

Method of Least Squares for STAT112

Charles Fleming

August 1, 2016

1 Mathematical Background

Published tables generally include the commonly used values, for one reason, in order to conserve space. For another reason, since it was very laborious to construct computationally difficult tables before the advent of computers, tables as a result were usually copyrighted, so that securing permission to publish them posed another incentive to abbreviate tables as much as possible. Consequently, it is not uncommon to discover, during the process of finding probabilities, values which are not in the table. Suppose a z-score had been calculated to have a value of 1.227 for which its corresponding probability is desired. The table of probabilities for the Standard Normal distribution which appears in Appendix 6 provides a listing for $P(z \leq 1.22) = .88877$ and $P(z \leq 1.23) = .89065$, but provides no listing for $P(z \leq 1.227)$. An entry in the table of probabilities depicting our dilemma might appear as:

c	$P(z \leq c)$
1.22	.88877
1.227	?
1.23	.89065

In the era before hand calculators and computers, a simple way to obtain a value of $P(X \leq 1.227)$ would have followed the method of linear interpolation by which an approximate value for $P(X \leq 1.227)$ would be computed from its neighbors by using straight lines like those shown in Figure 1.

Linear interpolation is one of many techniques falling under the auspices of the discipline called *numerical analysis*. It is a branch of mathematics and computer science in which methods for finding numerical values of symbolic numbers are studied and developed. For example, when one asks the question: “*What is $\sqrt{2}$?*”, he is asking for a decimal value of $\sqrt{2}$. Such a question posed a philosophical problem for the Greeks. Everything in

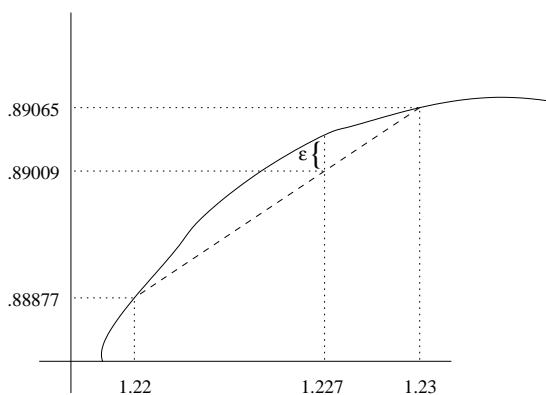


Figure 1

nature was supposed to have a rational explanation. Certainly, that sentiment is reflected in the development of Greek grammar. Numbers, too, had to be derivable in a rational way from whole numbers, but $\sqrt{2}$ is an irrational number. The Greeks could prove that $\sqrt{2}$ was irrational, but they found the conclusion upsetting to the extent that a student of Pythagoras whose belief in suggesting the validity of the existence of irrational numbers led, as it was told, to his execution.

The symbol, $\sqrt{2}$, represents a number which is the root of $f(x) = x^2 - 2$. It exists because $f(x) = x^2 - 2$ is a continuous function over the real numbers, and it is positive for $x=2$ and negative for $x=1$; therefore, somewhere in between one and two, there is a number, c , such that $f(c)=0$, and that number is given the symbol $\sqrt{2}$. The numbers, 1.41421 and 1.41422, are close approximations to $\sqrt{2}$ because $1.41421^2 = 1.999989924$ and $1.41422^2 = 2.000018208$. But more importantly, these two approximations bracket $\sqrt{2}$; that is $1.41421 \leq \sqrt{2} \leq 1.41422$. By successive approximations of trial and error ad infinitum, better decimal expansions of $\sqrt{2}$ can be obtained.

Although the cumulative distribution function of the Normal distribution is everywhere bending, yet, for very small portions of it, a straight line can be used to approximate it. That line defined by the two known points, (1.22,.88877) and (1.23,.89065), approximates the Standard Normal cumulative distribution in the vicinity at the point of interest, and it is drawn in Figure 1. By means of the straight line, it is an easy matter to find the ordinate for 1.227. Specifically, since the slope of the line is $\frac{.89065-.88877}{1.23-1.22} = \frac{y-.88877}{x-1.22}$, this expression can be rearranged so that $y = .88877 + \frac{.89065-.88877}{1.23-1.22}(x - 1.22)$. Therefore, when $x=1.227$, $y=.890086$. From another much different method which is available on a computer, a more precise answer is: .8900887. Even though the method of linear interpolation easily provides an approximation, the approximation is still erroneous even by a little amount which is denoted by the letter, ϵ , in Figure 1. Though it might be an arduous task, it is theoretically possible, nonetheless, to obtain a numerical value to any degree of precision by successive

applications of linear interpolation. All numerical techniques produce approximations to exact values. Though the decimal expansion of π is known to billions of digits and ϵ is miniscule, it is not an exact value for π . Whenever numerical calculations are made, a certain estimate of ϵ is given to indicate the precision of the approximation.

There are many other methods besides linear interpolation which are used for interpolating; there are the methods of splines, orthogonal polynomials, finite differences, to name a few, and there is the method of least squares. The appeal of an individual method depends on the computational resources and requirements of the analyst. A research laboratory with massive computational resources might prefer one method while a small satellite in orbit about the Earth might require another. All of the numerical methods which have been cited above are used in the analysis of data to some extent, but the method of least squares which was developed by Frederick Gauss in 1801 commands general use throughout the experimental sciences and has become an immensely popular statistical technique.

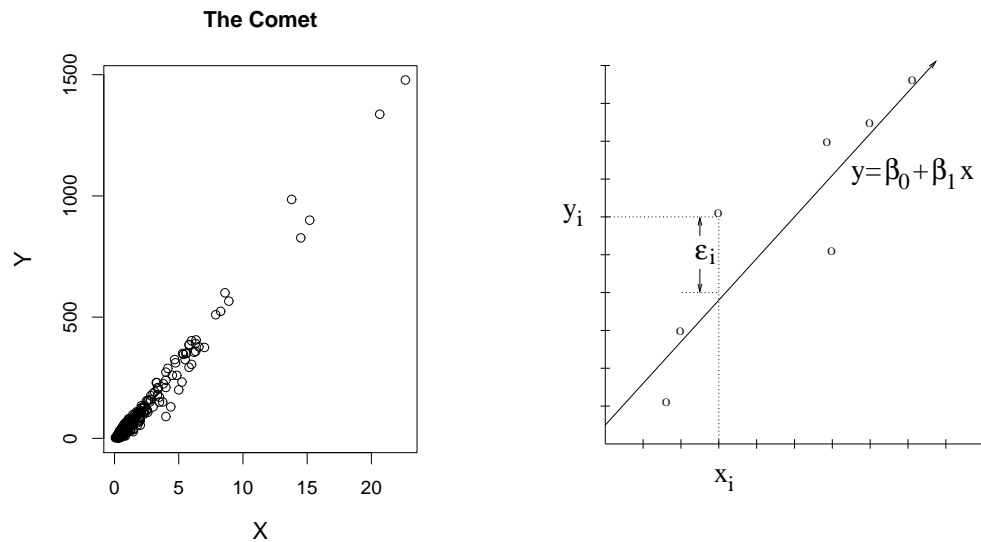


Figure 2

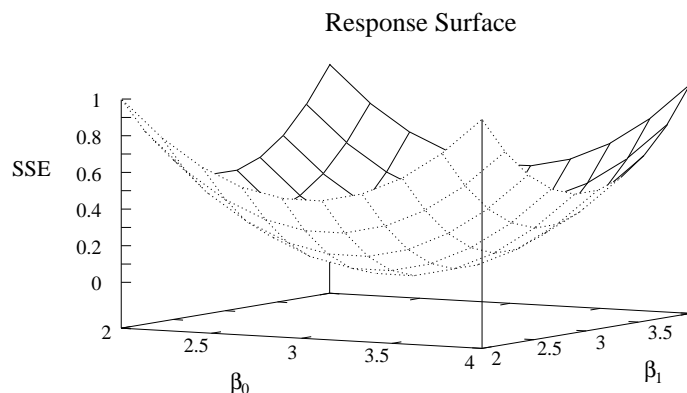
Upon making a picture of the data like the one shown on the left in Figure 2, a linear relationship in the data might readily appear. It suggests that a straight line can be used to describe the data. The line can only be an approximation as there will always be an error, ϵ_i , between the the line and the data point, i . Nonetheless, if the deviation of the line from the data is small, then it could serve as a good approximation to explain the data. As illustrated in the diagram shown on the right in Figure 2, at every value of x_i , there exists a deviation, ϵ_i , between the proposed line and the observed value, y_i .

If there does exist an inherent linear relationship between x and y , the linear relationship should be manifest in the plot of the data. In an attempt to deduce the equation of that linear relationship, it would seem reasonable to argue that the line which best fits the experimental data should also be the best candidate to explain the actual linear relationship. But in what sense does the idea of *best fit* mean was answered by the brilliant mathematician Frederick Gauss. Whatever error might be realized in measuring a value of x , it is presumed to be very small in comparison to the error in measuring a value of y , so that all the errors in taking measurements are attributed to y , and x is assumed to be fixed. For every i , the deviation of the observed value, y_i from the value of y of the proposed line should be zero, if there is no error in measuring the data and if the hypothesis that an inherent linear relationship exists between x and y is correct. But error does exist; no measurement can ever be exact. That line which best fits the data and which is, therefore, the best estimate of the true line is that one which minimizes the deviations, ϵ_i . Unfortunately, it is a vain attempt to minimize the ϵ_i 's, because the ϵ_i 's can be negative. To get rid of the influence of negative deviations, a feasible criterion for finding the best line stipulates that the line which minimizes the sum of squared errors defines the best line for explaining the inherent linear relationship of x and y .

2 Response Surface

Definition 1. *The sum of squared errors is defined to be: $SSE = \sum_{i=1}^n \epsilon_i^2$.*

Based on the proposition that x and y are related in a linear way, that is, $y = \beta_0 + \beta_1 x$, an estimate of that linear relationship is made by a line, $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$, which comes closest to fitting the data in the sense that it minimizes $SSE = \sum_{i=1}^n \epsilon_i^2$. For a given x_i which has been taken from the data, $\widehat{\beta}_0 + \widehat{\beta}_1 x_i$ does not exactly equal the observed y_i except by the deviation, ϵ_i . In other words, $\widehat{\beta}_0 + \widehat{\beta}_1 x_i + \epsilon_i = y_i$. Upon rearranging this last expression, $\epsilon_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$. If the hats are removed so that $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$, then ϵ_i becomes a function of β_0 and β_1 which in turn implies that $SSE = (y_1 - \beta_0 - \beta_1 x_1)^2 + (y_2 - \beta_0 - \beta_1 x_2)^2 + \dots + (y_n - \beta_0 - \beta_1 x_n)^2$ is also a function of β_0 and β_1 . For every β_0 and β_1 , there is a different SSE and for that matter a different line. There is an SSE for every line and there is a line for every SSE. That β_0 and that β_1 which minimize SSE will define the line, $y = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, which best fits the data. A picture of SSE as a function of β_0 and β_1 will make the concept of the method of least squares clearer.



The graph of SSE as a function of β_0 and β_1 shows a surface hovering in space. The surface has the shape of a bowl or a sheet which is suspended by four poles, and it is called the *response surface* of the model. Volumes of books have been written on response surface methodology as some response surfaces exhibit very complicated contours according to the complexity of the model. In the simple surface shown above, the graph clearly shows that the bowl has a bottom, a lowest point, a minimum value. Every conceivable straight line that passes through or about the data corresponds to a single point on the response surface, and conversely every single point on the response surface corresponds to a line drawn through the data. That line which minimizes the SSE corresponds to that point lying at the bottom of the bowl. There is one and only one such point which implies that there is one and only one line which best fits the data; hence the name: the *method of least squares*.

Why not minimize the sum of the absolute value errors, or the least median squared errors, or the least α quantile squared errors, or the least trimmed squared errors, etc.? There are, in fact, many criteria for determining that β_0 and that β_1 which determine the best fitting line to the data. The one that is chosen depends on the analyst who will arrive at his opinion after having adequately studied a problem. Among the numerous criteria which statisticians have advocated over the years, the method of least squares is by far the nicest one to use, because it admits a relatively easy derivation of simple formulas for the beta's. According to calculus, that β_0 and that β_1 which lie at the bottom of the bowl correspond to that point at which the tangent plane is horizontal, that is, where $\frac{\partial SSE}{\partial \beta_0} = \frac{\partial SSE}{\partial \beta_1} = 0$. These equations translate into the following formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Theorem 1. *The least squares estimators of β_0 and β_1 are:*

$$1. \hat{\beta}_0 = \frac{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i)}{n \sum_{i=1}^n (x_i - \bar{x})^2} = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$2. \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} = \frac{SS_{xy}}{SS_{xx}}$$

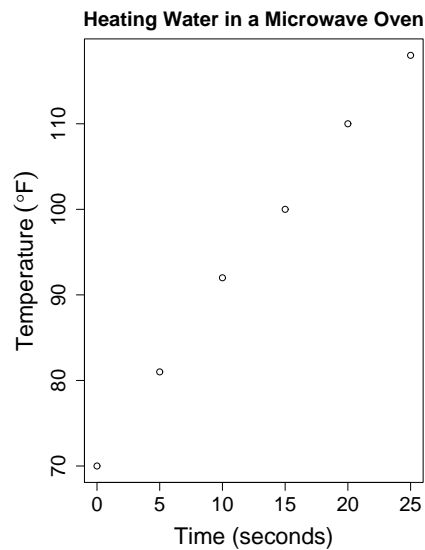
where $SS_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}$ and $SS_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$.

In the first item of Theorem 1, a short cut for calculating $\hat{\beta}_0$ is, also, written. The steps in calculating $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

1. Compute $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$
2. Compute $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

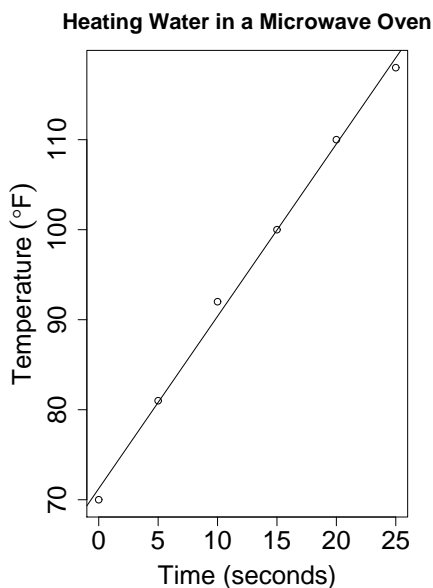
Example 1. Six plastic cups are filled with 8 oz. of water each. All cups of water are allowed to come to room temperature. The first cup of water is heated in a microwave oven for 5 seconds; the second cup of water is heated for 10 seconds; the third cup of water is heated for 15 seconds, and so on until the last cup of water is heated for 25 seconds. After each cup is heated, the final temperature of the water is recorded. The results of the experiment is shown in the table and graph below.

Time (sec)	Temperature ($^{\circ}F$)
0	70
5	81
10	92
15	100
20	110
25	118



It should go without saying that the first order of business is to make a picture of the data. When the set of data for the microwave experiment is plotted, the trend looks linear; therefore, it seems reasonable to assert that there is a linear relationship between temperature and the time spent in heating the water. If we let x denote time and y denote temperature, then the assertion that a linear relationship exists between x and y is expressed by: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Having decided on the linear nature of the model, performing the calculations begins the next step:

$$\begin{aligned}
 n &= 6 \\
 \sum x_i y_i &= 7975 \\
 \sum x_i^2 &= 1375 \\
 \sum x_i &= 75 \text{ hence } \bar{x} = 12.5 \\
 \sum y_i &= 571 \text{ hence } \bar{y} = 95.1667 \\
 SS_{xx} &= 1375 - \frac{75^2}{6} = 437.5 \\
 SS_{xy} &= 7975 - \frac{(75)(571)}{6} = 837.5 \\
 \hat{\beta}_1 &= \frac{837.5}{437.5} = 1.914 \\
 \hat{\beta}_0 &= \frac{571}{6} - 1.914\left(\frac{75}{6}\right) = 71.238
 \end{aligned}$$



The least squares fit is: $\hat{y} = 71.238 + 1.914x$.

The assertion that there exists a linear relation between x and y implies the formula $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. The process of calculating a least squares line requires many arithmetic computations which are susceptible to errors. A simple arithmetic mistake at any step in the calculations will lead to a wrong answer. In order to verify the accuracy of the computations, there is an easy way to check the arithmetic.

Corollary 1. $\sum_{i=1}^n \hat{\epsilon}_i = 0$ where $\hat{\epsilon}_i = y_i - \hat{y}_i$.

Proof.

$$\begin{aligned}
 \sum_{i=1}^n \hat{\epsilon}_i &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
 &= \sum_{i=1}^n y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i
 \end{aligned}$$

By definition of the mean, $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ so that $n\bar{y} = \sum_{i=1}^n y_i$. Similarly, $n\bar{x} = \sum_{i=1}^n x_i$. And since $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}$ can be substituted for $\widehat{\beta}_0$,

$$\begin{aligned} \sum_{i=1}^n y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i - n(\bar{y} - \widehat{\beta}_1\bar{x}) - \widehat{\beta}_1 \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n y_i - \sum_{i=1}^n y_i + \widehat{\beta}_1 \sum_{i=1}^n x_i - \widehat{\beta}_1 \sum_{i=1}^n x_i = 0 \end{aligned}$$

■

Law 1 (Fleming's Law). *Never trust a computer.*

One might wonder why in the age of programmable hand calculators and easy access to computers should the computer's arithmetic be checked for errors. After having entered the data into the statistics mode of a calculator or into some statistical software package on a computer should one not feel confident that the answers will be free of errors? A naive statistician may feel that way, but a computer scientist and a seasoned statistician will always verify the answer produced by a computer with a test like Corollary 1 which is known to be absolutely correct.

There are two principal reasons for fitting a line to the data.

- It is done for the purpose of interpolating.
- It is done to show an inherent relation between the variables.

Whatever evidence is presented to substantiate the claim that a linear relationship between the variables does exist, it must be cast in the form of confidence intervals. The use of confidence intervals makes it possible to defend an inference about the relationship. Otherwise any assertion about the relationship cannot be defended except by political might, a common practice to be sure but which anyone who subscribes to the principle of rational persuasion will oppose. Since the method of least squares is strictly a technique of numerical analysis in which random variables are absent and *a fortiori* confidence intervals are never mentioned, a way must be devised to introduce a natural property of the phenomenon into the method of least squares which will account for the randomness found in experimental data. What separates the field of numerical analysis and statistics is the utilization of random variables. When statisticians look upon the deviation, ϵ_i , they see a variable alluding to an event in which an error of size ϵ_i has occurred in measuring y_i . The deviation, ϵ_i , measures the size of that outcome; therefore, in the eyes of a statistician, ϵ_i is a random variable which comes with an associated probability distribution which is characteristic of the phenomenon.

In the theory of linear models, the linear relationship can be formulated in two ways.

Definition 2.

1. If $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ is proposed to explain the phenomenon and if y_i , x_i , and ϵ_i are assumed to be random variables, then this model is called a **random effects** model.
2. If $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ is proposed to explain the phenomenon and if only y_i and ϵ_i are random variables and the x_i 's are fixed, then this model is called a **fixed effects** model. Whatever randomness can be attributed to x is assumed to be negligible.

3 Fixed Effects Linear Model

Dealing with complicated computational demands of random effects models is such an excruciating experience that the prospects of doing so incline statisticians to avoid them whenever possible unless they have access to powerful computer programs and they know how to use them. The relative simplicity of fixed effects models, on the other hand, make them immensely popular. Computer software for calculating estimates of the parameters of a linear model is an integral element of any analyst's tool box. Linear fixed effects models abound in every facet of experimentation in every discipline. It is definitely a subject worth studying.

Definition 3. $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ and x_i is fixed is called a **fixed effects linear model**.

Definition 4. y_i is called the **response variable**; x_i is called the **independent variable**, or the **explanatory variable**, or the **predictor variable**.

The following are examples of linear and non-linear models. By linear, it is meant that the model is linear in the parameters.

Examples of a Linear Model

$$y_i = \beta_0 + \beta_1 x_i^2 + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 \cos(x_i) + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

Examples of a Non-linear Model

$$y_i = \beta_0 + (\beta_0 x_i)^{\beta_0} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_0 x_i \beta_1 + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

The language of statistics is full of jargon. The word, *regression*, as in *regression analysis*, for instance, is used as a synonym for the method of least squares by many authors for no clear reason. The origins of this association of regression with the method of least squares is obscure. Its connotation bears no resemblance to the meaning of *regression* which is understood in standard English nor to the Latin root from which it is derived. Supposedly, this peculiar meaning of *regression* was culled from the title of the article written by Francis Galton: “*Regression Toward Mediocrity in Heredity Stature*” Journal of the Anthropological Institute vol. 15, 1885, pp. 246-263 in which he utilized the method of least squares and it has remained a custom ever since. Many other examples of jargon abound in the statistical profession which prove the popularity of what must seem to be a puzzling language. In due time, when sufficient understanding of the underlying mathematics has been learned, the jargon can be overcome while acquiring a better appreciation of the benefits of using precise definitions based on good English. The method of least squares is one of those phrases found in statistics which has a precise meaning and, when ϵ is considered to be a random variable, the scope of the method of least squares broadens from encompassing only interpolation to include a method which admits the construction of confidence intervals. It is possible to construct confidence intervals in the use of linear models because the random nature which is ascribed to ϵ propagates through the equations of Theorem 1 to make $\hat{\beta}_0$ and $\hat{\beta}_1$ random variables.

Theorem 2. $\hat{\beta}_1$ is a random variable.

This claim is obvious because y_i is a random variable, and it appears in

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{SS_{xy}}{SS_{xx}}$$

Corollary 2. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ a random variable.

Whenever a linear model is proposed, the distribution of ϵ_i must always be stated, and it usually is assumed that $\epsilon_i \sim N(0, \sigma^2)$. Like the notion of randomness, the notion of noise has an intuitive meaning which is difficult to describe by mathematics. From theoretical arguments, noise may be modeled as a random variable which is distributed as $N(0, \sigma^2)$. On that account, because ϵ is meant to explain the random fluctuations in measurements which will naturally occur in an experiment, ϵ is assumed to follow a Normal distribution such that $E[\epsilon_i] = 0$ and $var(\epsilon_i) = \sigma^2$. If, on the other hand, $E[\epsilon_i] \neq 0$, then the model is deficient for it is not accounting for all of the pieces of information which are being observed about the phenomenon.

Theorem 3. If $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$, then $E[y_i] = \beta_0 + \beta_1 x_i$.

After calculating $\hat{\beta}_0$ and $\hat{\beta}_1$, they can be used in two ways. In the first way, $\hat{\beta}_0$ and $\hat{\beta}_1$ can be employed to estimate $E[y_p]$ by $\widehat{E[y_p]} = \hat{\beta}_0 + \hat{\beta}_1 x_p$, and in the second way, they can be used to estimate y_p by $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p + \epsilon_p$.

Definition 5. $\widehat{E[y_p]} = \hat{\beta}_0 + \hat{\beta}_1 x_p$ is called the **estimated expected value** of y_p .

That $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables make $\widehat{E[y_p]} = \hat{\beta}_0 + \hat{\beta}_1 x_p$ a random variable; therefore, it has an expectation, and a variance and a probability distribution which is associated with it.

It should be noted that the notation of authors becomes muddled with regard to $\widehat{E[y]}$. It is not uncommon to see $\widehat{E[y]}$ written as \hat{y} . Confusion arises unless one understands the author's reason for using a different convention because the other way of employing $\hat{\beta}_0$ and $\hat{\beta}_1$ is to put them back into the original statement of the model as $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p + \epsilon_p$, for the purpose of estimating a particular value of y_p . This raises a subtle point which some statisticians emphasize by calling \hat{y}_p a predicted value of y_p rather than an estimated value of y_p . In any case, \hat{y}_p is also a random variable, and it has associated with it a different probability distribution.

Definition 6. Let $y_p = \beta_0 + \beta_1 x_p + \epsilon_p$ where $\epsilon_p \sim N(0, \sigma^2)$, define the **predicted value** as $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p + \epsilon_p$

In summary, both $\widehat{E[y_p]}$ and \hat{y}_p are random variables; each has an expected value and each has a variance which are tabulated in Table 1.

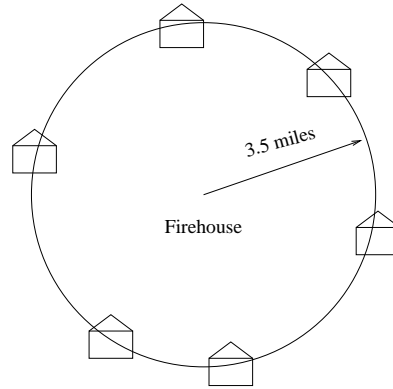
Table 1

	$\widehat{E[y_p]}$	\hat{y}_p
Expected Value	$E[\widehat{E[y_p]}] = \beta_0 + \beta_1 x_p$	$E[\hat{y}_p] = \beta_0 + \beta_1 x_p$
Variance	$var(\widehat{E[y_p]}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right)$	$var(\hat{y}_p) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right)$

By virtue of ascribing properties of a random variable to ϵ_i in formulating a model, it becomes possible to construct the confidence interval. For $\widehat{E[y_p]}$ and for \hat{y}_p , the lower and upper limits are given in Table 2.

Table 2: 100(1- α)% Confidence Intervals (a,b)

$E[\widehat{y}_p] = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$	$\widehat{y}_p = \widehat{\beta}_0 + \widehat{\beta}_1 x_p + \epsilon_p$
$a = E[\widehat{y}_p] - t_{n-2; \frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$	$a = \widehat{y}_p - t_{n-2; \frac{\alpha}{2}} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$
$b = E[\widehat{y}_p] + t_{n-2; \frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$	$b = \widehat{y}_p + t_{n-2; \frac{\alpha}{2}} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$
where $s^2 = \frac{sse}{n-2}$	where $s^2 = \frac{sse}{n-2}$



Example 2. *The following excellent example of illustrating the subtle difference in meanings of $E[\widehat{y}_p]$ and for \widehat{y}_p is taken from McClave, et al. The cost in thousands of dollars of repairing the damage which is caused by a fire is tabulated in Table 3 for fifteen houses according to the distance a house is from the nearest firehouse. Of course, a picture of the data must be drawn like the one shown in Figure 3 and placed with the tabulation of the data.*

$$n=15$$

$$\hat{\beta}_0 = 10.277929$$

$$\hat{\beta}_1 = 4.919331$$

$$SS_{xx} = 34.784$$

$$SSE=69.75$$

$$x_p = 3.5$$

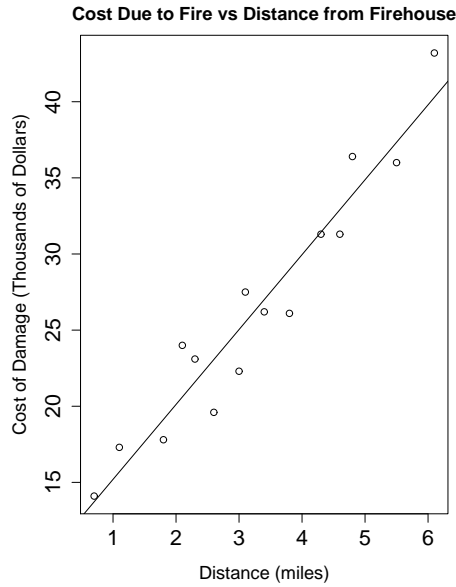


Figure 3

Table 3

Distance, x miles	Cost, y	Predicted Cost, \hat{y}	$\hat{\epsilon}^2 =$ $(y - \hat{y})^2$	Distance, x miles	Cost, y	Predicted Cost, \hat{y}	$\hat{\epsilon}^2 =$ $(y - \hat{y})^2$
3.4	26.2	27.0	0.64	1.8	17.8	19.1	1.77
4.6	31.3	32.9	2.58	2.3	23.1	21.5	2.27
3.1	27.5	25.5	3.88	5.5	36.0	37.3	1.78
.7	14.1	13.7	0.14	3.0	22.3	25.0	7.48
2.6	19.6	23.0	12.02	4.3	31.3	31.4	0.01
2.1	24.0	20.6	11.50	1.1	17.3	15.6	2.59
6.1	43.2	40.2	8.49	4.8	36.4	33.8	6.29
3.8	26.1	28.9	8.24				

Suppose a fire occurred 3.5 miles from a firehouse and that this event actually happened n times. Let c_i =the cost of damage to a house on the i^{th} fire. The sample mean, $\bar{c} = \frac{c_1+c_2+\dots+c_n}{n}$ is unknown, but we can estimate it by: $\hat{\bar{c}} = \widehat{E[y_i]} = 10.277929 + 4.919331(3.5) = 27.4956$.

Find the 95% CI about $\bar{c} = E[\widehat{E[y_i]}]$. Based on the data, $\hat{\beta}_0 = 10.277929$ and $\hat{\beta}_1 = 4.919331$.

Before beginning to solve this problem, a glance at Table 2 shows that SSE must be calculated, in order to construct a confidence interval. There is no easy way to compute SSE without the aid of a computer for corresponding to each x_i , a predicted value of y_i must be computed by means of the equation, $\widehat{\beta}_0 + \widehat{\beta}_1 x_i$. For example, $\widehat{E}[y_i] = 10.277929 + 4.919331(3.5) = 27.0$. A listing of predicted values of y is shown in Table 3. The squared deviations corresponding to each x_i are also listed. The first of them is: $(26.2 - 27.0)^2 = .64$. The sum of the squared deviations is $SSE = \sum_{i=1}^{15} \epsilon_i^2 = 69.75$. According to Table 2, $s^2 = \frac{SSE}{n-2} = \frac{69.75}{13} = 5.36546$. Now, all the necessary pieces to construct a confidence interval are ready.

1. $\alpha = .05 \rightarrow \frac{\alpha}{2} = .025 \quad n=15$

2. $t_{13,.025} = 2.160$

3. Compute:

(a) $\widehat{E}[y_p] = 10.277929 + 4.919331(3.5) = 27.4956$

(b) $s^2 = \frac{sse}{n-2} = \frac{69.75}{13} = 5.36546 \rightarrow s = 2.316$

(c) $\sum x_i = 49.2 \rightarrow \bar{x} = 3.28$

(d) $\sum x_i^2 = 196.16$

(e) $SS_{xx} = 34.784$

4. $a = 27.4956 - 2.160\sqrt{5.36546}\sqrt{\frac{1}{15} + \frac{(3.5-3.28)^2}{34.784}} = 27.4956 - 1.3053 = 26.1903$.

5. $b = 27.4956 + 2.160\sqrt{5.36546}\sqrt{\frac{1}{15} + \frac{(3.5-3.28)^2}{34.784}} = 27.4956 + 1.3053 = 28.8009$.

6. 95% CI of the average cost of a fire, $E[\widehat{E}[y]] = \bar{c}$ is (26.1903, 28.8009).

Example 3. Same situation as above except that a fire has not yet occurred at 3.5 miles. What is the 95% CI of the cost of the damage to my house. That is: $\widehat{y}_{my \text{ house}} = \widehat{\beta}_0 + \widehat{\beta}_1 3.5 = 10.277929 + 4.919331(3.5) = 27.4956$.

1. $\alpha = .05 \rightarrow \frac{\alpha}{2} = .025$

2. $t_{13,.025} = 2.160$

3. Use same computations as was used in item 3 of Example 2.

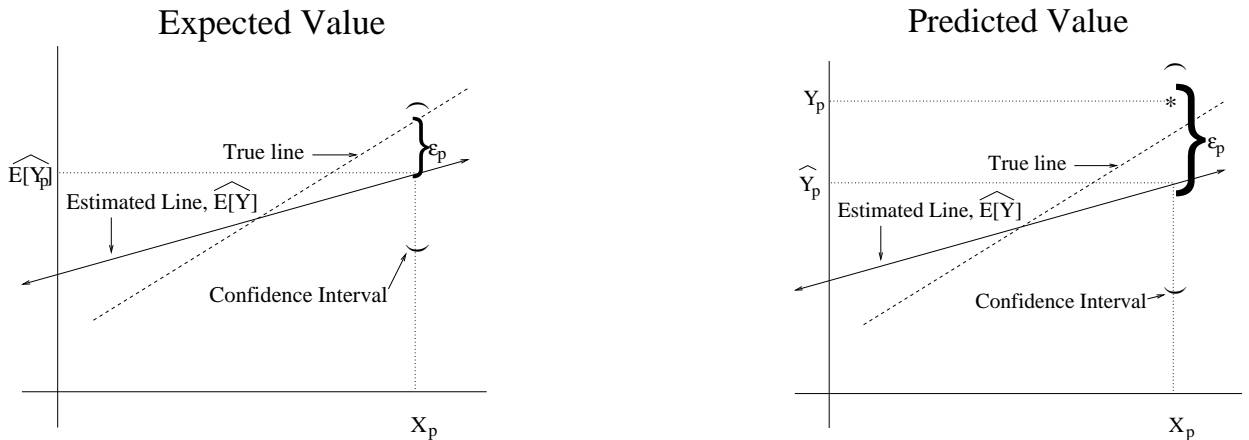
4. $a = 27.4956 - 2.160\sqrt{5.36546}\sqrt{1 + \frac{1}{15} + \frac{(3.5-3.28)^2}{34.784}} = 27.4956 - 5.17076 = 22.3248$

$$5. b = 27.4956 + 2.160\sqrt{5.36546}\sqrt{1 + \frac{1}{15} + \frac{(3.5-3.28)^2}{34.784}} = 27.4956 + 5.17076 = 32.66636.$$

6. 95% CI of the cost of a particular house is (22.3248, 32.6664).

In estimating the average claim, \bar{c} , for a damaged house located at 3.5 miles from a firehouse by $\widehat{E}[y_i]$, an insurance company for instance would be estimating the amount of money it would be paying, on the average, to a community of houses which is situated that far from a firehouse. Although some claims will be more or less than the average, an insurance company is interested in the average cost of damage. On the other hand, an individual is only interested in the cost which he will incur from damage to his house which lies 3.5 miles from the firehouse. The insurance company is interested in the average cost, while an individual is interested in a particular cost.

That there is more uncertainty in predicting the cost of the damage to a particular house than predicting the average cost of damage to a community of houses at the same distance from the firehouse is a consequence of averaging. The \widehat{y}_p does not only require the use of the fitted line $\widehat{\beta}_0 + \widehat{\beta}_1 x_p$ which is uncertain in itself, but it depends on the deviation, ϵ_p , from the fitted line to the true value of y_p .



Depicted in the left hand figure shown above, is the relation of the line of the estimated expected value, $\widehat{E}[y]$, to the true line which we are seeking. That same line of the estimated expected value is used again to estimate a particular y_p as shown in the right hand figure. Although both $\widehat{E}[y_p]$ and \widehat{y}_p correspond to the same value because we are using the same line in both cases to estimate them, the error between \widehat{y}_p and y_p must account for the additional uncertainty due to y_p not lying on the true line, hence the confidence interval for $\widehat{E}[y_p]$ is longer than the one for $\widehat{E}[\widehat{y}_p]$.

Because $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are random variables, they have probability distributions associated with them as well as expected values and variances. The next theorem gives the explicit

formulas for the expected values and variances which will play an important practical role in the final analysis of a linear model which is called the analysis of variance.

Theorem 4.

1. $E[\widehat{\beta}_0] = \beta_0$

2. $E[\widehat{\beta}_1] = \beta_1$

3.
$$var(\widehat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nSS_{xx}}$$

4.
$$var(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SS_{xx}}$$

5.
$$cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = -\frac{\sigma^2 \bar{x}}{SS_{xx}}$$

The covariance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ which is denoted by $cov(\widehat{\beta}_0, \widehat{\beta}_1)$ is a measure of how much $\widehat{\beta}_0$ and $\widehat{\beta}_1$ depend on each other. That $cov(\widehat{\beta}_0, \widehat{\beta}_1) \neq 0$ implies that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are not independent but correlated. A more detailed discussion of the covariance will occur in a subsequent statistics course.

There is no short-cut in calculating SSE other than using a computer. Manual computations of SSE are long and tedious, yet SSE is essential for constructing confidence intervals and testing hypotheses pertaining to a linear model. By assumption, $\epsilon \sim N(0, \sigma^2)$. The next theorem gives the unbiased estimator of σ^2 .

Theorem 5. For a linear fixed effects model with r parameters where $SSE = \widehat{\epsilon}_1^2 + \widehat{\epsilon}_2^2 + \dots + \widehat{\epsilon}_n^2$, $\widehat{\sigma}^2 = \frac{SSE}{n-r}$ is the unbiased estimator of σ^2 . That is, $E[\widehat{\sigma}^2] = \sigma^2$. Also, $var(\widehat{\sigma}^2) = 2(n-r)\sigma^4$ and $\frac{(n-r)\widehat{\sigma}^2}{\sigma^2} \sim X_{n-r}^2$.

Corollary 3. For a linear model with two parameters, $\widehat{\sigma}^2 = \frac{SSE}{n-2}$.

Definition 7. The distribution, X_n^2 , is pronounced, **chi-squared**.

The X^2 distribution owes its origins to the Standard Normal distribution. Specifically, if $z_i \sim N(0, 1)$, then $z_i^2 \sim X_1^2$. More generally, if $Z^2 = \sum_{i=1}^n z_i^2$ then $Z^2 \sim X_n^2$, that is, X_n^2 is generated by a sum of n Standard Normal distributions that are squared. The probability density function of a X_n^2 distribution is:

$$f(x) = \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

Theorem 6. $E[X_n^2] = n$ and $var(X_n^2) = 2n$.

Chi-squared X_n^2
$E[X_n^2] = n$
$var(X_n^2) = 2n$

We will use the X^2 distribution extensively in the chapter on goodness-of-fit.

A confidence interval for a parameter is constructed for the purpose of substantiating an inference about the parameter from a set of experimental data. Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables, they each follow a probability distribution, and consequently it is possible to construct confidence intervals about them. All the formulas for β_0 and β_1 which we are about to present here pertain to a two parameter model. They are sufficient for our purposes to introduce the theory of linear models. However, more elaborate formulas involving matrices are required for general linear models. They are derived using the theory of linear algebra and will appear in subsequent statistics courses.

Theorem 7. For $\hat{\beta}_0$, the lower and upper limits of the $100(1-\alpha)\%$ confidence interval are:

$$a = \hat{\beta}_0 - \frac{st_{n-2, \frac{\alpha}{2}} \sqrt{\sum x_i^2}}{\sqrt{nSS_{xx}}}$$

$$b = \hat{\beta}_0 + \frac{st_{n-2, \frac{\alpha}{2}} \sqrt{\sum x_i^2}}{\sqrt{nSS_{xx}}}$$

where $s = \sqrt{\frac{SSE}{n-2}}$

Theorem 8. For $\hat{\beta}_1$, the lower and upper limits of the $100(1-\alpha)\%$ confidence interval are:

$$a = \hat{\beta}_1 - \frac{st_{n-2, \frac{\alpha}{2}}}{\sqrt{SS_{xx}}}$$

$$b = \hat{\beta}_1 + \frac{st_{n-2, \frac{\alpha}{2}}}{\sqrt{SS_{xx}}}$$

Example 4. Let us return to the example of measuring the temperature of water versus time after it was heated in a microwave oven found on page 6.

$\hat{\beta}_0 = 71.238$ and $\hat{\beta}_1 = 1.914$ so that $\widehat{E}[y_p] = 71.238 + 1.914x_p$.

Table 4

Time (sec)	Temperature ($^{\circ}F$)	$\widehat{E}[y_i]$	$\hat{\epsilon}_i = y_i - \hat{y}_i$	$\hat{\epsilon}_i^2 = (y_i - \hat{y}_i)^2$
0	70	71.238	-1.238	1.5326
5	81	80.810	+1.190	.0361
10	92	90.381	+1.619	2.6212
15	100	99.952	+0.048	.0023
20	110	109.524	+0.476	.2266
25	118	119.095	-1.095	1.1990
Total			0	5.6178

A tabulation of the predicted values and the residuals of the microwave model appear in Table 4. Since $\sum_{i=1}^6 \hat{\epsilon}_i = 0$, the arithmetic appears to be correct.

It is not enough to assert a model for explaining the data. It must be a good model; one which can be defended. Collecting data is difficult and expensive. The analysis of the data is easy with the use of a computer. Performing diagnostics of the model is technically very difficult. A common technique to assess the validity of a model is to make a plot of the residuals versus predicted values.

Definition 8. $\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is called a **residual** of the fitted model.

- If there is no pattern in the plot of the residuals versus predicted values like that appearing in Figure 4, then the ϵ_i 's are evidently random, and the plot therefore substantiates the assumption that ϵ_i represents random noise which $\epsilon_i \sim N(0, \sigma^2)$ implies.
- On the other hand, if there is a discernible pattern in the plot of the residuals versus predicted values, then the assumption that $\epsilon_i \sim N(0, \sigma^2)$ is not valid and the model is fundamentally flawed.

Plot of Residuals versus Predicted Values

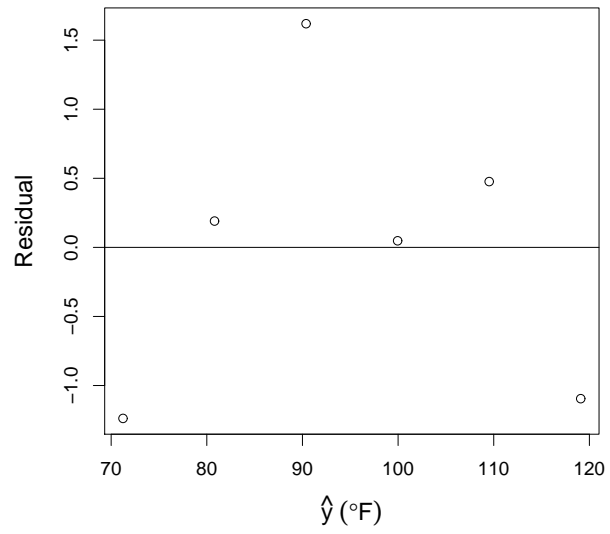


Figure 4

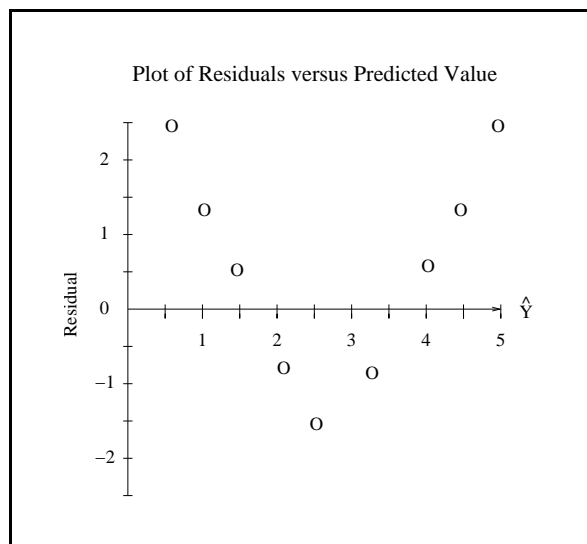


Figure 5

The plot of residuals versus predicted values for the microwave experiment displayed in Figure 4 shows a random pattern, as a result the proposed model shows promise. Three aspects of a model must be examined when determining if a model is good one.

1. The plot of the data must show a conspicuous linear relationship between x and y, i.e. $\beta_1 \neq 0$.
2. The plot of residuals versus predicted values must exhibit a random pattern.
3. The theoretical derivation of the model for describing the phenomenon of concern must be defensible.
4. Reject $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

Suppose a plot of residuals versus predicted values from a different experiment like the one shown in Figure 5 exhibits an obvious pattern, then we must conclude that whatever model is being proposed, it is seriously flawed.

Simply reporting values for $\hat{\beta}_0$ and $\hat{\beta}_1$ is not enough. They ought to be accompanied with their respective confidence intervals.

Problem 1. Find 95% CI about $\hat{\beta}_0$.

1. $\alpha = .05 \rightarrow \frac{\alpha}{2} = .025$ and $n=6$
2. $t_{n-2, \frac{\alpha}{2}} = t_{4, .025} = 2.776$
3. $\sum x_i^2 = 1375$ and $SS_{xx} = 437.5$ were already calculated.
4. $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{5.6178}{4} = 1.40445$ since $\sum \hat{\epsilon}_i^2 = 5.6178$
5. $a = \hat{\beta}_0 - \frac{st_{n-2, \frac{\alpha}{2}} \sqrt{\sum x_i^2}}{\sqrt{nSS_{xx}}} = 71.238 - \frac{1.185\sqrt{1375}}{\sqrt{6(437.5)}} 2.776 = 68.86$
6. $b = \hat{\beta}_0 + \frac{st_{n-2, \frac{\alpha}{2}} \sqrt{\sum x_i^2}}{\sqrt{nSS_{xx}}} = 71.238 + \frac{1.185\sqrt{1375}}{\sqrt{6(437.5)}} 2.776 = 73.62$
7. 95% CI for $\beta_0 = (68.86, 73.62)$

Problem 2. Find 95% CI about $\hat{\beta}_1$.

1. $\alpha = .05 \rightarrow \frac{\alpha}{2} = .025$ and $n=6$
2. $t_{n-2, \frac{\alpha}{2}} = t_{4, .025} = 2.776$
3. $\sum x_i^2 = 1375$ and $SS_{xx} = 437.5$ were already calculated.

$$4. \hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{5.6178}{4} = 1.40445 \text{ since } \sum \hat{\epsilon}_i^2 = 5.6178$$

$$5. a = \hat{\beta}_1 - \frac{st_{n-2, \frac{\alpha}{2}}}{\sqrt{SS_{xx}}} = 1.914 - \frac{1.185}{\sqrt{437.5}} 2.776 = 1.7567$$

$$6. b = \hat{\beta}_1 + \frac{st_{n-2, \frac{\alpha}{2}}}{\sqrt{SS_{xx}}} = 1.914 + \frac{1.185}{\sqrt{437.5}} 2.776 = 2.0713$$

$$7. 95\% \text{ CI for } \beta_1 = (1.7567, 2.0713)$$

Question 1. Is $0 \in (1.7567, 2.0713)$? No.

This question brings to mind the equivalence of confidence intervals and testing of hypotheses. It suggests the idea of testing hypotheses about $\hat{\beta}_0$ and $\hat{\beta}_1$.

Problem 3. Test $H_0 : \beta_0 = 71$ vs $H_1 : \beta_0 \neq 71$ at $\alpha = .05$

1. Is $71 \in 95\% \text{ CI for } \beta_0$ i.e. is $71 \in (68.86, 73.62)$?

2. Yes, therefore, cannot reject the null hypothesis.

Problem 4. Test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ at $\alpha = .05$

1. Is $0 \in 95\% \text{ CI for } \beta_1$ i.e. is $0 \in (1.7567, 2.0713)$?

2. No, therefore, reject the null hypothesis.

No one really cares about β_0 . In sharp contrast, however, the confirmation of whether or not $\beta_1 = 0$ is crucial. If the null hypothesis that $\beta_1 = 0$ cannot be rejected, then by implication, the proposition that there exists an inherent linear relationship of y with x is false. If $\beta_1 = 0$, then, no matter what x might be, y is unaffected. That is, no matter how long a microwave oven is used to heat a cup of water, the temperature of the water will always stay at room temperature. If $\beta_1 \neq 0$, then the microwave oven will be deemed useless. An experimenter is probably hoping that, according to his thesis, $\beta_1 \neq 0$ in order to demonstrate that there does indeed exist a linear relationship between the explanatory and response variables. The fact that $0 \notin 95\% \text{ CI for } \beta_1$ in the microwave oven experiment demonstrates that the temperature does depend on the length of time that the microwave oven is turned on.

Numerical computations are lengthy and tedious when estimating β_0 and β_1 , and for that reason among others, it is imperative to draw a picture of the data before commencing with the computations. Regardless of the theory and personal expectations, one must use common sense in evaluating a model. For instance, the predicted temperature of the water if it is heated in a microwave oven for 2 minutes (120 seconds) according to the fitted line is $71.283 + 1.914(120) = 301^\circ F$, but this is impossible. The lesson to be learned is that the method of least squares is a technique which comes from numerical analysis, and it is

best used in the problem of interpolating, not in extrapolating. The use of linear models is valid only within the scope of the data; otherwise, ridiculous results might be produced as in the case of our model which would make the preposterous prediction that after five days of heating, the final temperature of the water will be $826,929^{\circ}F$.

4 Analysis of Variance

The test of hypothesis, $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ at $\alpha = .05$ is a special case of testing the hypothesis of a general formulation of linear models in which there could be many more parameters besides the two which we have examined. Models could have $\beta_2, \beta_3, \dots, \beta_{r-1}$ as parameters so that the way to test the hypothesis that there is no linear effect must be answered by: $H_0 : \beta_1 = \beta_2 = \dots = \beta_{r-1} = 0$ vs. $H_1 : \text{at least one } \beta_k \neq 0$ where r is the number of parameters. Thus far, we have been dealing with linear models having two parameters, β_0 and β_1 , therefore $r=2$. For the more general model, a more sophisticated technique must be used to test the hypothesis. The implementation of this generalized technique is greatly facilitated by means of the analysis of variance table shown below.

The analysis of variance table shown in Table 5 is applicable to the analysis of variance for any linear model, whereas, in Table 6, the analysis of variance applies only to a two-parameter model, and that is the one which we will use.



Ronald Aylmer Fisher
1890-1962

Table 5: **Analysis of Variance for Fitting Regression for the General Linear Model**

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	$SSM = n\bar{y}^2$		
Regression	r-1	$SSR_{(m)} = \hat{\beta}'\mathbf{X}'\mathbf{Y} - n\bar{y}^2$	$MSSR = \frac{SSR_{(m)}}{r-1}$	$F = \frac{MSSR}{MSSE}$
Residual Error	n-r	$SSE = SST - SSM - SSR_{(m)}$	$\hat{\sigma}^2 = MSSE = \frac{SSE}{n-r}$	
Total	n	$SST = \sum_{i=1}^n y_i^2$		

Table 6: **Analysis of Variance for Fitting Regression for Two Parameters**

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	$SSM = n\bar{y}^2$		
Regression	1	$SSR_{(m)} = \hat{\beta}_1 SS_{xy}$	$MSSR = SSR_{(m)}$	$F = \frac{MSSR}{MSSE}$
Residual Error	n-2	$SSE = SST - SSM - SSR_{(m)}$	$\hat{\sigma}^2 = MSSE = \frac{SSE}{n-2}$	
Total	n	$SST = \sum_{i=1}^n y_i^2$		

The only substantial difference in these two tables is the entry for SSR. In the general case, SSR is expressed in terms of the design matrix, \mathbf{X} , and the vectors $\hat{\beta}$ and \mathbf{Y} . In the simple case in which there are only two parameters, the expression for SSR simplifies to $\hat{\beta}_1 SS_{xy}$ which can be easily computed even by hand.

Definition 9. $F_{r-1, n-r; \alpha}$ is a quantile of the F distribution with $r-1$ and $n-r$ degrees of freedom.

The criterion for testing the hypothesis, $H_0 : \beta_1 = \beta_2 = \dots = \beta_{r-1} = 0$ vs. $H_1 : \text{at least one } \beta_k \neq 0$ at the level of significance, α , appears in the accompanying table:

H_0	Test Statistic	H_1	Reject When
$\beta_1 = \beta_2 = \dots = \beta_{r-1} = 0$	F	$\beta_k \neq 0$	$F > F_{r-1, n-r; \alpha}$

For a two-parameter model, $r=2$ and the criterion for rejecting null hypothesis is:

H_0	Test Statistic	H_1	Reject When
$\beta_1 = 0$	F	$\beta_1 \neq 0$	$F > F_{1,n-2;\alpha}$

The F distribution has two degrees of freedom, ν_1 and ν_2 , and it is written as F_{ν_1, ν_2} . A tabulation of F quantiles for $\alpha = .05$ is given in Appendix 6.

The F distribution, like the X^2 distribution owes its origins to the Standard Normal distribution via the X^2 distribution. Specifically, if $X \sim X_{\nu_1}^2$ and $Y \sim X_{\nu_2}^2$, then $\frac{\left(\frac{X}{\nu_1}\right)}{\left(\frac{Y}{\nu_2}\right)} \sim F_{\nu_1, \nu_2}$. The probability density function for F is:

$$f(x) = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right) \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2} - 1}}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right) \left(1 + \frac{\nu_1}{\nu_2} x\right)^{\frac{\nu_1 + \nu_2}{2}}}$$

The F test statistic was named by George Snedecor in 1934 to honor Ronald Fisher.

F_{ν_1, ν_2}
$E[F_{\nu_1, \nu_2}] = \frac{\nu_2}{\nu_2 - 2}$
$var(F_{\nu_1, \nu_2}) = \frac{2\nu_2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 1)^2(\nu_2 - 4)}$

Example 5 (Analysis of Variance of the Model for the Microwave Data).

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	$6(95.1667)^2=54340.17$		
Regression	1	$1.914(837.5)=1603.215$	1603.215	F=1142
Residual Error	4	$55949-54340.17-1603.215=$ 5.615	$s^2 = 1.40375$	
Total	6	$\sum_{i=1}^6 y_i^2 = 55949$		

We demonstrated in Problem 4 on page 21 that because the confidence interval for β_1 did not contain zero, the hypothesis that $\beta_1 = 0$ was rejected. The same conclusion may be made by means of the ANOVA table.

1. $\alpha = .05$
2. $F_{1,4;.05} = 7.71$
3. Is $F = 1142 > 7.71$? Yes, therefore reject the null hypothesis.

The important quantity, s^2 , which was computed in Table 4 on page 18 is the same quantity as that which appears in the ANOVA table. The slight difference in these two is due to round-off error.

Definition 10. *Coefficient of determination is: $R^2 = 1 - \frac{SSE}{SS_{yy}}$.*

The coefficient of determination, R^2 , is a popular quantity made mentioned in meetings during which a critique of a model is often made. Someone will claim that because R^2 is high, his model is good, while another might complain that R^2 is not high enough to justify accepting the model. The diagnostics, the analysis of variance for testing the null hypothesis that $\beta_1 = 0$, the theoretical derivation of the model, and the plot of the data provide the means of making a good defense or a good attack on a proposed model. Although R^2 has some merits, as a general rule, it is better to avoid getting involved in a discussion in which R^2 is used to justify or discredit a model.

5 Simultaneous Comparison of Multiple Means

If two confidence intervals do not overlap, then we conclude that the means around which the confidence intervals were constructed are significantly different at a specified level of significance. Instead, if they do overlap, then, for practical purposes, the means of the populations are not different. When three or more means need to be compared, the problem

cannot be reduced to one of comparing the geometric position of multiple confidence intervals because we must enter higher dimensions where a geometric interpretation becomes infeasible.

Table 7: **Models Associated with Four Populations**

Population	Model
1	$y_{1j} = \beta_{01} + \epsilon_{1j}$ where $\epsilon_{1j} \sim N(0, \sigma_1^2)$
2	$y_{2j} = \beta_{02} + \epsilon_{2j}$ where $\epsilon_{2j} \sim N(0, \sigma_2^2)$
3	$y_{3j} = \beta_{03} + \epsilon_{3j}$ where $\epsilon_{3j} \sim N(0, \sigma_3^2)$
4	$y_{4j} = \beta_{04} + \epsilon_{4j}$ where $\epsilon_{4j} \sim N(0, \sigma_4^2)$

Instead a trick is used to make a simultaneous comparison of multiple means. The trick employs the method of least squares. The simple model, $y_j = \beta_0 + \beta_1 x_j + \epsilon_j$ where $\epsilon_j \sim N(0, \sigma^2)$, had been used to determine whether or not an inherent relationship between x and y exists. The answer to that question is determined by testing the hypothesis: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ at some level α . If the null hypothesis can be rejected, then we may conclude that $\beta_1 \neq 0$ and that an inherent relationship does exist between x and y . Suppose that the response variable does not change regardless of the value of x_i , then obviously the line is flat and hence $\beta_1 = 0$. In this case, the model simplifies to: $y_j = \beta_0 + \epsilon_j$ where $\epsilon_j \sim N(0, \sigma^2)$. The model now appears almost too trivial to be of any use, but, as we shall soon see, this model will be very useful.

To prevent any confusion in notation which might possibly occur later due to imprecise correspondences between models and their respective populations, we will introduce another subscript to the variables such that a model will be clearly associated with a certain population as in: $y_{1j} = \beta_{01} + \epsilon_{1j}$ where $\epsilon_{1j} \sim N(0, \sigma_1^2)$ for population 1. Similarly, the same formulation will be applied to other distinct populations like those listed in Table 7.

Before we simplify the notation any further, we should observe that the symbol, β_0 , appears in each model, and we should note that when $\beta_1 = 0$, the formula, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, for estimating $\hat{\beta}_0$ reduces to $\hat{\beta}_0 = \bar{y}$, where $\bar{y} = \frac{\sum_{j=1}^n y_j}{n}$. If measurements are made on subjects of population 1, then this formula should be associated with population 1 and that is done

by incorporating another subscript into the name such as $\hat{\beta}_{01} = \bar{y}_1 = \frac{\sum_{j=1}^{n_1} y_{1j}}{n_1} = \frac{y_{1.}}{n_1}$. The dot which appears in the subscript of \bar{y}_1 , signifies that the sum has been taken over the second index, j . The pattern can be extended to two indices as in $y_{..}$ which would be interpreted as,

Table 8: Analysis of Variance

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	$SSM = n\bar{y}_{..}^2$		
Regression	k-1	$SSR = \sum_{i=1}^k n_i(\bar{y}_i - \bar{y}_{..})^2$	$MSSR = \frac{SSR}{k-1}$	$F = \frac{MSSR}{MSSE}$
Residual Error	$\sum_{i=1}^k n_i - k$	$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$\widehat{\sigma}^2 = MSSE = \frac{SSE}{n-k}$	
Total	$\sum_{i=1}^k n_i$	$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2$		

$n = \sum_{i=1}^k n_i$ and n_i is the size of the sample which was drawn from population i, and k is the number of populations. H_0 is rejected if $F > F_{k-1, n-k, \alpha}$.

$y_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$, and $y_{...} = \sum_{i=1}^k \sum_{j=1}^{n_i} \sum_{l=1}^{m_i} y_{ijkl}$, and so on. \bar{y}_1 is nothing more than the sample mean from population 1. If \bar{y}_1 denotes the sample of population 1, then μ_1 will denote the population mean for population 1. With $\widehat{\beta}_{01} = \bar{y}_1$ in the sample case, it would appear that $\beta_{01} = \mu_1$ in the case of the population. We have been using μ to denote the population mean; therefore, it is appropriate to replace β_0 with μ and write the generic model for population, i, and observation, j, as, $y_{ij} = \mu_j + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma_i^2)$. The sample mean for population i is $\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$; the grand sample mean is denoted by $\bar{y}_{..} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^k n_i}$.

Table 9: Data for Job Satisfaction

Population	Education	Job Satisfaction	n_i	$\sum_{j=1}^{n_i} y_{ij}$	\bar{y}_i
1	Ph.D.	6 5 2 1 7 7 6	7	34	4.85
2	Masters	7 7 5 6 6 5 3 5 3 6 6 5 7 6 6 7 6 6 7 6	20	115	5.75
3	College	6 7 7 5 6 7 7 7 7 6 6 1 6 6	14	84	6
4	High School	5 6 1 5 5 6 7 1 4	9	40	4.44
Total			50	273	5.46

The problem of determining whether $\mu_1 = \mu_2 = \mu_3 = \mu_4$ can now be framed in terms of a linear model by which we may test the hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs $H_1 : otherwise$ for a specified α . By first referring to the ANOVA table for the general case, a simplified ANOVA table for comparing four means will be constructed. ANOVA Table 8 shows the formulas for the general case when testing the hypothesis $H_0 : \mu_1 = \dots = \mu_k$ vs $H_1 : otherwise$ at a specified level of α .

Example 6. *A clinical sociologist wants to know if there is a difference in job satisfaction according to level of education. The accompanying tabulation of data was taken from a survey of 50 people of both sexes in Washington, D.C.. Given the data for job satisfaction (1=lowest and 7=highest) and level of education in Table 9. The sociologist wants to test the hypothesis*

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ vs } H_1 : otherwise$$

at a level of significance $\alpha = .05$.

To that end, we note that $\bar{y}_{..} = \frac{273}{50} = 5.46$ and compute the various sums of squares.

$$SST = 6^2 + 5^2 + 2^2 + \dots + 7^2 + 1^2 + 4^2 = 1637$$

$$SSM = 50(5.46)^2 = 1490.58$$

$$SSR = 7(4.85 - 5.46)^2 + 20(5.75 - 5.46)^2 + 14(6 - 5.46)^2 + 9(4.44 - 5.46)^2 \approx 17.59$$

$$SSE = SST - SSM - SSR \text{ or}$$

$$SSE = (6 - 4.85)^2 + (5 - 4.85)^2 + (2 - 4.85)^2 + \dots + (7 - 4.44)^2 + (1 - 4.44)^2 + (4 - 4.44)^2 \approx 128.82$$

1. $\alpha = .05$
2. $F_{3,46,.05} = 2.806845$ From a computer.
3. $F = 2.0936$ from the ANOVA table.
4. Is $2.0936 > 2.806845$? No
5. Cannot reject null hypothesis. According to the data, the means of the four populations are, for all practical purposes, the same. Based on the sociologist's survey, there appears to be no difference in job satisfaction according to education.

The previous example in which $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs $H_1 : otherwise$ was tested for a specified α should remind us of the problem of comparing two means which was addressed in the discussion beginning on page 15 of *Testing Hypotheses*. In that chapter,

the T test statistic was used in the test of the hypotheses:

$H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.

$$T = \frac{\bar{x}_2 - \bar{x}_1}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

Table 10: **Analysis of Variance**

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	1490.58		
Regression	4-1=3	17.5906	5.863533	2.0936
Residual Error	50-4=46	128.8294	2.80064	
Total	50	1637		

This expression for the T test statistic will be shown to come from the Analysis of Variance of a linear model. It will be shown that the F test statistic found in the ANOVA table is equivalent to the T test statistic which was used in testing the difference of two means. Assume that X_{1j} are i.i.d., $X_{1j} \sim N(\mu_1, \sigma_1^2)$, X_{2j} are i.i.d., and $X_{2j} \sim N(\mu_2, \sigma_2^2)$. The criterion for testing the hypothesis regarding the difference of two means which was discussed in *Testing Hypothesis*, the formula for the T test statistics was merely presented without any explanation as to its origins. This time, we will derive the test for comparing two means by formulating the problem as a linear model:

$$\begin{aligned} y_{1j} &= \mu_1 + \epsilon_{1j} & \text{where } \epsilon_{1j} &\sim N(0, \sigma_1^2) \\ y_{2k} &= \mu_2 + \epsilon_{2k} & \text{where } \epsilon_{2k} &\sim N(0, \sigma_2^2) \end{aligned} \quad (1)$$

where j denotes any one of the observations from population 1 and k denotes any one of the observations from population 2. The number of y_{1j} 's might be different from the number of y_{2k} 's. Also, the ϵ_{1j} 's follow a different distribution from the ϵ_{2k} 's.

In as much as the definitions of y_{1j} and y_{2k} seem to bear little resemblance to X_{1j} and X_{2j} , their respective distributions are the same. To show this we will rely on the properties of the Normal distribution. It is easy to show that y_{1j} is the same as the random variables X_{1j} used in *Testing Hypotheses*. In line with the first property which states that if $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$, we will assign $\epsilon_{1j} \sim N(0, \sigma_1^2)$, $a=1$, and $b = \mu_1$, from which we may conclude that the distribution of $\epsilon_{1j} + \mu_1 \sim N(\mu_1, \sigma_1^2)$ which

is the same distribution of X_{1j} . The formulation of the problem of comparing two means by a linear model is, therefore, justified as claimed.

We will modify the entries of the Analysis of Variance Table 8 accordingly by reducing the number of means from four to two.

$$SSM = n\bar{y}_{..}^2 = (n_1 + n_2) \left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} \right)^2$$

$$SSR = \sum_{i=1}^k n_i(\bar{y}_i - \bar{y}_{..})^2 = n_1(\bar{y}_1 - \bar{y}_{..})^2 + n_2(\bar{y}_2 - \bar{y}_{..})^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{k=1}^{n_2} (y_{2k} - \bar{y}_2)^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$$

With additional algebraic simplifications, all the other sums of squares are shown in the ANOVA Table 11 for the model given in equation (1).

Table 11: **Analysis of Variance**

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	$(n_1 + n_2) \left(\frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} \right)^2$		
Regression	1	$\frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}$	$\frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{1}{n_1} + \frac{1}{n_2}}$	$\frac{(\bar{y}_1 - \bar{y}_2)^2}{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$
Residual Error	$n_1 + n_2 - 2$	$(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$	$\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = S_p^2$	
Total	$n_1 + n_2$	$SST = \sum_{j=1}^{n_1} y_{1j}^2 + \sum_{j=1}^{n_2} y_{2j}^2$		

Since the symbol, y , is arbitrary and, in order to give the equations a look which will agree with our customary use of notation, the symbol, x , will be used instead of y . To draw the equivalence between the test obtained from the linear model approach as expressed in the ANOVA Table 11 and the test of two means given on page 17 of *Testing Hypotheses* to completion, we invoke a relation between the F and the Student's t distributions. From an advanced course of mathematical statistics it can be proved that, $F_{1, \nu_2, \alpha} = (t_{\nu_2, \frac{\alpha}{2}})^2$; therefore, the F test statistic which appears in the ANOVA table can be written as: $T^2 = F = \frac{(\bar{x}_1 - \bar{x}_2)^2}{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$. By taking the square root, $T = \frac{(\bar{x}_1 - \bar{x}_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ which exactly is the same test statistic that was used in *Testing Hypotheses*. Besides showing the origins of the test statistic, the ANOVA table, also, shows the origins of the formula for S_p^2 , the degrees of freedom, and the peculiar factor $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

6 Paired Difference Test

In another illustration of employing the method of least squares for providing a concise derivation of a statistic, we will examine the origins of the paired difference test. The paired difference test and the test for comparing the means of two populations appear to be very similar. The similarity between them is close enough to cause confusion. When should a paired difference test be used or the test for comparing two means be used is a common question. The dilemma in deciding which test to use can be traced to the design of the experiment from which the set of data was obtained.

The population consists of those things which possess the information which is being sought. The definition of a population is completely arbitrary and is made according to the interests of the experimenter. For example, he might be interested in verifying that a treatment of his own design in curing a disease is significantly more effective on an element of the population than if the subject received no treatment at all. The defining characteristic for a paired difference test is that a subject of a population is measured twice; once before the treatment is administered and again afterwards. For example, consider the population of middle aged women who reside in the same affluent suburb in America and who wish to loose weight. An experiment could be designed by which eight women are selected at random and enrolled in an aerobics class under the presumption that supervised exercise will result in a healthy weight loss. If loss of weight is the matter of concern, then it would be natural to expect that they ought to be weighed before and again at the conclusion of the class. In this experiment, the weight of each woman is to be measured twice, in order to determine if the treatment of doing disciplined aerobic exercises under the direction of an instructor will in fact reduce her weight.

Table 12: **Weights of Women Before and After Aerobic Exercise Program**

ID	Before	After	d
1	198	194	4
2	154	151	3
3	124	126	-2
4	110	104	6
5	127	123	4
6	162	155	7
7	141	129	12
8	180	165	15
Mean	149.5	143.4	6.125

A picture of the data is helpful in seeing the design of the experiment. There is a pair of measurements for each subject. In most of the cases as illustrated in Figure 6, the weight of a participant before the exercise program, B, is usually greater than the weight at the completion of the program, A. It is suggested in the picture that by engaging in the exercise program, a loss of weight does occur.

Intuition will lead one to suspect that if the regimen is successful, then each woman will observe a loss of weight. The researcher realizes that some losses will be more than others, but, taking that into account, he will want to demonstrate that differences in weights before and after the class are not due to mere chance but are actually a product of exercising. Each difference can be associated with a random variable, d_i . Reminiscent of the discussion of paired difference test found on page 14 of *Testing Hypotheses*, the researcher will undoubtedly test the hypothesis: $H_0 : \mu_d = 0$ vs $H_1 : \mu_d \neq 0$ at a level of significance α . Based on the data, he will compute the T test statistic, $T = \frac{\bar{d}}{\frac{s}{\sqrt{n}}}$ and compare it against the appropriate t quantile.

When the second picture, Figure 7, is examined, one sees at once that the test for comparing two means necessarily includes information about the distribution of all weights. The horizontal solid line represents the average of the weights which were taken before the aerobics course, and the dashed horizontal line is the average of the weights which were taken after completing the course. Both averages include every one's weight: those who were light as well as those who were heavy. Even though the set of weights appears as two halves with each half having a mean, the weights ranging from heavy women to

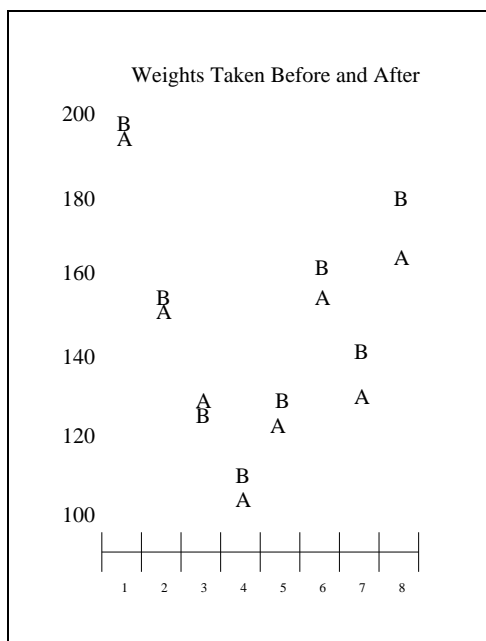


Figure 6

light women are intermingled in each mean, as a result, it is impossible to disassociate the distribution of weights of the whole group of women from any possible effect which the aerobic exercising might have had on a particular subject. By formulating the problem in such a way as to study paired differences, on the other hand, it is possible to make the differences independent of the other the subjects and to ignore the distribution of weights which would otherwise introduce undesirable confounding of information. The second picture, also, shows that the measurements are taken from subjects not of two populations but of only one population. In the design of an experiment for paired differences, all measurements are taken from subjects of the same population. Each individual subject is measured once before a treatment is applied and again after the application of the treatment; therefore, the test for comparing the means of two populations is actually inappropriate when ascertaining the effectiveness of a treatment on a population. It is appropriate for the researcher, in this example, to design the experiment with a paired difference test in mind.

To arrive at a test statistics for a paired difference test, the following linear model is used:

$$y_i = \mu_1 - \mu_2 + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2) \quad (2)$$

With this model as a basis, an experiment is designed such that a subject is measured twice and that all other subjects which are examined experience similar conditions. If, for

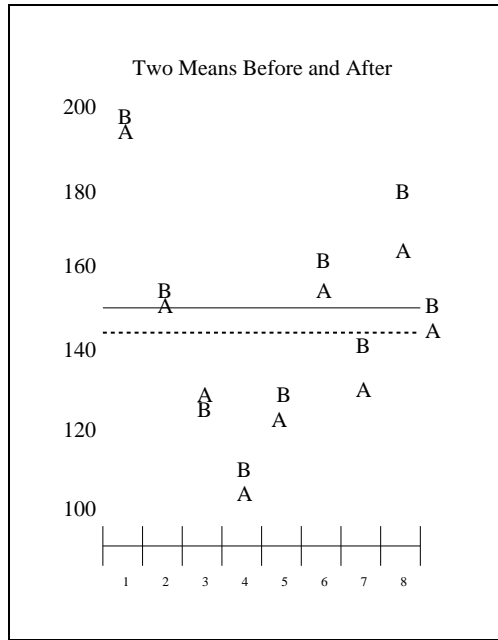


Figure 7

example, in the aerobics class, one woman had lost weight as a result of having underwent a serious surgical procedure, a well designed experiment would not include her in the sample because the condition which caused her loss of weight is not shared by the other women of the class.

The linear model which is expressed in equation (2) was conceived for the purpose of examining paired differences. It cannot be used to compare the means of two populations. Conversely, the linear model given by equation (1) for comparing two means cannot be used to examine pair differences. According to the model for comparing two means, there are two different populations, as implied by two differently distributed ϵ 's, however, in the paired difference case, there is only one population, one distribution for ϵ , and one sample in which each element is measured twice: once before a treatment has been administered and again afterwards. It should be noted that there is no x term in the model, consequently there is no β_1 . That is, β_1 is identically zero at the very beginning, and there will not be a row in the ANOVA table for it.

For testing the hypothesis: $H_0 : \mu_D = 0$ vs $H_1 : \mu_D \neq 0$ where $\mu_D = (\mu_1 - \mu_2)$ at a level of α , the ANOVA table given in Table 8, reduces to the following:

Table 13: Analysis of Variance of a Linear Model for Paired Differences

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	$SSM = n\bar{d}^2$	$MSSM = n\bar{d}^2$	$F = \frac{n\bar{d}^2}{s^2} = \frac{\bar{d}^2}{\frac{s^2}{n}}$
Residual Error	n-1	$SSE = \sum_{i=1}^n (d_i - \bar{d})^2$	$MSSE = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1} = s^2$	
Total	n	$SST = \sum_{i=1}^n d_{ij}^2$		

It was noted before that co-incidentally, $F_{1,\nu_2,\alpha} = (t_{\nu_2,\frac{\alpha}{2}})^2$ and, as a result, the F test statistic which appears in the ANOVA table for the paired difference experiment can be written in terms of T, that is, $T^2 = F = \frac{\bar{d}^2}{\frac{s^2}{n}}$. By taking the square root, the T test statistic which was derived from the ANOVA table for a paired difference experiment is exactly the same test statistic for paired differences which was used in section of paired difference test found in *Testing Hypotheses*.

According to the set of data regarding the differences in the weights of eight women who took an aerobics exercise class, the leader of the class wants to verify that the course did indeed bring about a loss of weight for a participant, and she wants the verification to be wrong with a probability no greater than .05. From the set of eight differences in weights, the analysis of the data produced the ANOVA Table 14

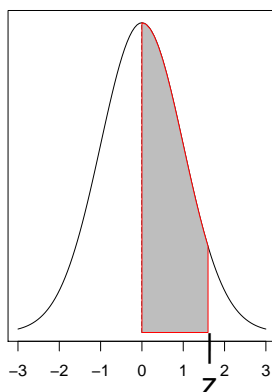
Table 14: Analysis of Variance for Paired Differences in Weights

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F statistic
Mean	1	$8(6.125)^2=300.125$	300.125	$F = \frac{300.125}{28.41071} = 10.56380$
Residual Error	7	198.875	28.41071	
Total	8	499		

Since $F = 10.56380 > F_{1,7,.05} = 5.59145$, the null hypothesis that no loss of weight occurred can be rejected at a level of significance, $\alpha = .05$. When the criterion for rejecting the null hypothesis which is derived from the linear model formulation of paired differences is compared to the corresponding criterion which was used in discussion of the paired difference test presented in *Testing Hypotheses*, both arrive at the same conclusions because the criteria are actually the same. For the paired difference test, $T = \frac{\bar{d}}{\frac{s}{\sqrt{n}}} = \frac{6.125}{\sqrt{\frac{28.4107}{8}}} = 3.250199$ and $t_{7,.025} = 2.36462$ and since $T = 3.250199 > t_{7,.025} = 2.36462$, $H_0 : \mu_d = 0$

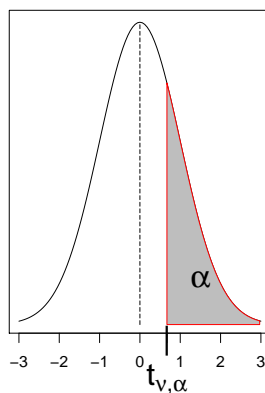
can be rejected at a level of significance, $\alpha = .05$. The criteria are the same because $T^2 = 3.250199^2 = 10.56380 = F$ and $t_{7,.025}^2 = 2.36462^2 = 5.5914 = F_{1,7,.05}$.

The formulation of the comparison of two means and the formulation of paired differences in terms of linear models illustrate the versatility of the method of least squares, and they help illustrate the reason why linear models enjoy such immense popularity. By means of linear models, it is possible to solve many problems more easily. A case in point is the preceding study of the paired difference test and the comparison of means. By using linear models, it is easy to see that the two tests correspond to two fundamentally different situations.



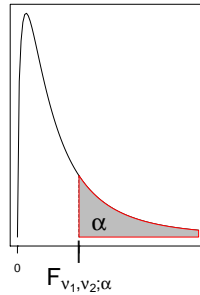
Cumulative Probabilities for a $N(0,1)$ Distribution: $\Phi(z) - .5$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.0279	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900



Quantiles for a Student's t Distribution

ν	$t_{\nu,.20}$	$t_{\nu,.15}$	$t_{\nu,.10}$	$t_{\nu,.05}$	$t_{\nu,.025}$	$t_{\nu,.01}$	$t_{\nu,.005}$
1	1.37638	1.96261	3.07768	6.31375	12.7062	31.82052	63.65674
2	1.06066	1.38621	1.88562	2.91999	4.30265	6.964560	9.92484
3	0.97847	1.24978	1.63775	2.35338	3.18245	4.54070	5.84091
4	0.94096	1.18957	1.53321	2.13185	2.77645	3.74695	4.60410
5	0.91954	1.15577	1.47588	2.01505	2.57058	3.36493	4.03216
6	0.90570	1.13416	1.43976	1.94318	2.44691	3.14267	3.70743
7	0.89603	1.11916	1.41492	1.89458	2.36462	2.99795	3.49948
8	0.88889	1.10815	1.39682	1.85955	2.30600	2.89646	3.35539
9	0.88340	1.09972	1.38303	1.83311	2.26216	2.82144	3.24984
10	0.87906	1.09306	1.37218	1.81246	2.22814	2.76377	3.16927
11	0.87553	1.08767	1.36343	1.79588	2.20099	2.71808	3.10581
12	0.87261	1.08321	1.35622	1.78229	2.17881	2.68100	3.05454
13	0.87015	1.07947	1.35017	1.77093	2.16037	2.65031	3.01228
14	0.86805	1.07628	1.34503	1.76131	2.14479	2.62449	2.97684
15	0.86624	1.07353	1.34061	1.75305	2.13145	2.60248	2.94671
16	0.86467	1.07114	1.33676	1.74588	2.11991	2.58349	2.92078
17	0.86328	1.06903	1.33338	1.73961	2.10982	2.56693	2.89823
18	0.86205	1.06717	1.33039	1.73406	2.10092	2.55238	2.87844
19	0.86095	1.06551	1.32773	1.72913	2.09302	2.53948	2.86093
20	0.85996	1.06402	1.32534	1.72472	2.08596	2.52798	2.84534
21	0.85907	1.06267	1.32319	1.72074	2.07961	2.51765	2.83136
22	0.85827	1.06145	1.32124	1.71714	2.07387	2.50832	2.81876
23	0.85753	1.06034	1.31946	1.71387	2.06866	2.49987	2.80734
24	0.85686	1.05932	1.31784	1.71088	2.06390	2.49216	2.79694
25	0.85624	1.05838	1.31635	1.70814	2.05954	2.48511	2.78744
26	0.85567	1.05752	1.31497	1.70562	2.05553	2.47863	2.77871
27	0.85514	1.05673	1.31370	1.70329	2.05183	2.47266	2.77068
28	0.85465	1.05599	1.31253	1.70113	2.04841	2.46714	2.76326
29	0.85419	1.05530	1.31143	1.69913	2.04523	2.46202	2.75639
30	0.85377	1.05466	1.31042	1.69726	2.04227	2.45726	2.75000
40	0.85070	1.05005	1.30308	1.68385	2.02108	2.42326	2.70446
50	0.84887	1.04729	1.29871	1.67591	2.00856	2.40327	2.67779
75	0.84644	1.04365	1.29294	1.66543	1.99210	2.37710	2.64298
100	0.84523	1.04184	1.29007	1.66023	1.98397	2.36422	2.62589
150	0.84402	1.04003	1.28722	1.65508	1.97591	2.35146	2.60900
∞	0.84162	1.03643	1.28155	1.64485	1.95996	2.32635	2.57583



Quantiles for an F_{1, ν_2} Distribution

ν_2	$F_{1, \nu_2; .20}$	$F_{1, \nu_2; .15}$	$F_{1, \nu_2; .10}$	$F_{1, \nu_2; .05}$	$F_{1, \nu_2; .025}$	$F_{1, \nu_2; .01}$	$F_{1, \nu_2; .005}$
1	9.47214	17.34972	39.86346	161.44764	647.78901	4052.1807	16210.72272
2	3.55556	5.20721	8.52632	18.51282	38.50633	98.50251	198.50125
3	2.68221	3.70301	5.53832	10.12796	17.44344	34.11622	55.55196
4	2.35072	3.16197	4.54477	7.70865	12.21786	21.19769	31.33277
5	2.17823	2.88783	4.06042	6.60789	10.00698	16.25818	22.78478
6	2.07290	2.72307	3.77595	5.98738	8.81310	13.74502	18.63500
7	2.00201	2.61337	3.58943	5.59145	8.07267	12.24638	16.23556
8	1.95109	2.53517	3.45792	5.31766	7.57088	11.25862	14.68820
9	1.91277	2.47664	3.36030	5.11736	7.20928	10.56143	13.61361
10	1.88289	2.43122	3.28502	4.96460	6.93673	10.04429	12.82647
11	1.85894	2.39494	3.22520	4.84434	6.72413	9.64603	12.22631
12	1.83933	2.36531	3.17655	4.74723	6.55377	9.33021	11.75423
13	1.82296	2.34065	3.13621	4.66719	6.41425	9.07381	11.37354
14	1.80911	2.31982	3.10221	4.60011	6.29794	8.86159	11.06025
15	1.79722	2.30198	3.07319	4.54308	6.19950	8.68312	10.79805
16	1.78692	2.28654	3.04811	4.49400	6.11513	8.53097	10.57546
17	1.77790	2.27304	3.02623	4.45132	6.04201	8.39974	10.38418
18	1.76994	2.26114	3.00698	4.41387	5.97805	8.28542	10.21809
19	1.76286	2.25057	2.98990	4.38075	5.92163	8.18495	10.07253
20	1.75653	2.24112	2.97465	4.35124	5.87149	8.09596	9.94393
21	1.75083	2.23262	2.96096	4.32479	5.82665	8.01660	9.82952
22	1.74567	2.22493	2.94858	4.30095	5.78630	7.94539	9.72706
23	1.74098	2.21795	2.93736	4.27934	5.74980	7.88113	9.63480
24	1.73669	2.21157	2.92712	4.25968	5.71664	7.82287	9.55127
25	1.73276	2.20573	2.91774	4.24170	5.68637	7.76980	9.47531
26	1.72915	2.20036	2.90913	4.22520	5.65862	7.72125	9.40593
27	1.72582	2.19541	2.90119	4.21001	5.63311	7.67668	9.34232
28	1.72273	2.19082	2.89385	4.19597	5.60956	7.63562	9.28377
29	1.71986	2.18657	2.88703	4.18296	5.58777	7.59766	9.22973
30	1.71719	2.18261	2.88069	4.17088	5.56753	7.56248	9.17968
40	1.69801	2.15420	2.83535	4.08475	5.42394	7.31410	8.82786
50	1.68666	2.13743	2.80866	4.03431	5.34032	7.17058	8.62576
60	1.67915	2.12635	2.79107	4.00119	5.28561	7.07711	8.49462
70	1.67382	2.11849	2.77860	3.97778	5.24703	7.01140	8.40266
80	1.66984	2.11263	2.76931	3.96035	5.21835	6.96269	8.33461
90	1.66676	2.10808	2.76211	3.94688	5.19621	6.92514	8.28222
100	1.66429	2.10446	2.75638	3.93614	5.17859	6.89530	8.24064
150	1.65694	2.09364	2.73927	3.90420	5.12626	6.80689	8.11767
200	1.65328	2.08826	2.73078	3.88837	5.10038	6.76330	8.05716