bullet} = 253.6$$

$$SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - X_{i\bullet}\right)^2 \qquad SS_b = n\sum_{i=1}^{m}\left(X_{i\bullet} - X_{\bullet\bullet}\right)^2$$

$$SS_W = 1991.6 \qquad SS_b = 863.3333$$

The value of the *F*-statistic is

$$TS = \frac{SS_b/(m-1)}{SS_W/(nm-m)} = \frac{863.3333/(3-1)}{1991.6/(5\times3-3)} = 2.600924$$

Critical value $F_{2,12,0.05} = 3.885294$

Since $2.600924 < 3.885294$ the null hypothesis that the mean mileage is the same for all 3 brands of gasoline cannot be rejected.

Or, p-value is greater than 0.05, the null hypothesis that the mean mileage is the same for all 3 brands of gasoline cannot be rejected.

```
pvalue<- 1-pf(2.600924,df1=2, df2=12)
> pvalue
[1] 0.1152489
```

The procedure performed by R:

```
summary(aov(date15$Mile~date15$Gas))
            Df Sum Sq Mean Sq F value Pr(>F)
date15$Gas   2  863.3   431.7   2.601  0.115
Residuals   12 1991.6   166.0
```

```
summary.lm(aov(date15$Mile~date15$Gas))

Call:
aov(formula = date15$Mile ~ date15$Gas)

Residuals:
   Min     1Q Median     3Q    Max 
 -20.6   -7.1   -0.6    7.4   19.4 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     240.600      5.761  41.761  2.3e-14 ***
date15$Gasgas2   -5.000      8.148  -0.614    0.551    
date15$Gasgas3   13.000      8.148   1.596    0.137    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.88 on 12 degrees of freedom
Multiple R-squared:  0.3024,	Adjusted R-squared:  0.1861 
F-statistic: 2.601 on 2 and 12 DF,  p-value: 0.1152
```

The following algebraic identity, called the *sum of squares identity*, is useful when doing the computations by hand.

The Sum of Squares Identity

$$\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^{2} = nmX_{\bullet\bullet}^{2} + SS_b + SS_W \tag{12.29}$$

When computing by hand, the quantity $SS_b$ defined by

$$SS_b = n\sum_{i=1}^{m}\left(X_{i\bullet} - X_{\bullet\bullet}\right)^2$$

should be computed first. Once $SS_b$ has been computed, $SS_W$ can be determined from the sum of squares identity. That is, $\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2$ and $X_{\bullet\bullet}^2$ should also be computed and then $SS_W$ determined from

$$SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 - nmX_{\bullet\bullet}^2 - SS_b$$

Let us do the computations of Example by hand.

| | | | | | |
|---|---|---|---|---|---|
| Gas 1: | 220 | 251 | 226 | 246 | 260 |
| Gas 2: | 244 | 235 | 232 | 242 | 225 |
| Gas 3: | 252 | 272 | 250 | 238 | 256 |

$m = 3 \qquad n = 5$

$X_{1\bullet} = 240.6 \qquad X_{2\bullet} = 235.6 \qquad X_{3\bullet} = 253.6$
$X_{\bullet\bullet} = 243.2667$

$$SS_b = 5\left[\left(240.6-243.26\right)^2+\left(235.6-243.26\right)^2+\left(253.6-243.26\right)^2\right] = 863.33$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 = 890535$$

$$SS_W = 890535 - 15\times\left(243.26\right)^2 - 863.33 = 1991.57$$

Let show that $E\left[SS_b/\left(m-1\right)\right] \geq \sigma^2$ with equality only when $H_0$ is true.
So, we must show that

$$E\left[\sum_{i=1}^{m}\left(X_{i\bullet}-X_{\bullet\bullet}\right)^2/\left(m-1\right)\right] \geq \sigma^2/n \qquad (12.30)$$

with equality only when $H_0$ is true. To verify this, let $\mu_\bullet = \sum_{i=1}^{m}\mu_i/m$ be the average of the means. Also, for $i = 1, 2, \ldots, m$, let

$$Y_i = X_{i\bullet} - \mu_i + \mu_\bullet \qquad (12.31)$$

Because $X_{i\cdot}$ is normal with mean $\mu_i$ and variance $\sigma^2/n$, it follows that $Y_i$ is normal with mean $\mu_{\cdot}$ and variance $\sigma^2/n$. Consequently, $Y_1, Y_2, \ldots, Y_m$ constitutes a sample from a normal population having mean $\mu_{\cdot}$ and variance $\sigma^2/n$. Let

$$\bar{Y} = Y_{\cdot} = \sum_{i=1}^{m} Y_i / m = X_{\cdot\cdot} - \mu_{\cdot} + \mu_{\cdot} = X_{\cdot\cdot}$$

be the average of these variables. Now,

$$X_{i\cdot} - X_{\cdot\cdot} = Y_i + \mu_i - \mu_{\cdot} - Y_{\cdot}$$

$$E\left[\sum_{i=1}^{m}\left(X_{i\cdot} - X_{\cdot\cdot}\right)^2\right] = E\left[\sum_{i=1}^{m}\left(Y_i - Y_{\cdot} + \mu_i - \mu_{\cdot}\right)^2\right]$$

$$= E\left[\sum_{i=1}^{m}\left(\left(Y_i - Y_{\cdot}\right)^2 + \left(\mu_i - \mu_{\cdot}\right)^2 + 2\left(\mu_i - \mu_{\cdot}\right)\left(Y_i - Y_{\cdot}\right)\right)\right]$$

$$= E\left[\sum_{i=1}^{m}\left(\left(Y_i - Y_{\cdot}\right)^2\right)\right] + \sum_{i=1}^{m}\left(\mu_i - \mu_{\cdot}\right)^2 + 2\sum_{i=1}^{m}\left(\mu_i - \mu_{\cdot}\right)E\left[Y_i - Y_{\cdot}\right]$$

$$= (m-1)\sigma^2/n + \sum_{i=1}^{m}\left(\mu_i - \mu_{\cdot}\right)^2 + 2\sum_{i=1}^{m}\left(\mu_i - \mu_{\cdot}\right)E\left[Y_i - Y_{\cdot}\right]$$

$$= (m-1)\sigma^2/n + \sum_{i=1}^{m}\left(\mu_i - \mu_{\cdot}\right)^2$$

where the next to last equality follows because the sample variance $\sum_{i=1}^{m}\left(Y_i - Y_{\cdot}\right)^2/(m-1)$ is an unbiased estimator of its population variance $\sigma^2/n$ and the final equality because $E\left[Y_i - Y_{\cdot}\right] = E\left[Y_i\right] - E\left[Y_{\cdot}\right] = \mu_{\cdot} - \mu_{\cdot} = 0$. Dividing by $m-1$ gives that

$$E\left[\sum_{i=1}^{m}\left(X_{i\cdot} - X_{\cdot\cdot}\right)^2/(m-1)\right] = \sigma^2/n + \sum_{i=1}^{m}\left(\mu_i - \mu_{\cdot}\right)^2/(m-1)$$

and the result follows because $\sum_{i=1}^{m}\left(\mu_i - \mu_{\cdot}\right)^2 \geq 0$ with equality only when all the $\mu_i$ are equal.

Table 2 sums up the results of this section.

TABLE 2 One-Way ANOVA

| Source of Variation | Sum of Squares | Degrees of Freedom | Value of Test Statistic |
|---|---|---|---|
| Between samples | $SS_b = n\sum_{i=1}^{m}\left(X_{i\bullet} - X_{\bullet\bullet}\right)^2$ | $m-1$ | $TS = \frac{SS_b/(m-1)}{SS_W/(nm-m)}$ |
| Within samples | $SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - X_{i\bullet}\right)^2$ | $nm-m$ | |

Significance level α test:

Reject $H_0$ if $TS \geq F_{m-1,nm-m,\alpha}$
do not reject otherwise

If $TS = v$ , then $p - \text{value} = P\left\{F_{m-1,nm-m} \geq v\right\}$ .