APPLIED ECONOMETRICS

Potluri Rao Roger LeRoy Miller

University of Washington

Wadsworth Publishing Company, Inc., Belmont, California

1971

Chapter 6

Hypothesis Testing

Empirical research is often called for in testing a verifiable statement. When faced by several theories, the researcher interested in empirically testing which of them is appropriate in a given situation will apply the tools of Hypothesis Testing.

To understand the relevant concepts and their proper use, let us consider a situation in which there are two theories. The researcher wants to choose the appropriate one for the policy purposes of his given situation. There may be several ways of deciding which of the theories is appropriate. One is by the criterion of empirical relevance. If the real world, as observed by measured facts, proves to be inconsistent with a theory, then the researcher may decide to discard the theory.

To make effective use of empirical investigation in this way, one must be able to distinguish one theory from the other empirically. If both predict the same observable phenomena, there is no way of distinguishing between them; observed relations may lead to the conclusion that both theories are appropriate or both inappropriate. Reaching such a conclusion is simply a redundant exercise.

Only when the two theories disagree on an observable relation can empirical research determine the appropriateness of a given theory. Typically, the researcher should construct a statement regarding some observable phenomenon in such a way that one theory implies that the statement is false and the other implies it is true. By empirical verification the researcher can then determine which of the two contesting theories is appropriate to his needs.

When the testing of two theories has been thus reduced, the statement is rewritten in the terminology of statistics as two hypotheses: one declaring that the statement is true, referred to as the Null hypothesis, and the other declaring that the statement is false, referred to as the Alternative hypothesis. In this terminology only one of the two hypotheses can be true in a given situation. Prior to any testing the researcher must have both the Null hypothesis and the alternative: hence the starting point in any test is a clear statement of the contrasting hypotheses and their relation to the corresponding theories. The researcher has need to test these as his only way of knowing which is the truth.

Having clearly stated the objective of his empirical investigation in the form of the Null and Alternative hypotheses, the researcher now turns to a test criterion—that is, how he intends to verify which of the two statements is the truth. He may set up a rule in such a way that if its terms are met by the data then he will accept the Null hypothesis as the truth, otherwise he will accept the Alternative.

When a rule is based on a statistic it may possibly give a wrong answer. When the Null hypothesis is in fact true, the rule may indicate that it is false, which is clearly a wrong answer. This is called the *Type I error*. Though the researcher would like to have a rule that does not give wrong answers, it is hard to find one which always performs correctly. However, one may select a rule that gives right answers more frequently than others. By choosing such a rule the researcher can be confident that if it is used repeatedly on different occasions he will at least in a majority of the cases be getting correct answers, even though he does not know when he is getting wrong answers.

Any rule may give a wrong answer of the opposite kind: when the Null hypothesis is actually false the rule may indicate that it is true. This is called the *Type II error*. Notice that the Type I error and the Type II error are different concepts. Any rule may give both kinds of wrong answers. Ideally, the researcher would prefer a rule that has the minimum chance of committing both of these errors. But unfortunately rules with a smaller probability of Type I error usually have a larger probability of Type II error and vice versa.

The problem of choosing a rule on the basis of *both* Type I and Type II errors is infrequent in applied econometrics, because we seldom know the probabilities of both of these two errors corresponding to any rule. It is true that their probabilities may be computed on the basis of the theoretical distributions of the estimates on which the rule is based, provided the Null and the Alternative hypotheses clearly state specific values of the corresponding parameters. But usually in applied econometrics only the Null hypothesis is specific, and the Alternative is usually the negation of the Null hypothesis with no specific values assigned to the respective parameters. In such cases it is possible to obtain the probability of the Type I error but not of the Type II error.

When the researcher rejects a Null hypothesis having a probability of Type I error of, say, 5 percent he is aware that his rejection on the basis of that rule will be wrong 5 percent of the time. Conversely, he is confident that he will get the correct answer in 95 percent of the cases. In other words, the rule is said to have a "95 percent level of confidence."

Unless the researcher knows the probability of the corresponding Type II error of a rule, he does not know the probability of getting a right answer when the rule indicates that the Null hypothesis should be accepted. It may be that the Null hypothesis is in fact false. That is, without knowing the probability of the Type II error the researcher cannot assess the chances of his getting the correct answer when the rule indicates acceptance of a Null hypothesis.

Since the Alternative hypothesis is usually the negation of the Null hypothesis rather than a statement regarding the specific values of parameters, the researcher knows the chances of obtaining a wrong answer when he rejects the Null hypothesis but not when he accepts. To maintain this distinction in reporting results, the applied econometrician either “rejects” or “does not reject” the Null hypothesis, rather than “rejecting” or “accepting” it. Thus “not rejecting” a Null hypothesis does not necessarily imply its acceptance.

The rule for rejecting or not rejecting a Null hypothesis on the basis of empirical research is usually based on some test statistic (call it *t*) computed from the data. Typically, a rule rejects the Null hypothesis when a test statistic exceeds a specified value, t_c , called the critical value. When the statistical distribution of the test statistic is known, then the researcher can compute the probability of the statistic exceeding the critical value when the Null hypothesis is true, which is the probability of committing the Type I error corresponding to the rule.

6.1 Test Based on One Regression coefficient

When the researcher wants to verify whether a theory is relevant to a given situation, he will want to test a Null hypothesis on the basis of data relating to the situation. Sometimes it is possible to formulate the Null hypothesis in the form of a specific value of a parameter. For example, consider a theory which states that, given the level of profits (P), investment (I) in an industry does not change with the sales (S) of that industry. This Null hypothesis may be readily translated into a specific value of a parameter by expressing the relation between investment and profits as a linear regression equation:

$$I_t = \beta_0 + \beta_1 S_t + \beta_2 P_t + \varepsilon_t . \qquad (6.1)$$

The Null hypothesis implies that the parameter value of β_1 is zero. The Alternative hypothesis may be stated simply as: the Null hypothesis is false. In the standard notation the Null hypothesis is expressed as

$$H_N: \beta_1 = 0 \qquad H_A: H_N \text{ is false.} \qquad (6.2)$$

Since specification of the objective of a test is complete only when the Null and Alternative hypotheses have been clearly stated, expression (6.2) is usually referred to as the *Null hypothesis* instead of the more accurate term, *Null and Alternative hypotheses*.

In this particular example, the rule to test the Null hypothesis against the Alternative is based on a statistic called the *t*-ratio, defined as

$$t = \hat{\beta}_1 \text{ / st. error of } \hat{\beta}_1. \qquad (6.3)$$

Since the estimate $\hat{\beta}_1$ and its standard error are obtained from the results of the least squares procedure, the *t*-ratio has a statistical distribution. When the Null hypothesis is true (β_1 = 0) and the error terms ($\varepsilon 's$) are generated by a normal distribution, the *t*-ratio follows the “Student’s *t*” distribution, with (T - K) degrees of freedom, where T is the number of observations and K is the number of parameters estimated (including the constant).

The theoretical distribution of the *t*-ratio provides the probability that this statistic will exceed a specified value, say, t_c, When the Null hypothesis is true, the probability of the statistic *t*-ratio exceeding a set critical value t_c , for various numbers of degrees of freedom, is furnished in most textbook.

Since the distributional properties of the *t*-ratio are known, we may set up the rule for testing the Null hypothesis on the basis of this statistic. The rule may be set so that whenever the *t*-ratio exceeds a set critical value, t_c, the Null hypothesis is rejected and not otherwise. Under this rule, the probability of a Type I error is solely the probability of *t* exceeding the value t_c, when the Null hypothesis is actually true. The critical value t_c, may be chosen so that the level of confidence associated with it is acceptable.

Note that the researcher cannot forever increase the value of t_c to reach higher levels of confidence, because by so doing he would be increasing the probability of committing the Type II error associated with the test.

The Null hypothesis is false when (β_1 > 0) and also when (β_1 < 0). When the Alternative hypothesis, (β_1 > 0), is in fact true, we expect the *t* statistic to be positive; hence, the rule would be: “Reject the Null hypothesis when *t* exceeds the critical value t_c” where *t*, is a positive quantity. Testing a Null hypothesis against an Alternative that assigns values to the parameter which are greater than the value implied by the Null hypothesis—for example in (6.2), ($\beta_1 > 0$)—is called the right-tail test. Similarly, when the alternative hypothesis, ($\beta_1 < 0$), is in fact true, the *t* statistic will be negative and it will be inappropriate to use the rule of *t* exceeding a positive quantity. When *t* is negative and large in magnitude, it offers evidence in favor of the Alternative. Therefore, the test rule for the Alternative hypothesis ($\beta_1 < 0$) should be different from that of the alternative ($\beta_1 > 0$).

When the Alternative assigns values to the parameter which are less than the parameter value implied by the Null hypothesis, the test is called the *left-tail* test; the rule is, then, "Reject the Null hypothesis when t is smaller than t_c" where t_c is a negative quantity.

When the Alternative hypothesis includes both the right tail and left tail tests, both these rules should be applied; hence, the test rule may be stated as, "Reject the Null hypothesis when the absolute value of the *t* statistic exceeds the value t_c in magnitude." Such a test is called the *two-tail* test.

Note that the probability of committing a Type I error in the case of a two-tail test is twice as much as in either of the one-tail tests for a given critical value t_c.

In our example (6.1), (6.2) we shall choose a 95 percent confidence level as acceptable. The critical value corresponding to 12 degrees of freedom (15 observations and 3 parameter estimates) may be obtained from the Table as 2.179. The rule for testing is, then, that if the computed statistic exceeds the value 2.179 we reject the Null hypothesis, but not otherwise.

The estimated regression equation (6.1) for the Indian engineering data is

$$\underset{}{I_t} = \underset{(98.112)}{-\,82.518} + \underset{(0.020)}{0.084}\, S_t + \underset{(0.418)}{0.048}\, P_t \qquad R^2 = 0.98 \qquad (6.4)$$

In this case, $\hat{\beta}_1$ is 0.084, and its standard error is 0.020; so the *t*-ratio is

$$t = 0.084 / 0.020 = 4.2. \qquad (6.5)$$

Since the computed *t*-ratio satisfies our test rule, we reject the Null hypothesis.

Hence, the Indian engineering data provide enough evidence for rejecting the Null hypothesis that "sales do not cause movements in investment." We are 95 percent confident that this test procedure will yield the correct answer when used in repeated samples.

This test procedure is based on one parameter and may be used for any specific value of the parameter and not necessarily zero as in the above example.

Consider the Null hypothesis

$$H_N: \beta_1 = 0.02 \qquad H_A: \ H_N \text{ is false.} \qquad (6.6)$$

In this case the test statistic is defined as

$$t = (\hat{\beta}_1 - 0.02) \text{ / standard error of } \hat{\beta}_1 \text{ .} \qquad (6.7)$$

When the Null hypothesis is true and error terms are normally distributed, the statistic defined by (6.7) follows the “Student’s *t*” distribution with (T - K) degrees of freedom. Once the distribution of the test statistic is known, the probability of committing a Type I error is known; hence, so are the critical value and the test rule.

In general, this test procedure may be used for any of the parameters of a regression equation, provided that the test involves only one parameter. Let the regression equation in the general case be

$$Y_t \ = \ \beta_0 \ + \ \beta_1 X_{1t} \ + \ \beta_2 X_{2t} \ + \ ... \ + \ \beta_k X_{kt} \ + \ \varepsilon_t \ , \qquad (6.8)$$

and let the Null hypothesis be based on the *i*-th parameter β_i as

$$H_N: \beta_i = \mu \qquad H_A: \ H_N \text{ is false,} \qquad (6.9)$$

where μ is the chosen constant [0 in (6.2); 0.02 in (6.6)].

The test statistic is defined as

$$t = (\hat{\beta}_i - \mu) \text{ / st. error of } \hat{\beta}_i \text{ .} \qquad (6.10)$$

When the Null hypothesis is true and the errors are generated by a normal distribution, the *t* statistic follows the Student’s *t* distribution with the appropriate degrees of freedom.

6.2 Consequences of Biased Estimates

In the preceding section the conventional *t*-statistic has been used as the test criterion. This statistic follows Student's *t* distribution only when the estimated regression equation is the truth and when errors are normally distributed. In many econometric investigations these assumptions need not hold, and the researcher will be interested in knowing the consequences of their violation. Though he can reasonably assume that the errors are normal, he may suspect that the estimated regression equation is not the truth. When such is the case, the resulting estimates of the parameter and the estimates of the variances may be biased, in which case the distribution of the statistic would be altered. In a typical econometric problem the researcher may have a misspecification of some sort which introduces bias in the estimate $\hat{\beta}_i$ and also in the estimate of the variance of $\hat{\beta}_i$.

To understand the concepts let us consider an extremely simple situation. Suppose that the Null hypothesis (6.9) is true. Let the estimate of the true regression equation be $\hat{\beta}_i$, which is assumed to be unbiased. The *t*-statistic is computed as

$$t = (\hat{\beta}_i - \mu) / \sqrt{\hat{V}(\hat{\beta}_i)} \ , \qquad (6.11)$$

where $\hat{V}$ is the estimate of the variance of $\hat{\beta}_i$ and μ is the chosen value of β_i under the Null hypothesis. The *t*-statistic has zero mean and follows Student's *t* distribution. The probability of Type I error corresponding to a specified critical value t_c, is given by the shaded area in Figure 6.1.

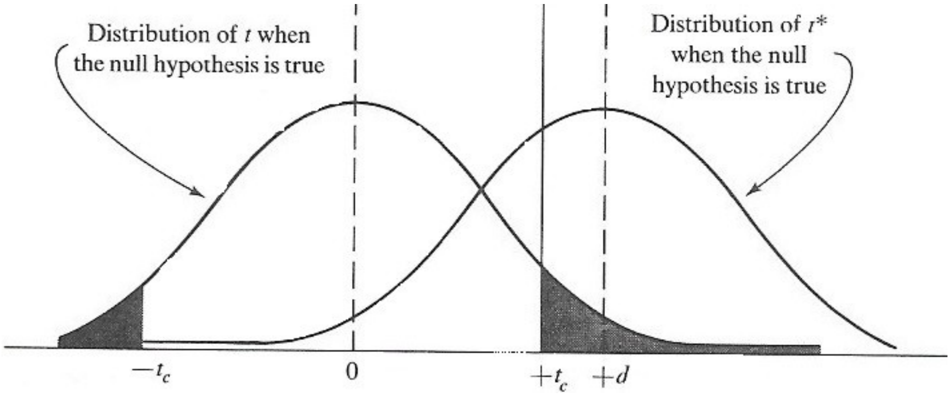

Figure 6.1. Distribution of *t* and *t** for the case of Positive Bias

Suppose there is another estimate—say, $\hat{\beta}_i^*$ —which is a biased estimate of β_i. When the Null hypothesis (6.9) is true,

$$E(\hat{\beta}_i^*) \; = \; \beta_i + d \; = \; \mu + d \,, \qquad (6.12)$$

where *d* is the amount of bias, which may be positive or negative. If the researcher uses the biased estimate $\hat{\beta}_i^*$ instead of $\hat{\beta}_i$ in computing the *t*-statistic in (6.11), the distribution of *t* changes. Let the *t*-statistic based on $\hat{\beta}_i^*$ be called t^*. Then

$$t^* \; = \; (\hat{\beta}_i^* - \mu) \; / \; \sqrt{\hat{V}(\hat{\beta}_i)} \; . \qquad (6.13)$$

The t^* - statistic has a mean value of *d* whereas *t* has a mean value of zero. We shall suppose for purposes of exposition that only the mean of the distribution is affected by the substitution of $\hat{\beta}_i^*$ for $\hat{\beta}_i$ in equation (6.11). The distribution of t^* for a positive value of *d* is given in Figure 6.1. The distribution of t^* for a negative value of *d* can be drawn similarly to the left of the *t* distribution, as shown in Figure 6.2.

When the researcher uses the same critical value t_c for the test rule, the probability of Type I error will not be the same for the use of $\hat{\beta}_i^*$ as for the use of $\hat{\beta}_i$ in computing the test statistic. The probability of Type I error is the probability of the test statistic exceeding the critical value; as may be seen from Figures 6.1 and 6.2, this differs for the distribution of *t* and of t^*.

When the researcher is not aware that he is using $\hat{\beta}_i^*$, he may be under the impression that the probability of Type I error associated with his test criterion is the shaded area in Figure 6.1, whereas the implied Type I error is that corresponding to the t^* distribution. This changes the probability of Type I error. In comparing the probability of actual Type I error (based on t^*) and the probability of theoretical Type I error (based on *t*), it can be seen that the difference

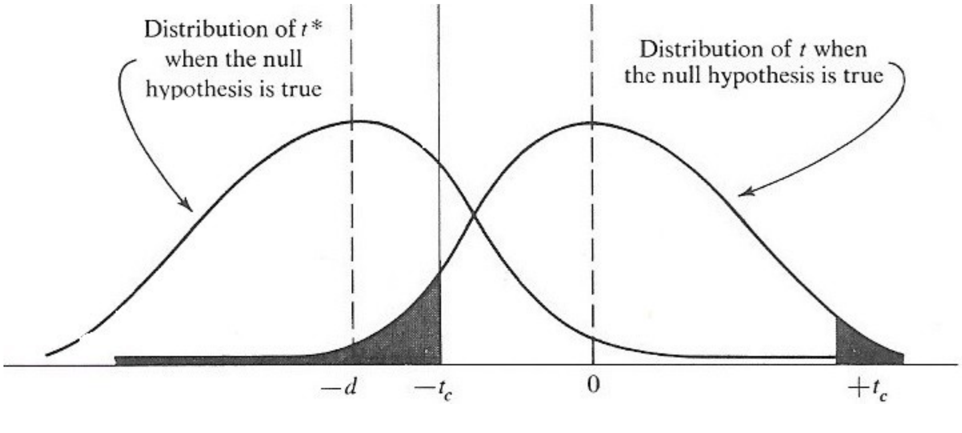

Figure 6.2. Distribution of *t* and *t** for the Case of Negative Bias

between them increases with the value of *d* for the right-tail test and decreases for the left-tail test. This result is symmetric for negative values of *d*.

The consequences of bias in only the regression estimate is to underspecify the probability of Type I error implied by the test criterion for the case of positive bias and a right-tail test or of negative bias and a left-tail test. When the bias is small, the theoretical and the actual Type I error may be so close that the difference is of no real consequence. When the researcher suspects that his estimate is biased, he may opt for a conservative choice of the critical value to allow for such discrepancy. The situation becomes crucial only when bias in the estimates is of such size that the Null hypothesis would be rejected under the theoretical *t* distribution but not under the t^* distribution. Should such a situation arise, the researcher is advised to seek more information or to try to reduce the bias of his estimate by inclusion of proxy variables or by other means.

6.3 Consequence of Biased Estimates of Coefficient Variance

Now let us turn to the situation in which $\hat{V}(\hat{\beta}_i)$, the estimate of variance of $\hat{\beta}_i$, is biased. Once again, for the sake of exposition, let us suppose that instead of $\hat{V}(\hat{\beta}_i)$ the researcher uses a biased estimate in equation (6.11) to compute the *t*-statistic. Let the statistic with $\hat{V}^{/}(\hat{\beta}_i)$ be

$$t^{/} \;=\; (\hat{\beta}_i - \beta_i) \;/\; \sqrt{\hat{V}^{/}(\hat{\beta}_i)} \;. \qquad (6.14)$$

The $t^{/}$ is centered around zero because the estimate $\hat{\beta}_i$ is assumed to be unbiased. When $\hat{V}^{/}(\hat{\beta}_i)$ is an overestimate of $\hat{V}(\hat{\beta}_i)$, the variance of *t*' will be smaller than that of *t*. Similarly, when $\hat{V}^{/}(\hat{\beta}_i)$ is smaller than $\hat{V}(\hat{\beta}_i)$, the variance of $t^{/}$ will be larger than that of *t*.

The distributions of *t* and $t^{/}$ are presented in Figure 6.3. The critical value is t_c. The probability of a Type I error corresponding to t_c for a right-tail test is given by the areas under the respective distributions to the right of t_c. When $\hat{V}^{/}$ is larger than $\hat{V}$; a frequent case in linear regressions, the probability of a Type T error associated with $t^{/}$ is smaller than that associated with *t*. When the researcher is not aware of the upward bias in the variance of his estimate, he may believe that he is testing with a lower level of confidence than is actually the case.

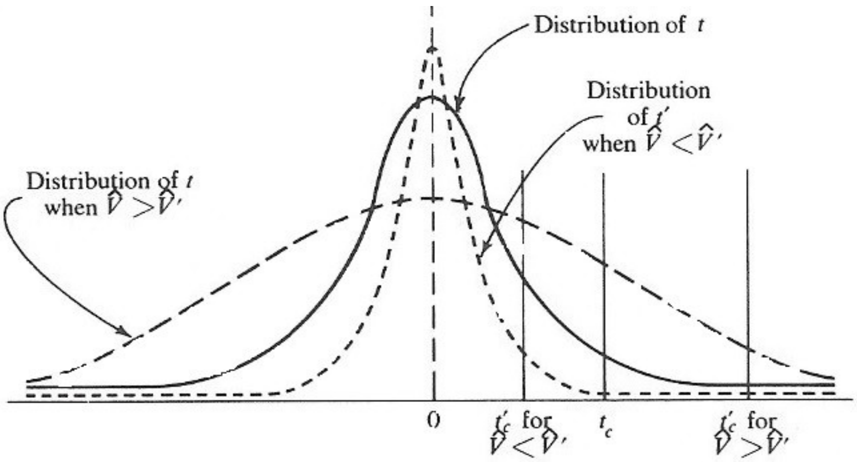

Figure 6.3. Distribution of *t'* for various estimates of $\hat{V}\left(\hat{\beta}_i\right)$

In a practical situation, however, he may have bias both in his regression estimates and in his estimate of variance. In a right-tail test, downward bias in the parameter estimate and upward bias in the estimate of variance tend to indicate a probability of Type I error larger than the actual probability. When the researcher rejects a Null hypothesis at the specified probability of Type I error, he will be rejecting it as well at other lower values, hence there is no problem. His test criterion is more “powerful” than he believes it to be.

When the bias in the estimate is positive and bias in the estimate of variance is also positive, the first tends to overstate and the other to understate the level of confidence. This also does not cause serious problems unless the test result is marginal: that is, for example, unless the researcher would reject the hypothesis at the 95 percent but not at the 99 percent level of confidence. In such a situation an applied econometrician cannot be sure whether the actual level of confidence is the same as the theoretical. If he rejects the Null hypothesis at the 95 percent level on the basis of the theoretical level of confidence, he is completely ignoring the consequences of bias in his estimates.

Once having an idea of the extent of bias, at least whether the bias is considerable or not, a researcher may conjecture the maximum and minimum levels of confidence associated with a critical value t_c. If he then rejects the Null hypothesis at the minimum actual probability of Type I error, he can be confident that, even though, the estimates are biased, his test is conclusive.

So far we have presented the consequences of bias in the parameter estimates and in the estimate of variance. Now let us consider some typical problems in econometrics in order to study the direction of such bias.

Let the true relation between the economic variables be

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t \qquad (6.15)$$

where the variables are deviations from their respective means. By misspecification the researcher estimates the following regression equation:

$$y_t = \hat{\beta}_1 x_{1t} + e_t . \qquad (6.16)$$

We have already shown that the estimate $\hat{\beta}_1$, is biased and that the expression for bias is

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 b_{21} . \qquad (6.17)$$

To study the nature of bias in the estimate of variance of $\hat{\beta}_1$, let us consider the bias in the estimate of variance of the error term. This variance is usually estimated from the sum of squares of the residuals.

The residual corresponding to the *t*-th observation is

$$e_t = y_t - \hat{\beta}_1 . x_{1t} . \qquad (6.18)$$

By substituting the true value of y_t given by equation (6.15) and the ordinary least squares estimate of β_1, in equation (6.16) we obtain

$$e_t = \beta_2 x_{2t} - \beta_2 b_{21} x_{1t} + \varepsilon_t - x_{1t} \Sigma x_{1t} \varepsilon_t / \Sigma x_{1t}^2 . \qquad (6.19)$$

When both sides of equation (6.19) are squared and the expected value of the summation over all the observations is taken,

$$E(\Sigma e_t^2) = \sigma^2 (T-1) + \beta_2^2 \Sigma x_2^2 (1 - r_{x1x2}^2) . \qquad (6.20)$$

The estimate of variance of the error term (*e*) based on the ordinary least squares residuals is biased whenever the equation is misspecified. This bias is nonnegative; that is, the variance of the error term is always overestimated (unless, of course, β_2 = 0 or r_{x1x2}^2 = 1). The extent of this bias depends on the coefficient of and the sample variance of the left-out variable. It also depends on the correlation between the left-out and the included variable (r_{x1x2}).

Notice that as this correlation increases from zero to unity this bias goes from its maximum to its minimum value.

When the left-out variable is perfectly correlated with the included variable, r_{x1x2}= 1, the included variable captures all the influence of the left-out variable in its coefficient, thus leaving no part of the left-out variable in the error terms. Since the residuals then represent only the error terms, and not any part of the left-out variable, the estimate of variance of errors computed from the residuals is unbiased. Note that these results are derived under the assumption that the *x*'s are held constant in repeated trials.

6.4 Test on a Linear Function of Parameters

Often the researcher is interested in testing a Null hypothesis on a linear function of the parameters rather than on the individual parameters. For example, he may want to test for returns to scale in the context of a Cobb-Douglas production function.

We shall study the general case of testing a linear combination of parameters in the linear regression

$$y_t \;=\; \beta_1 x_{1t} \;+\; \beta_2 x_{2t} \;+\; \ldots \;+\; \varepsilon_t \qquad (6.21)$$

Let the Null hypothesis be

$$H_N : \delta = c_1 \beta_1 + c_2 \beta_2 + \ldots \qquad H_A : H_N \text{ is false,} \qquad (6.22)$$

where δ, c_1, c_2 ... are specified (chosen) constants.

When the researcher is testing for constant returns to scale with a Cobb-Douglas production function, δ = 1 and all of the *c*'s = 1, so that the Null hypothesis (6.22) becomes

$$H_N :\; 1 \;=\; \beta_1 + \beta_2 + \ldots + \beta_k \qquad (6.23)$$

where there are *k* factors of production and the variables are in their logarithmic form.

The test statistic may be designed as

$$d \;=\; c_1 \hat{\beta}_1 \;+\; c_2 \hat{\beta}_2 \;+\; \ldots . \qquad (6.24)$$

When the $\hat{\beta}$'s are unbiased estimates,

$$E(d) \;=\; c_1 \beta_1 + c_2 \beta_2 + \;\ldots\; = \;\delta \qquad (6.25)$$

The variance of the statistic *d* is

$$V(d) \; = \; c_1^2 V(\hat{\beta}_1) + c_2^2 V(\hat{\beta}_2) + ... + 2 c_1 c_2 COV(\hat{\beta}_1, \hat{\beta}_2) + ... , \qquad (6.26)$$

where COV stands for covariance between the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. In linear regression models the covariance term is usually nonzero and plays a prominent role in the testing of the hypothesis.

When the error terms are normally distributed the *d*-statistic follows a normal distribution with mean and variance given by (6.25) and (6.26). The *d*-statistic (6.24) may be standardized by the following transformation:

$$z \; = \; (d - \delta) / \sqrt{V(d)} \, . \qquad (6.27)$$

The *z*-statistic follows the standard normal distribution with mean zero and variance 1.

The researcher does not know V(d) because it involves V($\hat{\beta}_i$), which is generally unknown. When an estimate of V($\hat{\beta}_i$) on the basis of the ordinary least squares estimation procedure is used instead of the true V($\hat{\beta}_i$) in computing V(*d*), the resulting statistic follows Student's *t* distribution with (T - K) degrees of freedom, where T is the number of observations and K is the number of parameters estimated by equation (6.21), including the constant term. When V(*d*) is replaced by $\hat{V}$(d), the *z*-statistic may be written as a *t*-statistic

$$t \; = \; (d - \delta) / \sqrt{\hat{V}(d)} \, . \qquad (6.28)$$

When the researcher has theoretical information which says that the Null hypothesis is true, then there is of course no point in testing. The researcher may incorporate the truth into his estimation procedure by estimating the production function such that 1 = $\hat{\beta}_1 + \hat{\beta}_2$. This can be accomplished by writing the regression equation under the truth as

$$\log q \; = \; \hat{\beta}_0 + \hat{\beta}_1 \log l + (1 - \hat{\beta}_1) \log k + e \, .$$

By rearrangement of terms

$$(\log q - \log k) \; = \; \hat{\beta}_0 + \hat{\beta}_1 (\log l - \log k) + e$$

Estimating this equation with y'(= log q - log k) as the dependent variable and x'(= log l - log k) as the independent variable yields $\hat{\beta}_1$ under the restriction of constant returns to scale. The estimate $\hat{\beta}_2$ is obtained as 1- $\hat{\beta}_1$.

Now let us turn to the production function in Indian woolen textiles to test whether there are constant returns to scale. The regression equation is

$$\log q \;=\; \beta_0 + \beta_1 \log l + \beta_2 \log k + \varepsilon \;. \qquad (6.29)$$

The Null hypothesis may be stated as

$$H_N:\; 1 = \beta_1 \;+ \beta_2 \qquad H_A:\; H_N \text{ is false.} \qquad (6.30)$$

The regression equation estimated from the data is

$$\underset{(0.154)}{\mathit{log}\ q = 1.652} \;+\; \underset{(0.138)}{0.708\ \mathit{log}\ l} \;+\; \underset{(0.161)}{0.413\ \mathit{log}\ k} \qquad R^2 = 0.947\ . \qquad (6.31)$$

The value of d (= $\hat{\beta}_1$ + $\hat{\beta}_2$) is 1.121, and the estimate of variance of *d* is

$$\begin{aligned}\hat{V}(d) \;&=\; \hat{V}(\hat{\beta}_1) + \hat{V}(\hat{\beta}_2) + 2\,C\hat{O}V(\hat{\beta}_1, \hat{\beta}_2)\\ &= 0.0189 + 0.0260 + 2\,(-0.0013) = 0.0423\ . \qquad (6.32)\end{aligned}$$

The *t*-statistic computed from *d* is

$$t \;=\; (d - 1)\ /\ \sqrt{\hat{V}(d)} = 0.121\ /\ 0.2058 = 0.588\ . \qquad (6.33)$$

The *t*-statistic in equation (6.33) follows the Student's *t* distribution with 12 degrees of freedom (15 observations minus 3 parameter estimates).

The critical value corresponding to the 95 percent confidence level is 2.201. the test rule is that we reject the Null hypothesis whenever the computed *t*-statistic exceeds the critical value 2.201 in magnitude. The computed *t*-statistic does not satisfy the rule. Therefore we do not reject the Null hypothesis that there are constant returns to scale in the woolen textile industry in India.

6.5 Simultaneous Test on Several Parameters

In some situations the theory may make a statement regarding several parameters at the same time. The theory is true only when the entire statement is true and not when just a part of it is true. In such a case the researcher has to treat the statement as a whole and should not try to test parts of it separately. For example, a theory may state that the parameters β_1 and β_2 in a regression equation take the specific values $\beta_1 = \mu_1$ and $\beta_2 = \mu_2$ If the researcher should test null hypotheses separately on β_1 and β_2, he would be misinterpreting the implications of the theory, that both of them must be simultaneously true.

Consider, for example, the case of tea imports into the United States. Let a theory state that the imports of Ceylonese tea do not depend on the price of Indian tea and that the price elasticity of Ceylonese imports is unity. This theory makes a statement regarding the price sensitivity of Ceylonese tea imports with respect to both Indian price and Ceylonese price. Hence, we cannot treat the statement as two different parts and test them separately.

Let the demand for Ceylonese imports be:

$$\log TEA \ = \ \beta_0+\beta_1 \log P_{cy}+\beta_2 \log P_I+\beta_3 \log P_{bz}+\beta_4 \log Y+\varepsilon \ . \qquad (6.34)$$

The Null hypothesis may be posed in the general framework as

$$H_N : \beta_1=\mu_1 \text{ and } \beta_2=\mu_2 \qquad H_A : H_N \text{ is false.} \qquad (6.35)$$

In the context of our example, μ_1 = -1 and μ_2 = 0.

The required test statistic may be based on the two separate values for the sum of squares of the residuals under the null and under the alternative hypotheses. The Alternative hypothesis implies no conditions on the parameters; hence, the residual sum of squares under the Alternative hypothesis, RSS(H_A), using ordinary least squares estimation, is the minimum value of sum of squares.

When the regression equation is estimated under the presumption that the Null hypothesis is true, the residual sum of squares, RSS(H_N), will be larger than the RSS(H_A). Since the Null hypothesis implies some specific values for the parameters, there is no need for estimating these parameters. In estimating the regression under the Null hypothesis the researcher is forcing the estimation procedure, so that the estimates are, in fact, the true values of the parameters under the Null hypothesis. This necessarily increases the residual sum of squares, because the least residual sum of squares is obtained under no restrictions.

The increase in the residual sum of squares due to imposing the condition that the Null hypothesis is true provides a basis for the test. When the Null hypothesis is actually true, then imposing the condition that the Null hypothesis is true should not increase the residual sum of squares. Because of sampling fluctuations we cannot hope to obtain a zero increase; hence we should test to see whether the increase in the residual sum of squares is significantly different from zero.

The increase in the sum of squares depends on the number of observations, on the number of restrictions implied by the Null hypothesis, and also on the units of measurement of the dependent variable. Since we want the test statistic to be independent of these factors we shall standardize the increase in the residual sum of squares as

$$F = \frac{[RSS(H_N) - RSS(H_A)]/n}{RSS(H_A)/(T-K)} \ . \qquad (6.36)$$

The numerator shows the increase in the residual sum of squares due to the restrictions imposed on the parameters, adjusted for the number of restrictions *n*. As the number of restrictions increases, the difference between RSS(H_N) and RSS(H_A) also increases. By dividing by the number of restrictions we are, in a way, correcting for this.

The denominator is the residual sum of squares based on (T - K) degrees of freedom; hence, it is divided by the number of degrees of freedom. Since the numerator and the denominator are in the same units, namely the square of the dependent variable, even if the unit of measurement of the dependent variable changes we still obtain the same value for the statistic.

The *F*-statistic depends crucially on two parameters: the number of restrictions (*n*), called the number of degrees of freedom in the numerator, and the degrees of freedom of the regression equation when no restrictions are imposed on the estimation, called the number of degrees of freedom in the denominator. As the number of degrees of freedom change in the numerator and the denominator, the distribution of the *F*-statistic changes when the Null hypothesis is true. Under the assumption that the errors are normally distributed, the theoretical properties of the *F*-statistic follow Snedecor's *F*-distribution with the corresponding number of degrees of freedom of the numerator and the denominator.

The researcher may specify an arbitrary critical value for the *F*-statistic as *F*. When the statistic exceeds the critical value the researcher rejects the Null hypothesis. The probability of Type I error associated with the critical value *F*, can be obtained by the area under the *F*-distribution corresponding to the value of *F* that implies rejection of the Null hypothesis when in fact it is true. The Type I errors corresponding to specified critical values are readily available in tabular form in any standard textbook in statistics.

Now we turn to the problem of obtaining the sum of squares of residuals under the null and the alternative hypotheses. When the Alternative hypothesis is true, there are no restrictions on the estimates, and the residual sum of squares is exactly that obtained under the ordinary least squares estimation of the specified regression equation (6.34). To obtain the residual sum of squares under the Null hypothesis the researcher has to impose the condition that the resulting estimates are the same as the parameter values implied by the Null hypothesis. In equation (6.34) with four independent variables the Null hypothesis $\beta_1 = \mu_1$ and $\beta_2 = \mu_2$ may be imposed as

$$\log TEA \;=\; \hat{\beta}_0 + \mu_1 \log P_{cy} + \mu_2 \log P_I + \hat{\beta}_3 \log P_{bz} + \hat{\beta}_4 \log Y + e\ . \qquad (6.37)$$

Since μ_1 and μ_2 are specified constants, the corresponding terms may be taken to the left side to obtain

$$\log TEA - \mu_1 \log P_{cy} - \mu_2 \log P_I \;=\; \hat{\beta}_0 + \hat{\beta}_3 \log P_{bz} + \hat{\beta}_4 \log Y + e\ , \qquad (6.38)$$

which may be rewritten as:

$$Y' = \hat{\beta}_0 + \hat{\beta}_3 \log P_{bz} + \hat{\beta}_4 \log Y + e \qquad (6.39)$$

where $Y' = \log TEA - \mu_1 \log P_{cy} - \mu_2 \log P_I$.

Obtaining equation (6.39) by ordinary least squares is identical to imposing the restrictions $\hat{\beta}_1 = \mu_1$, and $\hat{\beta}_2 = \mu_2$ in equation (6.37). By defining a new variable, Y', and running regression equation (6.39) the researcher has obtained the residual sum of squares under the Null hypothesis. Once the residual sums of squares under the null and the alternative hypotheses are known, the *F*-statistic can be computed by using expression (6.36).

In the specific example of Ceylonese tea, the regression equations under the alternative and null hypotheses are

$$\underset{}{\log \text{TEA}} = \underset{(2.000)}{2.837} \;-\; \underset{(0.987)}{1.481} \log P_{cy} + \underset{(0.690)}{1.181} \log P_I + \underset{(0.134)}{0.186} \log P_{bz} + \underset{(0.370)}{0.257} \log \text{Y}$$

$$\text{RSS}(H_A) = 0.4277 \qquad (6.40)$$

$$(\log \text{TEA} + \log P_{cy} - 0\,.\,P_I) = -\underset{(0.820)}{0.738} + \underset{(0.155)}{0.199} \log P_{bz} + \underset{(0.165)}{0.261} \log \text{Y}$$

$$\text{RSS}(H_N) = 0.6788 \qquad (6.41)$$

The Null hypothesis imposes two restrictions on the estimates. The *F*-statistic may be computed as

$$F = \frac{(0.6788 - 0.4277)/2}{(0.4277)/17} = 4.99 \qquad (6.42)$$

Since the numbers of degrees of freedom in the numerator total 2 and in the denominator 17, the *F*-statistic follows *F*(2, 17). Corresponding to the 95 percent level of confidence we obtain the critical value F, as 3.59. The test rule is, "Whenever the computed *F*-statistic exceeds the critical value (F_c = 3.59) we reject the Null hypothesis." The computed statistic (*F* = 4.99) exceeds the critical value; therefore at the 95 percent level of confidence we reject the Null hypothesis.

Notice that if the researcher had treated the stated Null hypothesis as two independent statements regarding the parameters β_1, and β_2 he would have reached a different answer. Since the theory makes a statement on the parameters jointly rather than independently, a test procedure should be based on the total statement. Testing only a part of a complete statement may lead to wrong conclusions.

A Null hypothesis frequently found in empirical research is the case in which μ_1 = 0 and μ_2 = 0. The residual sum of squares under the Null hypothesis is obtained by merely deleting the corresponding independent variables from the regression equation.

This test procedure can be extended to regression equations with several independent variables and several restrictions on the parameters. Formula (6.36) is given for a general case with *n* restrictions implied by the Null hypothesis and *K* parameters, including the constant term implicit in our discussion. It may be noted that when *n* = 1, the test procedure is identical to that which we used for testing a single parameter. When the number of degrees of freedom in the numerator is 1, the F-statistic is nothing but the square of the *t*-statistic with the number of degrees of freedom of the denominator.

6.6 Test for Equivalence of Two Parameters

Another Null hypothesis involving several parameters that is frequently encountered by empirical researchers is

$$H_N : \beta_1 = \beta_2 \qquad H_A : H_N \text{ is false .} \qquad (6.43)$$

This Null hypothesis (6.43) is similar to the one discussed earlier. The *F*-statistic can be used once the residual sum of squares has been obtained under the Null and the Alternative hypotheses.

The residual sum of squares under the Alternative hypothesis is identical to that obtained under the ordinary least squares estimation. The sum of squares of the residuals under the Null hypothesis can be obtained by imposing the condition that the resulting estimates for β_1 and β_2 are the same. This may be accomplished in the writing of the regression equation. For example, in the case of three independent variables;

$$y_t = \hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} + \hat{\beta}_3 x_{3t} + e_{3t} \qquad (6.44)$$

Since the estimates of β_1 and β_2 are the same, under Null hypothesis (6.43) equation (6.44) may be rewritten as

$$y_t = \hat{\beta}_1 (x_{1t} + x_{2t}) + \hat{\beta}_3 x_{3t} + e_{3t} \qquad (6.45)$$

which becomes

$$y_t = \hat{\beta}_1 x'_t + \hat{\beta}_3 x_{3t} + e_{3t} \qquad (6.46)$$

By defining the variable x' as $x_1 + x_2$ and using ordinary least squares estimation for regression equation (6.46), we are able to estimate equation (6.44) with the restrictions implied by the Null hypothesis.

This Null hypothesis may be generalized to a test involving several parameters.

Consider for example

$H_N : \beta_1 = \beta_2 = \beta_3$ $\qquad$ $H_A : H_N$ is false.

The Null hypothesis involves two restrictions on parameters, namely $\beta_1 = \beta_2$ and $\beta_2 = \beta_3$. The residual sum of squares under the Null hypothesis is obtained by defining x' as $x_1 + x_2 + x_3$.

The residual sum of squares given by equation (6.46) is the RSS(HN). The *F*-statistic given by (6.36) can be computed for testing the Null hypothesis.

To illustrate the test procedure, let us consider the rice production function for the Guntur district in India:

$$Q = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 R + \beta_4 t + \varepsilon \qquad (6.47)$$

Assume that the Null hypothesis states that the marginal yield of a dry acre is the same as that of an irrigated acre. The Null hypothesis may be expressed as:

$$H_N : \beta_1 = \beta_2 \qquad H_A : H_N \text{ is false.} \qquad (6.48)$$

To compute the test statistic, *F*, we need to estimate equation (6.47) under the Null hypothesis and the Alternative hypothesis. Under the Alternative hypothesis the parameter values can take any value; hence, we estimate the equation without imposing any restrictions:

$$Q = -739.950 + 0.578\ I + 0.218\ D + 46.588\ R - 40.388\ t$$
$$\text{RSS}(H_A) = 1{,}614{,}627. \qquad (6.49)$$

When the researcher has external theoretical information that states that the Null hypothesis is true, he may incorporate the truth into his estimation by forcing the coefficients of the appropriate variables to be the same by using this procedure.

When the Null hypothesis is true we estimate the equation in such a way that we obtain the same regression coefficient for the two independent variables (*I* and *D*) as

$$Q = -\ 669.241 + 0.520\ (I + D) + 47.603\ R - 32.246\ t$$
$$\text{RSS}(H_N) = 1{,}626{,}856\ . \qquad (6.50)$$

The residual sums of squares under the Null and the Alternative hypotheses are 1,626,856 and 1,614,627 respectively. The corresponding *F*-statistic is

$$F = \frac{(1{,}626{,}856 - 1{,}614{,}627)/1}{(1{,}614{,}627)/16} = 0.1211\ . \qquad (6.51)$$

Corresponding to one degree of freedom in the numerator and sixteen degrees of freedom in the denominator, the critical value for the *F* distribution for the 95 percent level of confidence is obtained as 4.49. The test rule is, "Whenever the computed *F*-statistic exceeds the value of 4.49 we reject the Null hypothesis."

The computed statistic 0.121 does not satisfy the rule. Hence, the Null hypothesis is not rejected, at the 95 percent level of confidence. We do not reject the Null hypothesis that the marginal yields of irrigated acre and of dry acre are the same.

This test procedure may also be used when the Null hypothesis implies that one parameter is a constant proportion of another. For example, let the Null hypothesis be

$$H_N:\ k.\beta_1=\beta_2 \qquad H_A: H_N \text{ is false.} \qquad (6.52)$$

where k is a given constant.

Under Null hypothesis (6.52) the estimated regression equation (6.45) now becomes

$$y_t\ =\ \hat{\beta}_1(x_{1t}+k\,x_{2t})\ +\ \hat{\beta}_3 x_{3t}+e_{3t}\ . \qquad (6.53)$$

Testing of the Null hypothesis in many practical situations involves either the *t*- or the *F*-statistic, and the necessary ingredients for computing these statistics may be obtained by simple least squares regression. Once the researcher translates the restrictions implied by the Null hypothesis into an estimable regression equation, the rest is a mechanical standard process of estimation.

6.7 Testing across Several Sets of Data

All the test procedures discussed so far relate to a linear regression for a given set of data. When the researcher estimates a regression equation separately for several sets of data, he may want to test whether some or all of the parameters are the same for all different sets of data. Instances are numerous in econometric research. The researcher may, for example, want to test whether the demand schedule for Indian imports in the US. is the same before and after India's independence. He will then estimate one regression for pre-independence and another for post-independence, and test the Null hypothesis that they are the same for both periods. As another example, he may be interested in testing the Null hypothesis that the marginal productivity of labor is the same in all southern as in all northern states of the United States.

This problem is reduced essentially to the problem of units of measurement. By appropriate choice of units of measurement we can always express constant proportionality between parameters in the form of Null hypothesis (6.48).

Consider the following regression equations for two sets of data in which the same definitions of the variables and the same general regression model are used:

$$Y_t\ =\ \beta_0\ +\ \beta_1 X_{1t}\ +\ \beta_2 X_{2t}\ + \beta_3 X_{3t}\ +\ \varepsilon_{1t} \qquad (6.54)$$

$$Y_{t'}\ =\ \beta_0^*\ +\ \beta_1^* X_{1t'}\ +\ \beta_2^* X_{2t'}\ +\ \beta_3^* X_{3t'}\ +\ \varepsilon_{2t'}\ . \qquad (6.55)$$

Equation (6.54) corresponds to the first set of data and (6.55) corresponds to the second set. In this case the *t* and *t*' subscripts (respectively distinguishing the two sets of data) are used only to identify separate observations and do not necessarily imply time-series data. The following results hold for cross-sectional as well as time-series data.

The researcher is interested in testing the Null hypothesis

$$H_N:\ \beta_1=\beta_1^* \qquad H_A: H_N \text{ is false .} \qquad (6.56)$$

The test procedure may be based on the residual sum of squares under the Null and the Alternative hypotheses. To pose this as a special case of preceding problems, let us introduce a dummy variable, *D*, to distinguish between the two sets of data. Let *D* take on a value of zero when corresponding to the first set of data and of unity for the second set. Since the variables used in equations (6.54) and (6.55) have the same definitions, we may combine the two sets of data into a pooled set having the original variables and then utilize a dummy variable to distinguish the sets.

Taking advantage of the information contained in the dummy variable, we may rewrite equations (6.54) and (6.55) as one regression equation:

$$Y_t\ =\ \beta_0+\alpha_0 D+\beta_1 X_{1t}+\alpha_1(X_{1t}.D)+\beta_2 X_{2t}+\alpha_2(X_{2t}.D)+\beta_3 X_{3t}+\alpha_3(X_{3t}.D)+\varepsilon_t \qquad (6.57)$$

When the dummy variable takes the value zero, the data correspond to the first set, and equation (6.57) is identical to equation (6.54). When the dummy variable is unity the data correspond to the second set, and equation (6.57) becomes

$$Y_t\ =\ (\beta_0+\alpha_0)+(\beta_1+\alpha_1)X_{1t}+(\beta_2+\alpha_2)X_{2t}+(\beta_3+\alpha_3)X_{3t}+\varepsilon_t \qquad (6.58)$$

which can be rewritten as

$$Y_t\ =\ \beta_0^*\ +\ \beta_1^* X_{1t}\ +\ \beta_2^* X_{2t}\ +\ \beta_3^* X_{3t}\ +\ \varepsilon_t\ . \qquad (6.59)$$

Equation (6.58) is nothing but equation (6.55) rewritten in terms of the parameters, $\beta 's$ and $\alpha 's$. By using equation (6.57) instead of the two equations (6.54) and (6.55) we neither gain nor lose any information in terms of interpretation of the parameters. As we have already shown, estimation of equation (6.57) from the pooled data gives the same regression estimates as those obtained by estimating equations (6.54) and (6.55) separately to the respective sets of data. The sum of squares of the residuals in equation (6.57) is equal to the addition of the two sums of squares of residuals from the estimation of equations (6.54) and (6.55) separately. Also, the number of degrees of freedom corresponding to equation (6.57) is equal to the sum of the separate numbers of degrees of freedom corresponding to the two regression equations (6.54) and (6.55).

To repeat: whether the researcher estimates the two equations (6.54) and (6.55) separately to the two sets of data, or estimates equation (6.57) to the pooled data, he will obtain identical information regarding the estimates, the residual sum of squares, and the number of degrees of freedom. There is no loss or gain in estimation or interpretation.

In the form of equation (6.57), Null hypothesis (6.56) is analytically equivalent to Null hypothesis (6.43) discussed above. The test procedure developed for Null hypothesis (6.43) is equally applicable for (6.56) when the latter is considered to be the regression equation for a single set of data.

To compute the test statistic we need the residual sum of squares under both the null and the alternative hypotheses. Under the Alternative hypothesis no restrictions are imposed on the estimates. The ordinary least squares estimation of equation (6.57) for the pooled data gives the residual sum of squares under the Alternative hypothesis. If the researcher has already computed the regression equations for the two sets of data, he may obtain the RSS(H_A) by simply summing the two residual sums of squares and the respective numbers of degrees of freedom to obtain the residual sum of squares and the number of degrees of freedom corresponding to equation (6.57).

The residual sum of squares under the Null hypothesis may be obtained by estimating equation (6.57) with the restrictions implied by the Null hypothesis.

The Null hypothesis implies that $\beta_1 = \beta_1^*$, which is the same as $\alpha_1 = 0$ in equation (6.57). This condition may be easily imposed by deleting the term (X_{1t} . D) in equation (6.57). The residual sum of squares obtained in estimating the following equation for the pooled data is the RSS(H_N):

$$Y_t \ = \ \beta_0 + \alpha_0 D + \beta_1 X_{1t} + \beta_2 X_{2t} + \alpha_2 (X_{2t} . D) + \beta_3 X_{3t} + \alpha_3 (X_{3t} . D) + \varepsilon_t \qquad (6.60)$$

Equation (6.60) implies the same estimate corresponding to the coefficient of X_{1t}, for the two regressions (6.54) and (6.55), whereas all estimates of other parameters are different for the two equations. It may be noticed that the number of restrictions imposed by the Null hypothesis in estimation of equation (6.57) also equals the number of terms deleted in equation (6.57) to obtain equation (6.60). When a Null hypothesis implies several, restrictions, the corresponding residual sum of squares may be obtained by deleting the appropriate terms in equation (6.57).

Once the researcher obtains the RSS(H_N) and RSS(H_A), the corresponding F-statistic given by (6.36) may be computed. The appropriate critical value, F_c, corresponding to a given level of confidence may be used for the test criterion.

The *F*-test may be used for any number of restrictions, provided that the researcher can obtain the residual sum of squares under the Null hypothesis. A frequent case in applied econometrics is the case in which, instead of testing for a subgroup of parameters, the researcher is interested in testing the Null hypothesis that all parameters are the same for the two sets of data. For example, in the case under consideration examine the Null hypothesis on equations (6.54) and (6.55):

$$H_N: \ \beta_i = \beta_i^* \quad \text{for all i(i= 0, 1, 2, 3)} \qquad H_A : H_N \text{ is false} \qquad (6.61)$$

This Null hypothesis implies four restrictions; we may use an *F*-test. The residual sum of squares under the Alternative hypothesis is the same as in the above examples. The residual sum of squares under the Null hypothesis is obtained by estimating (6.57) after deleting all the appropriate terms applied by the Null hypothesis. As can be seen, the equation implied by the null hypothesis is

$$Y_t \ = \ \beta_0 \ + \ \beta_1 X_{1t} \ + \ \beta_2 X_{2t} \ + \ \beta_3 X_{3t} \ + \ \varepsilon_t \, , \qquad (6.62)$$

which is in the same form as equations (6.54) and (6.55).

The RSS(H_N) is obtained by simply estimating equation (6.62) for the pooled data with no dummy variables. The RSS (H_A) is obtained by estimating equation (6.62) separately for the two sets of data. When the Null hypothesis implies that all parameters are the same for the two sets of data, there is no need to introduce a dummy variable, because the researcher can now obtain all ingredients necessary to compute the *F*-test by estimating the equations separately to the two sets of data and estimating the same equation to the pooled data.

The *F*-test may be used in the context of linear regression equations with several independent variables and for several sets of data. With (M) sets of data the researcher has to introduce (M - 1) dummy variables to distinguish the sets. By rewriting the regression equations corresponding to various sets of data in the form of equation (6.57) he can clearly see the restrictions implied by the Null hypothesis and may estimate the RSS(H_A) by the standard regression techniques.

The reader may already have noticed that in hypothesis testing, whether the Null hypothesis is based on one set of data or on several sets, on one parameter or on several parameters, the test procedure is to obtain the residual sum of squares by using the least squares estimation of a regression equation. Once the researcher obtains the basic form of the regression (for example, equation (6.57)) all regression equations that need to be computed are obtained by deleting the appropriate independent variables. With the availability of regression programs, the cost of obtaining an additional regression equation is negligible.

The test procedures discussed so far have been based on regression coefficients. In econometric research, however, one also encounters hypotheses not readily expressible in the form of linear regression models. Some such hypotheses and the relevant test procedures are now considered.

6.8 Test for Association

Often in empirical investigations the researcher proceeds on the basis of conjectures, regardless of theoretical reasoning. Such investigations are used to find out whether any systematic relation exists between two observed variables even if there is no reason for any causal relationship. These results are “empirical observations” and need not imply any theoretical relations. For example, a statistician may observe, on the basis of empirical investigation, that the proportion of lung cancer among smokers is higher than among nonsmokers. This empirical observation does not imply that smoking is the cause of lung cancer; any such causal relation would need to be established by the medical profession on the basis of its theories. Lack of any underlying theory does not, however, invalidate empirical observations.

When based on strong evidence, such observations call for explanation, and the search for explanation has resulted in much of the feedback from empirical work to theory. In econometric research, the making of empirical observations on the basis of evidence, even if no theory already exists, is a task equal in importance to the testing of a specified theory for appropriateness.

Any systematic relation between two qualitative variables is called association. When there is no systematic relation between the variables, they are called independent; that is, the value of one variable is not associated with the observed value of the other. Association does not imply causal relationship. Any variable qualitative in nature may be categorized into one of two groups: with, or without, a specified attribute. For example, information on a set of farmers may be classified as “with” or “without” education. When the variable is agricultural productivity, the yield may be classified ass “high” and “low.” Even though the information on the two variables is, in principle, quantitative, they may be treated as qualitative variables. When the above two variables, education and agricultural productivity, are precisely measured, the researcher may use a better test procedure. When the variables are inadequately defined, or when it is less “sinful” to treat the information as qualitative rather than quantitative, or when the information is purely qualitative, then the researcher may decide to use the test for association between the two variables.

In our example, if the per-acre yield is independent of the education of the farmers, and if farmers are classified as producing high or low yield, we would expect to observe the same proportion of high-yield farmers among the educated as among the illiterate; and a similar situation for the low-yield farmers.

When the researcher actually observes the same percentage of educated and of illiterate farmers having high yield, then empirically the two variables are independent. Because of sampling fluctuations, however, the researcher may rarely find the percentages to be exactly equal. He is then interested in testing an Null hypothesis that the difference is due to the sampling fluctuations. If the evidence rejects the Null hypothesis, then the difference in percentages is greater than can be attributed to sampling fluctuations, and some systematic relationship is implied.

In our example, the Null hypothesis may be stated as

$$H_N: \text{education and yield are independent} \qquad H_A: H_N \text{ is false} \qquad (6.63)$$

To test the Null hypothesis we need a test statistic. This may be obtained on the basis of what the researcher would expect to observe if the Null hypothesis were true. Under the Null hypothesis the proportion of high-yield farmers is the same for the non-educated as for the educated category, and this is the same proportion of high yield as for the total sample. Since the proportion of high-yield farmers in the total sample is known, the researcher can readily compute the number of farmers with high yield to be expected in each category if the Null hypothesis were true.

The data corresponding to 256 districts in India are given in Table 6.1.

Each number in the table represents the attributes corresponding to its column and row. For example, 40 families are illiterate and have high yield.

Table 6.1. Data on Education and Yield per Acre in India

Level of education	Yield per acre			
	Low	High	Total	Proportion
Illiterate	16	40	56	0.219
Educated	49	151	200	0.781
Total	65	191	256	1.0
Proportion	0.254	0.746	1.0	

In all, 191 families out of 256 have high yield. The proportion of high yield in the total sample is p = 191/256 = 0.746. If the two variables are independent, then we expect 0.746 of the total educated farmers to have high yield and the same proportion of illiterate farmers to have high yield. The number of expected observations corresponding to each category may be presented as in Table 6.2.

Table 6.2. Expected Numbers in Each Category
When Education and Yield Are Independent

Level of education	Yield per acre		
	Low	High	Total
Illiterate	14	42	56
Educated	51	149	200
Total	65	191	256

The observed numbers differ from those expected under the Null hypothesis. The researcher wants to test whether the difference is due to sampling fluctuations or not.

In all, there are four separate categories, and their corresponding expected and observed values may be arranged as in Table 6.3.

Table 6.3. The Difference between the Observed and Expected
Numbers in the Four Categories

Number i	Expected E_i	Observed O_i	Difference d_i
1	14	16	2
2	42	40	2
3	51	49	2
4	149	151	2

The test statistic is defined as

$$\text{Chi-square } (\chi^2) \ = \ \sum_{i=1}^{4} \left(d_i^2 / E_i \right) . \qquad (6.64)$$

The test statistic defined by (6.64) follows the chi-squared (χ^2) distribution with 1 degree of freedom. By setting up a critical value for the chi-square (χ_c^2) the researcher may use as a test rule: "Whenever the chi-squared value exceeds the critical value Z, the Null hypothesis is rejected." The levels of confidence associated with given levels of critical value are presented in most textbooks. For the 95 percent level of confidence the critical value of chi-square with one degree of freedom is 3.84. Hence, our test rule is, "Whenever the computed χ^2 exceeds 3.84 the Null hypothesis is rejected." By definition in (6.64) the computed χ^2 statistic is

$$\chi^2 = 4/14 \ + 4/42 \ + \ 4/51 \ + \ 4/149 \ = \ 0.44 \qquad (6.65)$$

The test statistic does not satisfy the test rule. We do not reject the Null hypothesis on the basis of available evidence. That is, within the population covered by the data there is not enough evidence to reject the statement, "Education of farmers and their yield per acre have no systematic empirical relation."

This chi-squared test may be applied to numerous instances in economic and business studies. When the researcher finds a systematic relation between the variables, he may be interested in knowing whether the association is positive or negative. When low literacy and low yield are more frequently observed than would be expected if the two variables were independent, then it may be called *positive association*; a reverse case is *negative association*.

Sometimes the systematic relation between two variables may be the result of a third variable influencing them both. In such cases the researcher can make a valid inference from the data only if the third variable is held constant. This may be accomplished by separating the data into several groups within which the third variable is kept constant; the association within each subgroup is then called the *partial association*. If the systematic relation between the variables is not due to the presence of the third variable, the same systematic relation should be observed within all the subgroups classified according to the third variable. If the relationship changes in a systematic way according to which values the third variable takes, then it becomes apparent that the observed systematic relation cannot be independent of the third variable. In this case the observed association is called a *spurious association*. The researcher should be alert for such misleading results.

6.9 Test for Correlation

Association between two quantitative variables is called a *correlation*. We used a 2 x 2 table in the context of qualitative variables, but when the data on variables are more precise than mere "high" and "low" we should be able to improve the test procedure by taking advantage of this information. The correlation between two quantitative variables provides such a test.

Let X and Y be two variables under investigation, and let the data correspond to T observations. The correlation between the variables is defined as

$$r_{XY} = \frac{\Sigma(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\Sigma(X_t - \bar{X})^2 . \Sigma(Y_t - \bar{Y})^2}} , \qquad (6.66)$$

where $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.

When the value of X does not depend upon the value of Y, then the correlation between the two is zero, and they are called independent. Suppose the values of X and Y are taken from a population with zero correlation. Since the correlation coefficient *r* is based on a sample of only *T* observations from this population it need not be zero, because of sampling fluctuations.

The researcher is interested in testing whether the correlation between the two variables in the population is zero. If the observed correlation in the sample is more than would be expected due to sampling fluctuations, then he may reject the Null hypothesis that the two variables are independent in the population. The Null hypothesis is

H_N: correlation between X and Y in the population is zero
H_A: H_N is false. (6.67)

The test statistic is defined as

$$t = \frac{r\sqrt{(T-2)}}{\sqrt{1-r^2}} . \qquad (6.68)$$

If the Null hypothesis were true and the variables are drawn from a normal distribution, the *t*-statistic defined in (6.68) follows Student's *t* distribution with (T - 2) degrees of freedom. The test rule is to reject the Null hypothesis whenever the computed *t*-statistic exceeds the critical value t_c. The alternative hypothesis implies a two-tailed test. The critical value for a two-tailed test must have two values, + t_c, and - t_c, or the test criterion may be restated as $|t| > t_c$. The probability of Type I error for the two-tailed test associated with the given critical value may be obtained from standard tables.

6.10 The d-Test

A common problem in empirical investigation is to study the effectiveness of a policy in the short run. The time period for such studies is often so limited that almost all the independent variables determining the value of the dependent variable —the variable under investigation—remain unchanged. A regression analysis for such situations seems inappropriate and uneconomical. A simple test based on the differences of the dependent variable before and after the policy implementation provides an answer.

When no other variable except the policy has changed between two time periods, and there is a difference in the observed values of the dependent variable, the difference is attributable to the policy. In real life we rarely observe a situation in which the difference is zero even if the policy did not change. The researcher is, therefore, interested in testing whether the difference is due to randomness of the data or whether there is any systematic difference.

For example, consider the case of free reserves of commercial banks. Suppose the researcher wants to study whether a recent increase in the Central Bank's rediscount value (bank rate) has altered the free reserves of the commercial banks. Suppose no change has occurred in other variables that would determine the bank's policy as to the free reserves. The researcher may collect information from the various banks on their free reserves before and after the change in the rediscount rate. Let the data correspond to *T* commercial banks, and let the difference in free reserves corresponding to the *i*th bank be

$$d_i = \text{(free reserves before the change) — (free reserves after the change).} \qquad (6.69)$$

The Null hypothesis, based on the mean difference for all banks in the total population rather than for the *T* selected banks, states that the mean difference for the population is zero:

H_N: mean of *d* in the population is zero
H_A: H_N is false. (6.70)

The test statistic is based on the observed mean and variance of the *d*'s, corresponding to the sample of T observations:

$$t = \frac{\bar{d}}{\sqrt{\frac{\Sigma(d-\bar{d})^2}{T-1}}} . \qquad (6.71)$$

Under the Null hypothesis, if the *d*'s were drawn randomly from a normal distribution, then the *t*-statistic follows Student's *t* distribution with (T - 1) degrees of freedom. The test rule rejects the Null hypothesis whenever the computed *t*-statistic exceeds the critical value.

6.11 Transformation of the Variables

Some functional forms are expressible as a linear function after a suitable transformation of the variables. For example, equation (6.73) is not a linear function in X_1, but by defining $X_1^{/}$, a new variable, as equal to X_1^2, we may express the equation in linear form as in (6.74).

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_{1t} \qquad (6.72)$$

$$Y_t = \alpha_0 + \alpha_1 X_{1t}^2 + \alpha_2 X_{2t} + \varepsilon_{2t} \qquad (6.73)$$

$$Y_t = \alpha_0 + \alpha_1 X_{1t}^{/} + \alpha_2 X_{2t} + \varepsilon_{2t} \qquad (6.74)$$

Since equation (6.74) is a linear function of its independent variables, $X_1^{/}$ and X_2 , the researcher may estimate it by the standard least squares procedure. In this case the problem is not with the technique of estimation. Since the researcher is generally interested in knowing which of the alternative non-linear forms of the variable X_1 is empirically appropriate, he may treat the alternative forms of the variable as alternative definitions of a specified variable.

When a variable has alternative definitions, its empirically appropriate definition may be obtained by studying the residual sum of squares of the regressions under the various definitions. As long as the dependent variable and the number of parameters are estimated the same, the residual sums of squares are comparable in different equations with different definitions of an independent variable.

For example, coal from different sources have different BTU. Using BTU instead of the amount of coal would improve the estimates. It is the quality of the inputs, not the quantity that counts.

This case is simple because we are dealing with the functional form of an independent variable. A problem frequently encountered in empirical research is the choice between a linear regression and a log-linear regression equation. To answer this question let us consider first the case in which the alternative functional forms are

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_{1t} , \qquad (6.75)$$

$$\log Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \varepsilon_{2t} . \qquad (6.76)$$

In this case the researcher cannot play the game of minimum residual sum of squares, because the dependent variables are different in the two equations.

We can trace the source of our trouble to a scaling factor. The variance of Y changes with the units of measurement of Y, but the variance of log Y does not, because log cY = log c + log Y and the addition of a constant (log c) does not alter the variance. A direct comparison of residual sum of squares is therefore meaningless because by a proper choice of units of measurement one residual sum of squares may be made smaller than the other.

By standardizing the variable Y in such a way that its variance does not change with units of measurement we may bring these two equations onto a common footing. If we do the transformation so that the "Jacobian" of transformation is the same for Y* and log Y*, where Y* is the transformed Y, we can directly compare the residual sums of squares. A transformation of Y that allows such a comparison of the residual sum of squares may be defined as

$$Y_t^* = c . Y_t \ , \qquad (6.77)$$

where

$$c = \exp\left(- \frac{\Sigma \log Y_t}{T} \right) \qquad (6.77a)$$

is the inverse of the geometric mean of Y. By standardizing Y by its geometric mean and defining the standardized value as Y* we may express the two equations (6.75) and (6.76) in terms of Y* rather than Y as

$$Y_t^* = \beta_0 + \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + \varepsilon_{1t}^* \qquad (6.78)$$

$$\log Y_t^* = \alpha_0 + \alpha_1 X_{1t}^* + \alpha_2 X_{2t}^* + \varepsilon_{2t}^* \qquad (6.79)$$

Since the residual sums of squares in these two equations, (6.78) and (6.79), are directly comparable, we choose the functional form yielding the minimum residual sum of squares as the empirically appropriate functional form.

The researcher may use a nonparametric test to see whether the difference between the residual sums of squares in these two functional forms is significant. The test is based on a statistic defined as

$$d \;=\; \frac{T}{2}\left|\log\frac{\Sigma e_{1t}^{*2}}{\Sigma e_{2t}^{*2}}\right| \,, \tag{6.80}$$

where Σe_{1t}^{*2} and Σe_{2t}^{*2} are the residual sums of squares in estimating equations (6.78) and (6.79) respectively. The d statistic follows a chi-squared distribution with one degree of freedom. When the *d* statistic exceeds the chosen critical value, the researcher may reject the null hypothesis that these two functions are empirically equivalent.

This may seem to be an ad hoc procedure, but actually it is similar to so-called “Maximum likelihood estimation,” except that in this case we are not interested in the functional form that maximizes the likelihood value over the entire space. We are choosing one of the two well specified functional forms with the larger likelihood value. The likelihood function may have a higher value outside the ranges of our inquiry, but we are not concerned with them because it may be difficult to interpret their relevance in a practical situation.

Once we understand the procedure of comparison between two regression equations with different dependent variables, we may extend our results to the choice of the linear versus the log-linear functional form. The choice is between the equations

$$Y_t^* \;=\; \beta_0+\beta_1 X_{1t}^*+\beta_2 X_{2t}^*+\varepsilon_{1t}^* \,, \tag{6.81}$$

$$\log Y_t^* \;=\; \gamma_0+\gamma_1 \log X_{1t}^*+\gamma_2\, logX_{2t}^*+\varepsilon_{3t}^* \,. \tag{6.82}$$

If the researcher has standardized his dependent variable by dividing it by its geometric mean, then the two equations are comparable. Using the residual sum of squares as the criterion, we choose the one with the minimum residual sum of squares. In this choice we are combining the rules for the transformation of a dependent variable and of independent variables. To illustrate this procedure let us consider a numerical example.

A production function was estimated for Indian woolen textiles as a log-linear function. We could just as well have expressed the production process as a linear function of the inputs.

$$Q = 4.9669 + 2.3965\, K + 2.7031\, L \quad \Sigma e_{1t}^{*2} = \;1471.0702 \tag{6.83}$$

$$\log Q = 1.6524 + 0.4133 \log K + 0.7082 \log L \;\; \Sigma e_{2t}^{*2} \;=\; 1.7367 \tag{6.84}$$

where Σe_{1t}^{*2} and Σe_{2t}^{*2} are the residual sums of squares in these two equations respectively. Since the dependent variables in equations (6.83) and (6.84) are not the same we cannot directly compare the sums of squares of the residuals. To make them comparable we have to transform the dependent variable Q as

$$Q^* = \mathrm{c} \,.\, Q, \qquad (6.85)$$

where c is the inverse of the geometric mean of Q. The geometric mean of Q is 19.2459, hence c is 0.05196. To be able to compare the residual sums of squares we have to estimate the following regression equations:

$$\begin{aligned} Q^* &= \beta_0^* + \beta_1^* K + \beta_2^* L + \varepsilon_{1t}^* \ , && (6.86) \\ \log Q^* &= \gamma_0^* + \gamma_1^* \log K + \gamma_2^* \log L + \varepsilon_{2t}^* \ . && (6.87) \end{aligned}$$

For our exercise we do not need the estimates but only the residual sums of squares. We know that the variance of log Q* and of log Q is the same, because these Q's are constant multiples of each other. The residual sum of squares in equation (6.84) is the same as in (6.87). We also know that multiplication of the entire equation (6.83) by a constant c is the same as changing the units of measurement of all the variables by the same scale. Hence, the residual sum of squares in (6.86) is the same as in (6.83) if they are expressed in the same units. Thus the sums of squares of the residuals in equations (6.86) and (6.87)—respectively Σe_{1t}^{*2} and Σe_{2t}^{*2} —are

$$\begin{aligned} \Sigma e_{1t}^{*2} &= c^2 . \Sigma e_{1t}^2 = 3.9715 \ , && (6.88) \\ \Sigma e_{2t}^{*2} &= \Sigma e_{2t}^2 = 1.7367 \ . && (6.89) \end{aligned}$$

The sum of squares of the residuals in equation (6.87) is smaller than that of (6.86). Hence, the log-linear function appears empirically more appropriate than the linear function for Indian woolen textiles.

To test whether these two functions are empirically equivalent, let us compute the d statistic:

$$d = \frac{15}{2} \left| \log \frac{3.9715}{1.7367} \right| = 6.2 \ , \qquad (6.90)$$

where the sample size is 15. The d statistic follows the chi-squared distribution with one degree of freedom. The critical value for the 90 percent level of confidence is 2.706. The computed statistic, 6.2, exceeds the critical value. Hence we reject the null hypothesis that these two functions are empirically equivalent with 90 percent confidence.