5

TWO-VARIABLE REGRESSION: INTERVAL ESTIMATION AND HYPOTHESIS TESTING

> Beware of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confession obtained under duress may not be admissible in the court of scientific opinion.[1]

As pointed out in Chapter 4, estimation and hypothesis testing constitute the two major branches of classical statistics. The theory of estimation consists of two parts: point estimation and interval estimation. We have discussed point estimation thoroughly in the previous two chapters where we introduced the OLS and ML methods of point estimation. In this chapter we first consider interval estimation and then take up the topic of hypothesis testing, a topic intimately related to interval estimation.

5.1 STATISTICAL PREREQUISITES

Before we demonstrate the actual mechanics of establishing confidence intervals and testing statistical hypotheses, it is assumed that the reader is familiar with the fundamental concepts of probability and statistics. Although not a substitute for a basic course in statistics, **Appendix A** provides the essentials of statistics with which the reader should be totally familiar. Key concepts such as **probability, probability distributions, Type I and Type II errors, level of significance, power of a statistical test,** and **confidence interval** are crucial for understanding the material covered in this and the following chapters.

[1] Stephen M. Stigler, "Testing Hypothesis or Fitting Models? Another Look at Mass Extinctions," in Matthew H. Nitecki and Antoni Hoffman, eds., *Neutral Models in Biology,* Oxford University Press, Oxford, 1987, p. 148.

5.2 INTERVAL ESTIMATION: SOME BASIC IDEAS

To fix the ideas, consider the hypothetical consumption-income example of Chapter 3. Equation (3.6.2) shows that the estimated marginal propensity to consume (MPC) β_2 is 0.5091, which is a single (point) estimate of the unknown population MPC β_2. How reliable is this estimate? As noted in Chapter 3, because of sampling fluctuations, a single estimate is likely to differ from the true value, although in repeated sampling its mean value is expected to be equal to the true value. [*Note:* $E(\hat{\beta}_2) = \beta_2$.] Now in statistics the reliability of a point estimator is measured by its standard error. Therefore, instead of relying on the point estimate alone, we may construct an interval around the point estimator, say within two or three standard errors on either side of the point estimator, such that this interval has, say, 95 percent probability of including the true parameter value. This is roughly the idea behind **interval estimation.**

To be more specific, assume that we want to find out how "close" is, say, $\hat{\beta}_2$ to β_2. For this purpose we try to find out two positive numbers δ and α, the latter lying between 0 and 1, such that the probability that the **random interval** $(\hat{\beta}_2 - \delta, \hat{\beta}_2 + \delta)$ contains the true β_2 is $1 - \alpha$. Symbolically,

$$\Pr\,(\hat{\beta}_2 - \delta \le \beta_2 \le \hat{\beta}_2 + \delta) = 1 - \alpha \tag{5.2.1}$$

Such an interval, if it exists, is known as a **confidence interval;** $1 - \alpha$ is known as the **confidence coefficient;** and α $(0 < \alpha < 1)$ is known as the **level of significance.**[2] The endpoints of the confidence interval are known as the **confidence limits** (also known as *critical* values), $\hat{\beta}_2 - \delta$ being the **lower confidence** *limit* and $\hat{\beta}_2 + \delta$ the **upper confidence** *limit*. In passing, note that in practice α and $1 - \alpha$ are often expressed in percentage forms as 100α and $100(1 - \alpha)$ percent.

Equation (5.2.1) shows that an **interval estimator,** in contrast to a point estimator, is an interval constructed in such a manner that it has a specified probability $1 - \alpha$ of including within its limits the true value of the parameter. For example, if $\alpha = 0.05$, or 5 percent, (5.2.1) would read: The probability that the (random) interval shown there includes the true β_2 is 0.95, or 95 percent. The interval estimator thus gives a range of values within which the true β_2 may lie.

It is very important to know the following aspects of interval estimation:

1. Equation (5.2.1) does not say that the probability of β_2 lying between the given limits is $1 - \alpha$. Since β_2, although an unknown, is assumed to be some fixed number, either it lies in the interval or it does not. What (5.2.1)

[2] Also known as the **probability of committing a Type I error.** A Type I error consists in rejecting a true hypothesis, whereas a Type II error consists in accepting a false hypothesis. (This topic is discussed more fully in **App. A.**) The symbol α is also known as the **size of the (statistical) test.**

states is that, for the method described in this chapter, the probability of constructing an interval that contains β_2 is $1 - \alpha$.

2. The interval (5.2.1) is a **random interval;** that is, it will vary from one sample to the next because it is based on $\hat{\beta}_2$, which is random. (Why?)

3. Since the confidence interval is random, the probability statements attached to it should be understood in the long-run sense, that is, repeated sampling. More specifically, (5.2.1) means: If in repeated sampling confidence intervals like it are constructed a great many times on the $1 - \alpha$ probability basis, then, in the long run, on the average, such intervals will enclose in $1 - \alpha$ of the cases the true value of the parameter.

4. As noted in 2, the interval (5.2.1) is random so long as $\hat{\beta}_2$ is not known. But once we have a specific sample and once we obtain a specific numerical value of $\hat{\beta}_2$, the interval (5.2.1) is no longer random; it is fixed. In this case, we **cannot** make the probabilistic statement (5.2.1); that is, we cannot say that the probability is $1 - \alpha$ that a given *fixed* interval includes the true β_2. In this situation β_2 is either in the fixed interval or outside it. Therefore, the probability is either 1 or 0. Thus, for our hypothetical consumption-income example, if the 95% confidence interval were obtained as $(0.4268 \le \beta_2 \le 0.5914)$, as we do shortly in (5.3.9), we **cannot** say the probability is 95% that this interval includes the true β_2. That probability is either 1 or 0.

How are the confidence intervals constructed? From the preceding discussion one may expect that if the **sampling or probability distributions** of the estimators are known, one can make confidence interval statements such as (5.2.1). In Chapter 4 we saw that under the assumption of normality of the disturbances u_i the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are themselves normally distributed and that the OLS estimator $\hat{\sigma}^2$ is related to the χ^2 (chi-square) distribution. It would then seem that the task of constructing confidence intervals is a simple one. And it is!

5.3 CONFIDENCE INTERVALS FOR REGRESSION COEFFICIENTS β_1 AND β_2

Confidence Interval for β_2

It was shown in Chapter 4, Section 4.3, that, with the normality assumption for u_i, the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are themselves normally distributed with means and variances given therein. Therefore, for example, the variable

$$\begin{aligned} Z &= \frac{\hat{\beta}_2 - \beta_2}{\operatorname{se}(\hat{\beta}_2)} \\ &= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\sigma} \end{aligned} \tag{5.3.1}$$

as noted in (4.3.6), is a standardized normal variable. It therefore seems that we can use the normal distribution to make probabilistic statements about β_2 provided the true population variance σ^2 is known. If σ^2 is known, an important property of a normally distributed variable with mean μ and variance σ^2 is that the area under the normal curve between $\mu \pm \sigma$ is about 68 percent, that between the limits $\mu \pm 2\sigma$ is about 95 percent, and that between $\mu \pm 3\sigma$ is about 99.7 percent.

But σ^2 is rarely known, and in practice it is determined by the unbiased estimator $\hat{\sigma}^2$. If we replace σ by $\hat{\sigma}$, (5.3.1) may be written as

$$\begin{aligned} t &= \frac{\hat{\beta}_2 - \beta_2}{\text{se}\,(\hat{\beta}_2)} = \frac{\text{estimator} - \text{parameter}}{\text{estimated standard error of estimator}} \\ &= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}} \end{aligned} \tag{5.3.2}$$

where the se $(\hat{\beta}_2)$ now refers to the estimated standard error. It can be shown (see Appendix 5A, Section 5A.2) that the t variable thus defined follows the t distribution with $n - 2$ df. [Note the difference between (5.3.1) and (5.3.2).] Therefore, instead of using the normal distribution, we can use the t distribution to establish a confidence interval for β_2 as follows:

$$\Pr\,(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1 - \alpha \tag{5.3.3}$$

where the t value in the middle of this double inequality is the t value given by (5.3.2) and where $t_{\alpha/2}$ is the value of the t variable obtained from the t distribution for $\alpha/2$ level of significance and $n - 2$ df; it is often called the **critical** t value at $\alpha/2$ level of significance. Substitution of (5.3.2) into (5.3.3) yields

$$\Pr\left[-t_{\alpha/2} \le \frac{\hat{\beta}_2 - \beta_2}{\text{se}\,(\hat{\beta}_2)} \le t_{\alpha/2}\right] = 1 - \alpha \tag{5.3.4}$$

Rearranging (5.3.4), we obtain

$$\Pr\,[\hat{\beta}_2 - t_{\alpha/2}\,\text{se}\,(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2}\,\text{se}\,(\hat{\beta}_2)] = 1 - \alpha \qquad \mathbf{(5.3.5)}^3$$

[3] Some authors prefer to write (5.3.5) with the df explicitly indicated. Thus, they would write

$$\Pr\,[\hat{\beta}_2 - t_{(n-2),\alpha/2}\,\text{se}\,(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{(n-2)\alpha/2}\,\text{se}\,(\hat{\beta}_2)] = 1 - \alpha$$

But for simplicity we will stick to our notation; the context clarifies the appropriate df involved.

Equation (5.3.5) provides a $100(1-\alpha)$ percent **confidence interval** for β_2, which can be written more compactly as

$$100(1-\alpha)\% \text{ confidence interval for } \beta_2:$$

$$\hat{\beta}_2 \pm t_{\alpha/2}\,\text{se}\,(\hat{\beta}_2) \tag{5.3.6}$$

Arguing analogously, and using (4.3.1) and (4.3.2), we can then write:

$$\Pr\,[\hat{\beta}_1 - t_{\alpha/2}\,\text{se}\,(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2}\,\text{se}\,(\hat{\beta}_1)] = 1-\alpha \tag{5.3.7}$$

or, more compactly,

$$100(1-\alpha)\% \text{ confidence interval for } \beta_1:$$

$$\hat{\beta}_1 \pm t_{\alpha/2}\,\text{se}\,(\hat{\beta}_1) \tag{5.3.8}$$

Notice an important feature of the confidence intervals given in (5.3.6) and (5.3.8): In both cases *the width of the confidence interval is proportional to the standard error of the estimator.* That is, the larger the standard error, the larger is the width of the confidence interval. Put differently, the larger the standard error of the estimator, the greater is the uncertainty of estimating the true value of the unknown parameter. Thus, the standard error of an estimator is often described as a measure of the **precision** of the estimator, i.e., how precisely the estimator measures the true population value.

Returning to our illustrative consumption–income example, in Chapter 3 (Section 3.6) we found that $\hat{\beta}_2 = 0.5091$, $\text{se}\,(\hat{\beta}_2) = 0.0357$, and df = 8. If we assume $\alpha = 5\%$, that is, 95% confidence coefficient, then the t table shows that for 8 df the **critical** $t_{\alpha/2} = t_{0.025} = 2.306$. Substituting these values in (5.3.5), the reader should verify that the 95% confidence interval for β_2 is as follows:

$$0.4268 \le \beta_2 \le 0.5914 \tag{5.3.9}$$

Or, using (5.3.6), it is

$$0.5091 \pm 2.306(0.0357)$$

that is,

$$0.5091 \pm 0.0823 \tag{5.3.10}$$

The interpretation of this confidence interval is: Given the confidence coefficient of 95%, in the long run, in 95 out of 100 cases intervals like

(0.4268, 0.5914) will contain the true β_2. But, as warned earlier, we cannot say that the probability is 95 percent that the specific interval (0.4268 to 0.5914) contains the true β_2 because this interval is now fixed and no longer random; therefore, β_2 either lies in it or does not: The probability that the specified fixed interval includes the true β_2 is therefore 1 or 0.

Confidence Interval for β_1

Following (5.3.7), the reader can easily verify that the 95% confidence interval for β_1 of our consumption–income example is

$$9.6643 \le \beta_1 \le 39.2448 \tag{5.3.11}$$

Or, using (5.3.8), we find it is

$$24.4545 \pm 2.306(6.4138)$$

that is,

$$24.4545 \pm 14.7902 \tag{5.3.12}$$

Again you should be careful in interpreting this confidence interval. In the long run, in 95 out of 100 cases intervals like (5.3.11) will contain the true β_1; the probability that this particular fixed interval includes the true β_1 is either 1 or 0.

Confidence Interval for β_1 and β_2 Simultaneously

There are occasions when one needs to construct a *joint confidence interval* for β_1 and β_2 such that with a confidence coefficient $(1 - \alpha)$, say, 95%, that interval includes β_1 and β_2 simultaneously. Since this topic is involved, the interested reader may want to consult appropriate references.[4] We will touch on this topic briefly in Chapters 8 and 10.

5.4 CONFIDENCE INTERVAL FOR σ^2

As pointed out in Chapter 4, Section 4.3, under the normality assumption, the variable

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \tag{5.4.1}$$

[4] For an accessible discussion, see John Neter, William Wasserman, and Michael H. Kutner, *Applied Linear Regression Models*, Richard D. Irwin, Homewood, Ill., 1983, Chap. 5.

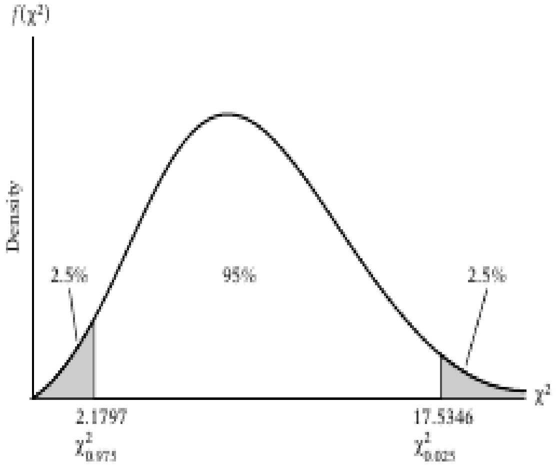

FIGURE 5.1 The 95% confidence interval for χ^2 (8 df).

follows the χ^2 distribution with $n - 2$ df.[5] Therefore, we can use the χ^2 distribution to establish a confidence interval for σ^2

$$\Pr\left(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}\right) = 1 - \alpha \qquad \textbf{(5.4.2)}$$

where the χ^2 value in the middle of this double inequality is as given by (5.4.1) and where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are two values of χ^2 (the **critical** χ^2 values) obtained from the chi-square table for $n - 2$ df in such a manner that they cut off $100(\alpha/2)$ percent tail areas of the χ^2 distribution, as shown in Figure 5.1.

Substituting χ^2 from (5.4.1) into (5.4.2) and rearranging the terms, we obtain

$$\Pr\left[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right] = 1 - \alpha \qquad \textbf{(5.4.3)}$$

which gives the $100(1 - \alpha)\%$ confidence interval for σ^2.

To illustrate, consider this example. From Chapter 3, Section 3.6, we obtain $\hat{\sigma}^2 = 42.1591$ and df $= 8$. If α is chosen at 5 percent, the chi-square table for 8 df gives the following critical values: $\chi^2_{0.025} = 17.5346$, and $\chi^2_{0.975} = 2.1797$. These values show that the probability of a chi-square value exceeding 17.5346 is 2.5 percent and that of 2.1797 is 97.5 percent. Therefore, the interval between these two values is the 95% confidence interval for χ^2, as shown diagrammatically in Figure 5.1. (Note the skewed characteristic of the chi-square distribution.)

[5] For proof, see Robert V. Hogg and Allen T. Craig, *Introduction to Mathematical Statistics,* 2d ed., Macmillan, New York, 1965, p. 144.

Substituting the data of our example into (5.4.3), the reader should verify that the 95% confidence interval for σ^2 is as follows:

$$19.2347 \le \sigma^2 \le 154.7336 \qquad \textbf{(5.4.4)}$$

The interpretation of this interval is: If we establish 95% confidence limits on σ^2 and if we maintain a priori that these limits will include true σ^2, we shall be right in the long run 95 percent of the time.