BBE SEM IV:
BASIC ECONOMETRICS

MULTIPLE REGRESSION ANALYSIS:
THE PROBLEM OF INFERENCE

Textbook: Damodar N. Gujarati (2004) *Basic Econometrics*, 4th edition, The McGraw-Hill Companies

8.1 THE NORMALITY ASSUMPTION

- We continue to assume that the u_i follow the normal distribution with zero mean and constant variance σ^2.
- With normality assumption we find that the OLS estimators of the partial regression coefficients are best linear unbiased estimators (BLUE).

8.1 THE NORMALITY ASSUMPTION

Moreover, the estimators $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_1$ are themselves normally distributed with means equal to true β_2, β_3, and β_1 and the variances given in Chapter 7. Furthermore, $(n-3)\hat{\sigma}^2/\sigma^2$ follows the χ^2 distribution with $n-3$ df, and the three OLS estimators are distributed independently of $\hat{\sigma}^2$. The proofs follow the two-variable case discussed in Appendix 3. As a result and following Chapter 5, one can show that, upon replacing σ^2 by its unbiased estimator $\hat{\sigma}^2$ in the computation of the standard errors, each of the following variables

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)}$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\text{se}(\hat{\beta}_2)}$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{\text{se}(\hat{\beta}_3)}$$

follows the t distribution with $n-3$ df.

8.1 THE NORMALITY ASSUMPTION

Note that the df are now $n - 3$ because in computing $\sum \hat{u}_i^2$ and hence $\hat{\sigma}^2$ we first need to estimate the three partial regression coefficients, which therefore put three restrictions on the residual sum of squares (RSS) (following this logic in the four-variable case there will be $n - 4$ df, and so on). Therefore, the t distribution can be used to establish confidence intervals as well as test statistical hypotheses about the true population partial regression coefficients. Similarly, the χ^2 distribution can be used to test hypotheses about the true σ^2. To demonstrate the actual mechanics, we use the following illustrative example.

8.2 EXAMPLE 8.1: CHILD MORTALITY EXAMPLE REVISITED

In Chapter 7 we regressed child mortality (CM) on per capita GNP (PGNP) and the female literacy rate (FLR) for a sample of 64 countries. The regression results given in (7.6.2) are reproduced below with some additional information:

$$
\begin{aligned}
\widehat{\text{CM}}_i &= 263.6416 \quad - \quad 0.0056\ \text{PGNP}_i - \quad 2.2316\ \text{FLR}_i \\
\text{se} &= (11.5932) \quad (0.0019) \quad (0.2099) \\
t &= (22.7411) \quad (-2.8187) \quad (-10.6293) \qquad \textbf{(8.2.1)} \\
p\ \text{value} &= (0.0000)^* \quad (0.0065) \quad (0.0000)^* \\
R^2 &= 0.7077 \qquad \bar{R}^2 = 0.6981
\end{aligned}
$$

where * denotes extremely low value.

8.2 EXAMPLE 8.1: CHILD MORTALITY EXAMPLE REVISITED

In Eq. (8.2.1) we have followed the format first introduced in Eq. (5.11.1), where the figures in the first set of parentheses are the estimated standard errors, those in the second set are the t values under the null hypothesis that the relevant population coefficient has a value of zero, and those in the third are the estimated p values. Also given are R^2 and adjusted R^2 values. We have already interpreted this regression in Example 7.1.

What about the statistical significance of the observed results? Consider, for example, the coefficient of PGNP of −0.0056. Is this coefficient statistically significant, that is, statistically different from zero? Likewise, is the coefficient of FLR of −2.2316 statistically significant? Are both coefficients statistically significant? To answer this and related questions, let us first consider the kinds of hypothesis testing that one may encounter in the context of a multiple regression model.

8.3 HYPOTHESIS TESTING IN MULTIPLE REGRESSION: GENERAL COMMENTS

Once we go beyond the simple world of the two-variable linear regression model, hypothesis testing assumes several interesting forms, such as the following:

1. Testing hypotheses about an individual partial regression coefficient (Section 8.4)

2. Testing the overall significance of the estimated multiple regression model, that is, finding out if all the partial slope coefficients are simultaneously equal to zero (Section 8.5)

3. Testing that two or more coefficients are equal to one another (Section 8.6)

4. Testing that the partial regression coefficients satisfy certain restrictions (Section 8.7)

5. Testing the stability of the estimated regression model over time or in different cross-sectional units (Section 8.8)

6. Testing the functional form of regression models (Section 8.9)

Since testing of one or more of these types occurs so commonly in empirical analysis, we devote a section to each type.

8.4 HYPOTHESIS TESTING ABOUT INDIVIDUAL REGRESSION COEFFICIENTS

If we invoke the assumption that $u_i \sim N(0, \sigma^2)$, then, as noted in Section 8.1, we can use the t test to test a hypothesis about any *individual* partial regression coefficient. To illustrate the mechanics, consider the child mortality regression, (8.2.1). Let us postulate that

$$H_0: \beta_2 = 0 \qquad \text{and} \qquad H_1: \beta_2 \neq 0$$

The null hypothesis states that, with X_3 (female literacy rate) held constant, X_2 (PGNP) has no (linear) influence on Y (child mortality).[2] To test the null hypothesis, we use the t test given in (8.1.2). Following Chapter 5 (see Table 5.1), if the computed t value exceeds the critical t value at the chosen level of significance, we may reject the null hypothesis; otherwise, we may not reject it. For our illustrative example, using (8.1.2) and noting that $\beta_2 = 0$ under the null hypothesis, we obtain

$$t = \frac{-0.0056}{0.0020} = -2.8187 \qquad \textbf{(8.4.1)}$$

as shown in Eq. (8.2.1).

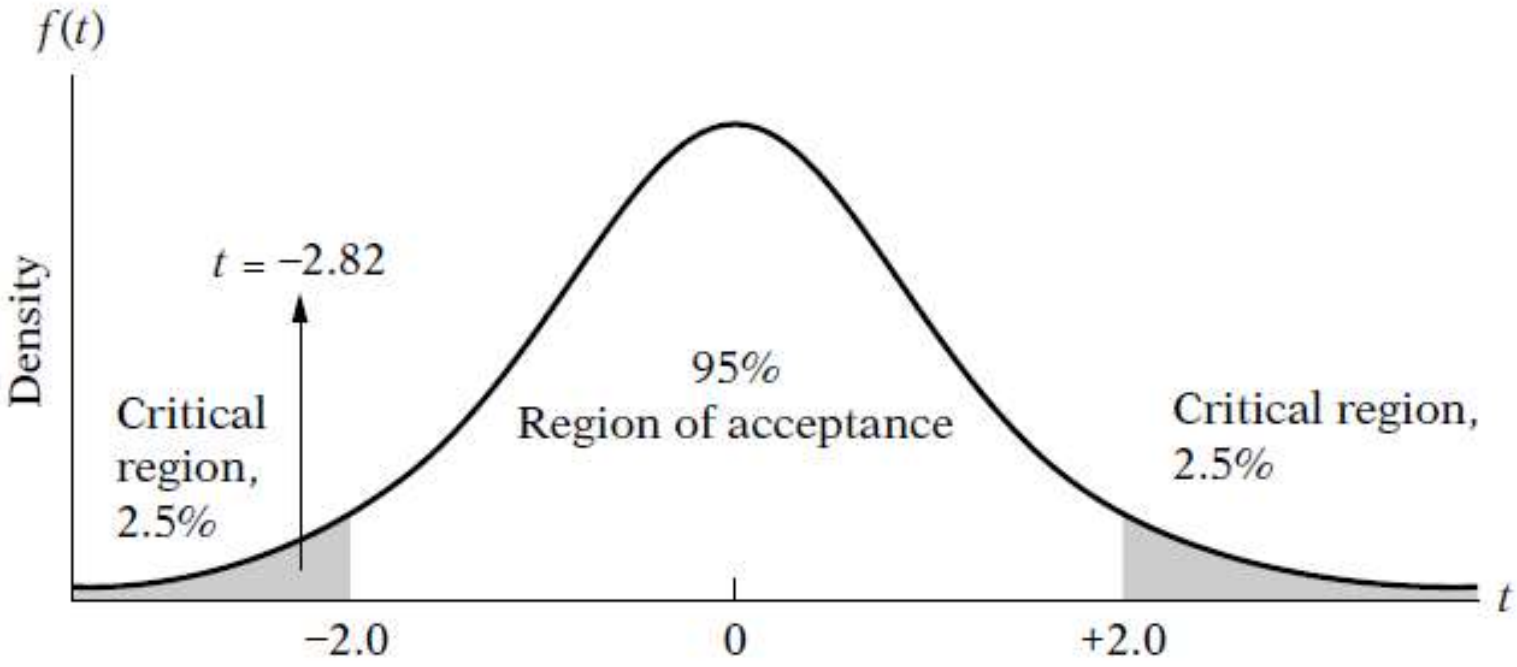
8.4 HYPOTHESIS TESTING ABOUT INDIVIDUAL REGRESSION COEFFICIENTS

Notice that we have 64 observations. Therefore, the degrees of freedom in this example are 61 (why?). If you refer to the t table given in **Appendix D**, we do not have data corresponding to 61 df. The closest we have are for 60 df. If we use these df, and assume α, the level of significance (i.e., the probability of committing a Type I error) of 5 percent, the critical t value is 2.0 for a two-tail test (look up $t_{\alpha/2}$ for 60 df) or 1.671 for a one-tail test (look up t_{α} for 60 df).

For our example, the alternative hypothesis is two-sided. Therefore, we use the two-tail t value. Since the computed t value of 2.8187 (in absolute terms) exceeds the critical t value of 2, we can reject the null hypothesis that PGNP has no effect on child mortality. To put it more positively, with the female literacy rate held constant, per capita GNP has a significant (negative) effect on child mortality, as one would expect a priori. Graphically, the situation is as shown in Figure 8.1.

8.4 HYPOTHESIS TESTING ABOUT INDIVIDUAL REGRESSION COEFFICIENTS

FIGURE 8.1



The 95% confidence interval for t (60 df).

8.4 HYPOTHESIS TESTING ABOUT INDIVIDUAL REGRESSION COEFFICIENTS

In practice, one does not have to assume a particular value of α to conduct hypothesis testing. One can simply use the p value given in (8.2.2), which in the present case is 0.0065. The interpretation of this p value (i.e., the exact level of significance) is that if the null hypothesis were true, the probability of obtaining a t value of as much as 2.8187 or greater (in absolute terms) is only 0.0065 or 0.65 percent, which is indeed a small probability, much smaller than the artificially adopted value of $\alpha = 5\%$.

8.4 HYPOTHESIS TESTING ABOUT INDIVIDUAL REGRESSION COEFFICIENTS

In Chapter 5 we saw the intimate connection between hypothesis testing and confidence interval estimation. For our example, the 95% confidence interval for β_2 is:

$$\hat{\beta}_2 - t_{\alpha/2}\,\text{se}\,(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2}\,\text{se}\,(\hat{\beta}_2)$$

which in our example becomes

$$-0.0056 - 2(0.0020) \leq \beta_2 \leq -0.0056 + 2(0.0020)$$

that is,

$$-0.0096 \leq \beta_2 \leq -0.0016 \qquad \textbf{(8.4.2)}$$

that is, the interval, −0.0096 to −0.0016 includes the true β_2 coefficient with 95% confidence coefficient. Thus, if 100 samples of size 64 are selected and 100 confidence intervals like (8.4.2) are constructed, we expect 95 of them to contain the true population parameter β_2. Since the interval (8.4.2) does not include the null-hypothesized value of zero, we can reject the null hypothesis that the true β_2 is zero with 95% confidence.

8.4 HYPOTHESIS TESTING ABOUT INDIVIDUAL REGRESSION COEFFICIENTS

$$\begin{aligned}
\widehat{\text{CM}}_i &= 263.6416 \quad - \quad 0.0056\ \text{PGNP}_i - \quad 2.2316\ \text{FLR}_i \\
\text{se} &= (11.5932) \quad (0.0019) \quad (0.2099) \\
t &= (22.7411) \quad (-2.8187) \quad (-10.6293) \qquad \textbf{(8.2.1)} \\
p\ \text{value} &= (0.0000)^* \quad (0.0065) \quad (0.0000)^* \\
R^2 &= 0.7077 \qquad \bar{R}^2 = 0.6981
\end{aligned}$$

Following the procedure just described, we can test hypotheses about the other parameters of our child mortality regression model. The necessary data are already provided in Eq. (8.2.1). For example, suppose we want to test the hypothesis that, with the influence of PGNP held constant, the female literacy rate has no effect whatsoever on child mortality. We can confidently reject this hypothesis, for under this null hypothesis the p value of obtaining an absolute t value of as much as 10.6 or greater is practically zero.

8.5 TESTING THE OVERALL SIGNIFICANCE OF THE SAMPLE REGRESSION

Throughout the previous section we were concerned with testing the significance of the estimated partial regression coefficients individually, that is, under the separate hypothesis that each true population partial regression coefficient was zero. But now consider the following hypothesis:

$$H_0: \beta_2 = \beta_3 = 0 \qquad \textbf{(8.5.1)}$$

This null hypothesis is a joint hypothesis that β_2 and β_3 are jointly or simultaneously equal to zero. A test of such a hypothesis is called a test of the **overall significance** of the observed or estimated regression line, that is, whether Y is linearly related to both X_2 and X_3.

The Analysis of Variance Approach to Testing the Overall Significance of an Observed Multiple Regression: The F Test

we cannot use the usual t test to test the joint hypothesis that the true partial slope coefficients are zero simultaneously. However, this joint hypothesis can be tested by the **analysis of variance** (ANOVA) technique first introduced in Section 5.9, which can be demonstrated as follows.

ANOVA TABLE FOR THE THREE-VARIABLE REGRESSION

Source of variation	SS	df	MSS
Due to regression (ESS)	$\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$	2	$\dfrac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{2}$
Due to residual (RSS)	$\sum \hat{u}_i^2$	$n - 3$	$\hat{\sigma}^2 = \dfrac{\sum \hat{u}_i^2}{n - 3}$
Total	$\sum y_i^2$	$n - 1$	

The Analysis of Variance Approach to Testing the Overall Significance of an Observed Multiple Regression: The F Test

Recall the identity

$$\sum y_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i} + \sum \hat{u}_i^2 \qquad \textbf{(8.5.2)}$$

$$\text{TSS} = \quad \text{ESS} \quad + \text{RSS}$$

TSS has, as usual, $n-1$ df and RSS has $n-3$ df for reasons already discussed. ESS has 2 df since it is a function of $\hat{\beta}_2$ and $\hat{\beta}_3$. Therefore, following the ANOVA procedure discussed in Section 5.9, we can set up Table 8.1.

Now it can be shown[6] that, under the assumption of normal distribution for u_i and the null hypothesis $\beta_2 = \beta_3 = 0$, the variable

$$F = \frac{(\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i})/2}{\sum \hat{u}_i^2/(n-3)} = \frac{\text{ESS/df}}{\text{RSS/df}} \qquad \textbf{(8.5.3)}$$

is distributed as the F distribution with 2 and $n-3$ df.

The Analysis of Variance Approach to Testing the Overall Significance of an Observed Multiple Regression: The F Test

Therefore, the F value of (8.5.3) provides a test of the null hypothesis that the true slope coefficients are simultaneously zero. If the F value computed from (8.5.3) exceeds the critical F value from the F table at the α percent level of significance, we reject H_0; otherwise we do not reject it. Alternatively, if the p value of the observed F is sufficiently low, we can reject H_0.

TABLE 8.3 ANOVA TABLE FOR THE CHILD MORTALITY EXAMPLE

Source of variation	SS	df	MSS
Due to regression	257,362.4	2	128,681.2
Due to residuals	106,315.6	61	1742.88
Total	363,678	63	

The Analysis of Variance Approach to Testing the Overall Significance of an Observed Multiple Regression: The F Test

Using (8.5.3), we obtain

$$F = \frac{128{,}681.2}{1742.88} = 73.8325 \qquad \textbf{(8.5.6)}$$

The p value of obtaining an F value of as much as 73.8325 or greater is almost zero, leading to the rejection of the hypothesis that together PGNP and FLR have no effect on child mortality. If you were to use the conventional 5 percent level-of-significance value, the critical F value for 2 df in the numerator and 60 df in the denominator (the actual df, however, are 61) is about 3.15 or about 4.98 if you were to use the 1 percent level of significance. Obviously, the observed F of about 74 far exceeds any of these critical F values.

We can generalize the preceding F-testing procedure as follows.

Testing the Overall Significance of a Multiple Regression: The F Test

Decision Rule. Given the k-variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

To test the hypothesis

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

(i.e., all slope coefficients are simultaneously zero) versus

H_1: Not all slope coefficients are simultaneously zero

compute

$$F = \frac{\text{ESS/df}}{\text{RSS/df}} = \frac{\text{ESS}/(k-1)}{\text{RSS}/(n-k)} \qquad \textbf{(8.5.7)}$$

If $F > F_\alpha(k-1, n-k)$, reject H_0; otherwise you do not reject it, where $F_\alpha(k-1, n-k)$ is the *critical* F value at the α level of significance and $(k-1)$ numerator df and $(n-k)$ denominator df. Alternatively, if the p value of F obtained from (8.5.7) is sufficiently low, one can reject H_0.

Testing the Overall Significance of a Multiple Regression: The F Test

Individual versus Joint Testing of Hypotheses. In Section 8.4 we discussed the test of significance of a single regression coefficient and in Section 8.5 we have discussed the joint or overall test of significance of the estimated regression (i.e., all slope coefficients are simultaneously equal to zero). **We reiterate that these tests are different.** Thus, on the basis of the t test or confidence interval (of Section 8.4) it is possible to accept the hypothesis that a particular slope coefficient, β_k, is zero, and yet reject the joint hypothesis that all slope coefficients are zero.

> The lesson to be learned is that the joint "message" of individual confidence intervals is no substitute for a joint confidence region [implied by the F test] in performing joint tests of hypotheses and making joint confidence statements.[8]

An Important Relationship between R^2 and F

There is an intimate relationship between the coefficient of determination R^2 and the F test used in the analysis of variance. Assuming the normal distribution for the disturbances u_i and the null hypothesis that $\beta_2 = \beta_3 = 0$, we have seen that

$$F = \frac{\text{ESS}/2}{\text{RSS}/(n-3)} \qquad \textbf{(8.5.8)}$$

is distributed as the F distribution with 2 and $n - 3$ df.

An Important Relationship between R^2 and F

More generally, in the k-variable case (including intercept), if we assume that the disturbances are normally distributed and that the null hypothesis is

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0 \qquad \textbf{(8.5.9)}$$

then it follows that

$$F = \frac{\text{ESS}/(k-1)}{\text{RSS}/(n-k)} \qquad \textbf{(8.5.7) = (8.5.10)}$$

follows the F distribution with $k-1$ and $n-k$ df. (*Note:* The total number of parameters to be estimated is k, of which one is the intercept term.)

An Important Relationship between R^2 and F

Let us manipulate (8.5.10) as follows:

$$
\begin{aligned}
F &= \frac{n-k}{k-1}\frac{\text{ESS}}{\text{RSS}} \\
&= \frac{n-k}{k-1}\frac{\text{ESS}}{\text{TSS}-\text{ESS}} \\
&= \frac{n-k}{k-1}\frac{\text{ESS/TSS}}{1-(\text{ESS/TSS})} \\
&= \frac{n-k}{k-1}\frac{R^2}{1-R^2} \\
&= \frac{R^2/(k-1)}{(1-R^2)/(n-k)}
\end{aligned}
$$

where use is made of the definition R^2 = ESS/TSS. Equation on the left shows how F and R^2 are related. These two vary directly. When R^2 = 0, F is zero ipso facto. The larger the R^2, the greater the F value. In the limit, when R^2 = 1, F is infinite. Thus the F test, which is a measure of the overall significance of the estimated regression, is also a test of significance of R^2. In other words, testing the null hypothesis (8.5.9) is equivalent to testing the null hypothesis that (the population) R^2 is zero.

An Important Relationship between R^2 and F

For the three-variable case (8.5.11) becomes

$$F = \frac{R^2/2}{(1 - R^2)/(n - 3)} \qquad \textbf{(8.5.12)}$$

By virtue of the close connection between F and R^2, the ANOVA Table 8.1 can be recast as Table 8.4.

For our illustrative example, using (8.5.12) we obtain:

$$F = \frac{0.7077/2}{(1 - 0.7077)/61} = 73.8726$$

which is about the same as obtained before, except for the rounding errors.

One advantage of the F test expressed in terms of R^2 is its ease of computation: All that one needs to know is the R^2 value. Therefore, the overall F test of significance given in (8.5.7) can be recast in terms of R^2 as shown in Table 8.4.

Testing the Overall Significance of a Multiple Regression in Terms of R^2

Decision Rule. Testing the overall significance of a regression in terms of R^2: Alternative but equivalent test to (8.5.7).

Given the k-variable regression model:

$$Y_i = \beta_i + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_x X_{ki} + u_i$$

To test the hypothesis

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

versus

$$H_1: \text{Not all slope coefficients are simultaneously zero}$$

compute

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \qquad \textbf{(8.5.13)}$$

If $F > F_{\alpha(k-1,n-k)}$, reject H_0; otherwise you may accept H_0 where $F_{\alpha(k-1,n-k)}$ is the critical F value at the α level of significance and $(k-1)$ numerator df and $(n-k)$ denominator df. Alternatively, if the p value of F obtained from (8.5.13) is sufficiently low, reject H_0.

When to Add a New Variable. The F-test procedure just outlined provides a formal method of deciding whether a variable should be added to a regression model. Often researchers are faced with the task of choosing from several competing models **involving the same dependent variable** but with different explanatory variables. As a matter of ad hoc choice (because very often the theoretical foundation of the analysis is weak), these researchers frequently choose the model that gives the highest adjusted R^2. Therefore, if the inclusion of a variable increases $\bar{R}^2$, it is retained in the model although it does not reduce RSS significantly in the statistical sense.

The question then becomes: When does the adjusted R^2 increase? It can be shown that $\bar{R}^2$ *will increase if the t value of the coefficient of the newly added variable is larger than 1 in absolute value*, where the t value is computed under the hypothesis that the population value of the said coefficient is zero [i.e., the t value computed from (5.3.2) under the hypothesis that the true β value is zero].[10] The preceding criterion can also be stated differently: $\bar{R}^2$ *will increase with the addition of an extra explanatory variable only if the* $F(=t^2)$ *value of that variable exceeds 1.*

Applying either criterion, the FLR variable in our child mortality example with a t value of -10.6293 or an F value of 112.9814 should increase $\bar{R}^2$, which indeed it does—when FLR is added to the model, $\bar{R}^2$ increases from 0.1528 to 0.6981.