CHAPTER 5

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

This chapter continues the treatment of linear regression with a single regressor. Chapter 4 explained how the OLS estimator $\hat{\beta}_1$ of the slope coefficient β_1 differs from one sample to the next—that is, how $\hat{\beta}_1$ has a sampling distribution. In this chapter, we show how knowledge of this sampling distribution can be used to make statements about β_1 that accurately summarize the sampling uncertainty. The starting point is the standard error of the OLS estimator, which measures the spread of the sampling distribution of $\hat{\beta}_1$. Section 5.1 provides an expression for this standard error (and for the standard error of the OLS estimator of the intercept), then shows how to use $\hat{\beta}_1$ and its standard error to test hypotheses. Section 5.2 explains how to construct confidence intervals for β_1. Section 5.3 takes up the special case of a binary regressor.

Sections 5.1–5.3 assume that the three least squares assumptions of Chapter 4 hold. If, in addition, some stronger conditions hold, then some stronger results can be derived regarding the distribution of the OLS estimator One of these stronger conditions is that the errors are homoskedastic, a concept introduced in Section 5.4. Section 5.5 presents the Gauss-Markov theorem, which states that, under certain conditions, OLS is efficient (has the smallest variance) among a certain class of estimators. Section 5.6 discusses the distribution of the OLS estimator when the population distribution of the regression errors is normal.

5.1 Testing Hypotheses About One of the Regression Coefficients

Your client, the superintendent, calls you with a problem. She has an angry taxpayer in her office who asserts that cutting class size will not help boost test scores, so that reducing them further is a waste of money. Class size, the taxpayer claims, has no effect on test scores.

The taxpayer's claim can be rephrased in the language of regression analysis. Because the effect on test scores of a unit change in class size is $\beta_{ClassSize}$, the taxpayer is asserting that the population regression line is flat—that is, the slope $\beta_{ClassSize}$ of the population regression line is zero. Is there, the superintedent asks, evidence in your sample of 420 observations on California school districts that this slope is nonzero? Can you reject the taxpayer's hypothesis that $\beta_{ClassSize} = 0$, or should you accept it, at least tentatively pending further new evidence?

This section discusses tests of hypotheses about the slope β_1 or intercept β_0 of the population regression line. We start by discussing two-sided tests of the slope β_1 in detail, then turn to one-sided tests and to tests of hypotheses regarding the intercept β_0.

Two-Sided Hypotheses Concerning β_1

The general approach to testing hypotheses about these coefficients is the same as to testing hypotheses about the population mean, so we begin with a brief review.

Testing hypotheses about the population mean. Recall from Section 3.2 that the null hypothesis that the mean of Y is a specific value $\mu_{Y,0}$ can be written as H_0: $E(Y) = \mu_{Y,0}$, and the two-sided alternative is H_1: $E(Y) \neq \mu_{Y,0}$.

The test of the null hypothesis H_0 against the two-sided alternative proceeds as in the three steps summarized in Key Concept 3.6. The first is to compute the standard error of $\overline{Y}$, $SE(\overline{Y})$, which is an estimator of the standard deviation of the sampling distribution of $\overline{Y}$. The second step is to compute the t-statistic, which has the general form given in Key Concept 5.1; applied here, the t-statistic is $t = (\overline{Y} - \mu_{Y,0})/SE(\overline{Y})$.

The third step is to compute the p-value, which is the smallest significance level at which the null hypothesis could be rejected, based on the test statistic actually observed; equivalently, the p-value is the probability of obtaining a statistic, by random sampling variation, at least as different from the null hypothesis value as is the statistic actually observed, assuming that the null hypothesis is correct

5.1 General Form of the *t*-Statistic

In general, the *t*-statistic has the form

$$t = \frac{\textbf{estimator} - \textbf{hypothesized value}}{\textbf{standard error of the estimator}}. \qquad (5.1)$$

(Key Concept 3.5). Because the t-statistic has a standard normal distribution in large samples under the null hypothesis, the p-value for a two-sided hypothesis test is $2\Phi(-|t^{act}|)$, where t^{act} is the value of the t-statistic actually computed and Φ is the cumulative standard normal distribution tabulated in Appendix Table 1. Alternatively, the third step can be replaced by simply comparing the t-statistic to the critical value appropriate for the test with the desired significance level. For example, a two-sided test with a 5% significance level would reject the null hypothesis if $|t^{act}| > 1.96$. In this case, the population mean is said to be statistically significantly different than the hypothesized value at the 5% significance level.

Testing hypotheses about the slope β_1. At a theoretical level, the critical feature justifying the foregoing testing procedure for the population mean is that, in large samples, the sampling distribution of $\overline{Y}$ is approximately normal. Because $\hat{\beta}_1$ also has a normal sampling distribution in large samples, hypotheses about the true value of the slope β_1 can be tested using the same general approach.

The null and alternative hypotheses need to be stated precisely before they can be tested. The angry taxpayer's hypothesis is that $\beta_{ClassSize} = 0$. More generally, under the null hypothesis the true population slope β_1 takes on some specific value, $\beta_{1,0}$. Under the two-sided alternative, β_1 does not equal $\beta_{1,0}$. That is, the **null hypothesis** and the two-sided **alternative hypothesis** are

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 \neq \beta_{1,0} \quad \text{(two-sided alternative)}. \qquad (5.2)$$

To test the null hypothesis H_0, we follow the same three steps as for the population mean.

The first step is to compute the **standard error of $\hat{\beta}_1$**, $SE(\hat{\beta}_1)$. The standard error of $\hat{\beta}_1$ is an estimator of $\sigma_{\hat{\beta}_1}$, the standard deviation of the sampling distribution of $\hat{\beta}_1$. Specifically,

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}, \qquad (5.3)$$

where

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}. \tag{5.4}$$

The estimator of the variance in Equation (5.4) is discussed in Appendix 5.1. Although the formula for $\hat{\sigma}^2_{\hat{\beta}_1}$ is complicated, in applications the standard error is computed by regression software so that it is easy to use in practice.

The second step is to compute the ***t*-statistic**,

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}. \tag{5.5}$$

The third step is to compute the ***p*-value**, the probability of observing a value of $\hat{\beta}_1$ at least as different from $\beta_{1,0}$ as the estimate actually computed ($\hat{\beta}_1^{act}$), assuming that the null hypothesis is correct. Stated mathematically,

$$\begin{aligned} p\text{-value} &= \Pr_{H_0}[|\hat{\beta}_1 - \beta_{1,0}| > |\hat{\beta}_1^{act} - \beta_{1,0}|] \\ &= \Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right] = \Pr_{H_0}(|t| > |t^{act}|). \end{aligned} \tag{5.6}$$

where $\Pr_{H_0}$ denotes the probability computed under the null hypothesis, the second equality follows by dividing by $SE(\hat{\beta}_1)$, and t^{act} is the value of the t-statistic actually computed. Because $\hat{\beta}_1$ is approximately normally distributed in large samples, under the null hypothesis the t-statistic is approximately distributed as a standard normal random variable, so in large samples,

$$p\text{-value} = \Pr(|Z| > |t^{act}|) = 2\Phi(-|t^{act}|). \tag{5.7}$$

A small value of the p-value, say less than 5%, provides evidence against the null hypothesis in the sense that the chance of obtaining a value of $\hat{\beta}_1$ by pure random variation from one sample to the next is less than 5% if, in fact, the null hypothesis is correct. If so, the null hypothesis is rejected at the 5% significance level.

Alternatively, the hypothesis can be tested at the 5% significance level simply by comparing the value of the t-statistic to ±1.96, the critical value for a two-sided test, and rejecting the null hypothesis at the 5% level if $|t^{act}| > 1.96$.

These steps are summarized in Key Concept 5.2.

KEY CONCEPT 5.2

TESTING THE HYPOTHESIS $\beta_1 = \beta_{1,0}$ AGAINST THE ALTERNATIVE $\beta_1 \neq \beta_{1,0}$

1. Compute the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$ [Equation (5.3)].
2. Compute the t-statistic [Equation (5.5)].
3. Compute the p-value [Equation (5.7)]. Reject the hypothesis at the 5% significance level if the p-value is less than 0.05 or, equivalently, if $|t^{act}| > 1.96$.

The standard error and (typically) the t-statistic and p-value testing $\beta_1 = 0$ are computed automatically by regression software.

Reporting regression equations and application to test scores. The OLS regression of the test score against the student–teacher ratio, reported in Equation (4.11), yielded $\hat{\beta}_0 = 698.9$ and $\hat{\beta}_1 = -2.28$. The standard errors of these estimates are $SE(\hat{\beta}_0) = 10.4$ and $SE(\hat{\beta}_1) = 0.52$.

Because of the importance of the standard errors, by convention they are included when reporting the estimated OLS coefficients. One compact way to report the standard errors is to place them in parentheses below the respective coefficients of the OLS regression line:

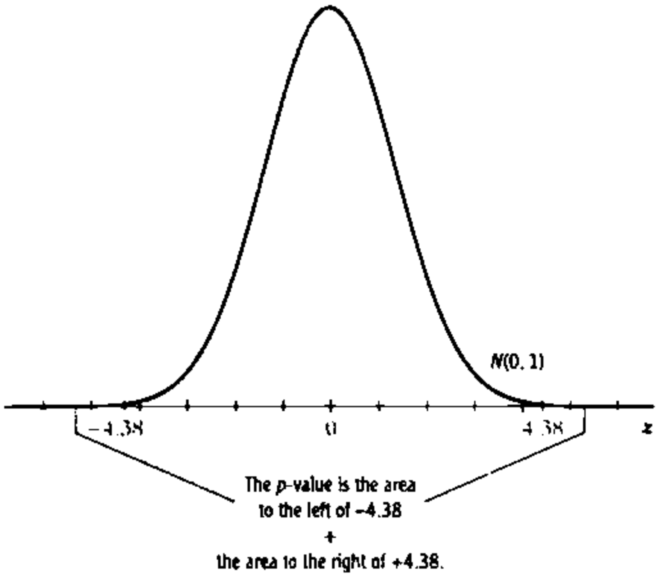
$$\underset{}{\widehat{TestScore}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR,\ R^2 = 0.051, SER = 18.6. \tag{5.8}$$

Equation (5.8) also reports the regression R^2 and the standard error of the regression (SER) following the estimated regression line. Thus Equation (5.8) provides the estimated regression line, estimates of the sampling uncertainty of the slope and the intercept (the standard errors), and two measures of the fit of this regression line (the R^2 and the SER). This is a common format for reporting a single regression equation, and it will be used throughout the rest of this book.

Suppose you wish to test the null hypothesis that the slope β_1 is zero in the population counterpart of Equation (5.8) at the 5% significance level. To do so, construct the t-statistic and compare it to 1.96, the 5% (two-sided) critical value taken from the standard normal distribution. The t-statistic is constructed by substituting the hypothesized value of β_1 under the null hypothesis (zero), the estimated slope, and its standard error from Equation (5.8) into the general formula

FIGURE 5.1 Calculating the p-Value of a Two-Sided Test When $t^{act} = -4.38$

The p-value of a two-sided test is the probability that $|Z| > |t^{act}|$, where Z is a standard normal random variable and t^{act} is the value of the t-statistic calculated from the sample. When $t^{act} = -4.38$, the p-value is only 0.00001.



in Equation (5.5); the result is $t^{act} = (-2.28 - 0)/0.52 = -4.38$. This t-statistic exceeds (in absolute value) the 5% two-sided critical value of 1.96, so the null hypothesis is rejected in favor of the two-sided alternative at the 5% significance level.

Alternatively, we can compute the p-value associated with $t^{act} = -4.38$. This probability is the area in the tails of standard normal distribution, as shown in Figure 5.1. This probability is extremely small, approximately 0.00001, or 0.001%. That is, if the null hypothesis $\beta_{ClassSize} = 0$ is true, the probability of obtaining a value of $\hat{\beta}_1$ as far from the null as the value we actually obtained is extremely small, less than 0.001%. Because this event is so unlikely, it is reasonable to conclude that the null hypothesis is false.

One-Sided Hypotheses Concerning β_1

The discussion so far has focused on testing the hypothesis that $\beta_1 = \beta_{1,0}$ against the hypothesis that $\beta_1 \neq \beta_{1,0}$. This is a two-sided hypothesis test, because under the alternative β_1 could be either larger or smaller than $\beta_{1,0}$. Sometimes, however, it is appropriate to use a one-sided hypothesis test. For example, in the student–teacher ratio/test score problem, many people think that smaller classes provide a better learning environment. Under that hypothesis, β_1 is negative: Smaller classes lead to higher scores. It might make sense, therefore, to test the null hypothesis that $\beta_1 = 0$ (no effect) against the one-sided alternative that $\beta_1 < 0$.

For a one-sided test, the null hypothesis and the one-sided alternative hypothesis are

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_1: \beta_1 < \beta_{1,0}, \quad \text{(one-sided alternative).} \qquad (5.9)$$

where $\beta_{1,0}$ is the value of β_1 under the null (0 in the student–teacher ratio example) and the alternative is that β_1 is less than $\beta_{1,0}$. If the alternative is that β_1 is greater than $\beta_{1,0}$, the inequality in Equation (5.9) is reversed.

Because the null hypothesis is the same for a one- and a two-sided hypothesis test, the construction of the t-statistic is the same. The only difference between a one- and two-sided hypothesis test is how you interpret the t-statistic. For the one-sided alternative in Equation (5.9), the null hypothesis is rejected against the one-sided alternative for large negative, but not large positive, values of the t-statistic: Instead of rejecting if $|t^{act}| > 1.96$, the hypothesis is rejected at the 5% significance level if $t^{act} < -1.645$.

The p-value for a one-sided test is obtained from the cumulative standard normal distribution as

$$p\text{-value} = \Pr(Z < t^{act}) = \Phi(t^{act}) \ (p\text{-value, one-sided left-tail test}). \quad (5.10)$$

If the alternative hypothesis is that β_1 is greater than $\beta_{1,0}$, the inequalities in Equations (5.9) and (5.10) are reversed, so the p-value is the right-tail probability, $\Pr(Z > t^{act})$.

When should a one-sided test be used? In practice, one-sided alternative hypotheses should be used only when there is a clear reason for doing so. This reason could come from economic theory, prior empirical evidence, or both. However, even if it initially seems that the relevant alternative is one-sided, upon reflection this might not necessarily be so. A newly formulated drug undergoing clinical trials actually could prove harmful because of previously unrecognized side effects. In the class size example, we are reminded of the graduation joke that a university's secret of success is to admit talented students and then make sure that the faculty stays out of their way and does as little damage as possible. In practice, such ambiguity often leads econometricians to use two-sided tests.

Application to test scores. The t-statistic testing the hypothesis that there is no effect of class size on test scores [so $\beta_{1,0} = 0$ in Equation (5.9)] is $t^{act} = -4.38$. This is less than -2.33 (the critical value for a one-sided test with a 1% significance level), so the null hypothesis is rejected against the one-sided alternative at the 1% level. In fact, the p-value is less than 0.0006%. Based on these data, you can reject the angry taxpayer's assertion that the negative estimate of the slope arose purely because of random sampling variation at the 1% significance level.

Testing Hypotheses About the Intercept β_0

This discussion has focused on testing hypotheses about the slope, β_1. Occasionally, however, the hypothesis concerns the intercept, β_0. The null hypothesis concerning the intercept and the two-sided alternative are

$$H_0: \beta_0 = \beta_{0,0} \text{ vs. } H_1: \beta_0 \neq \beta_{0,0} \quad \text{(two-sided alternative).} \tag{5.11}$$

The general approach to testing this null hypothesis consists of the three steps in Key Concept 5.2, applied to β_0 (the formula for the standard error of $\hat{\beta}_0$ is given in Appendix 5.1). If the alternative is one-sided, this approach is modified as was discussed in the previous subsection for hypotheses about the slope.

Hypothesis tests are useful if you have a specific null hypothesis in mind (as did our angry taxpayer). Being able to accept or to reject this null hypothesis based on the statistical evidence provides a powerful tool for coping with the uncertainty inherent in using a sample to learn about the population. Yet, there are many times that no single hypothesis about a regression coefficient is dominant, and instead one would like to know a range of values of the coefficient that are consistent with the data. This calls for constructing a confidence interval.

5.2 Confidence Intervals for a Regression Coefficient

Because any statistical estimate of the slope β_1 necessarily has sampling uncertainty, we cannot determine the true value of β_1 exactly from a sample of data. It is, however, possible to use the OLS estimator and its standard error to construct a confidence interval for the slope β_1 or for the intercept β_0.

Confidence interval for β_1. Recall that a 95% **confidence interval for β_1** has two equivalent definitions. First, it is the set of values that cannot be rejected using a two-sided hypothesis test with a 5% significance level. Second, it is an interval that has a 95% probability of containing the true value of β_1: that is, in 95% of possible samples that might be drawn, the confidence interval will contain the true value of β_1. Because this interval contains the true value in 95% of all samples, it is said to have a **confidence level** of 95%.

The reason these two definitions are equivalent is as follows. A hypothesis test with a 5% significance level will, by definition, reject the true value of β_1 in only 5% of all possible samples; that is, in 95% of all possible samples the true value of β_1 will *not* be rejected. Because the 95% confidence interval (as defined in the first definition) is the set of all values of β_1 that are *not* rejected at the 5% significance level, it follows that the true value of β_1 will be contained in the confidence interval in 95% of all possible samples.

As in the case of a confidence interval for the population mean (Section 3.3), in principle a 95% confidence interval can be computed by testing all possible values of β_1 (that is, testing the null hypothesis $\beta_1 = \beta_{1,0}$ for all values of $\beta_{1,0}$) at the 5% significance level using the t-statistic. The 95% confidence interval is then the collection of all the values of β_1 that are not rejected. But constructing the t-statistic for all values of β_1 would take forever.

An easier way to construct the confidence interval is to note that the t-statistic will reject the hypothesized value $\beta_{1,0}$ whenever $\beta_{1,0}$ is outside the range $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$. That is, the 95% confidence interval for β_1 is the interval $[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)]$. This argument parallels the argument used to develop a confidence interval for the population mean.

The construction of a confidence interval for β_1 is summarized as Key Concept 5.3.

Confidence interval for β_0. A 95% confidence interval for β_0 is constructed as in Key Concept 5.3, with $\hat{\beta}_0$ and $SE(\hat{\beta}_0)$ replacing $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$.

Application to test scores. The OLS regression of the test score against the student–teacher ratio, reported in Equation (5.8), yielded $\hat{\beta}_1 = -2.28$ and $SE(\hat{\beta}_1) = 0.52$. The 95% two-sided confidence interval for β_1 is $\{-2.28 \pm 1.96 \times 0.52\}$, or $-3.30 \leq \beta_1 \leq -1.26$. The value $\beta_1 = 0$ is not contained in this confidence interval.

CONFIDENCE INTERVAL FOR β_1

5.3

A 95% two-sided confidence interval for β_1 is an interval that contains the true value of β_1 with a 95% probability; that is, it contains the true value of β_1 in 95% of all possible randomly drawn samples. Equivalently, it is the set of values of β_1 that cannot be rejected by a 5% two-sided hypothesis test. When the sample size is large, it is constructed as

$$\text{95\% confidence interval for } \beta_1 = [\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)]. \qquad (5.12)$$

so (as we knew already from Section 5.1) the hypothesis $\beta_1 = 0$ can be rejected at the 5% significance level.

Confidence intervals for predicted effects of changing X. The 95% confidence interval for β_1 can be used to construct a 95% confidence interval for the predicted effect of a general change in X.

Consider changing X by a given amount, Δx. The predicted change in Y associated with this change in X is $\beta_1\Delta x$. The population slope β_1 is unknown, but because we can construct a confidence interval for β_1, we can construct a confidence interval for the predicted effect $\beta_1\Delta x$. Because one end of a 95% confidence interval for β_1 is $\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)$, the predicted effect of the change Δx using this estimate of β_1 is $[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)] \times \Delta x$. The other end of the confidence interval is $\hat{\beta}_1 + 1.96SE(\hat{\beta}_1)$, and the predicted effect of the change using that estimate is $[\hat{\beta}_1 + 1.96SE(\hat{\beta}_1)] \times \Delta x$. Thus a 95% confidence interval for the effect of changing x by the amount Δx can be expressed as

$$\text{95\% confidence interval for } \beta_1\Delta x = [\hat{\beta}_1\Delta x - 1.96SE(\hat{\beta}_1) \times \Delta x, \hat{\beta}_1\Delta x + 1.96SE(\hat{\beta}_1) \times \Delta x]. \qquad (5.13)$$

For example, our hypothetical superintendent is contemplating reducing the student–teacher ratio by 2. Because the 95% confidence interval for β_1 is $[-3.30, -1.26]$, the effect of reducing the student–teacher ratio by 2 could be as great as $-3.30 \times (-2) = 6.60$, or as little as $-1.26 \times (-2) = 2.52$. Thus decreasing the student–teacher ratio by 2 is predicted to increase test scores by between 2.52 and 6.60 points, with a 95% confidence level.

5.3 Regression When X Is a Binary Variable

The discussion so far has focused on the case that the regressor is a continuous variable. Regression analysis can also be used when the regressor is binary—that is, when it takes on only two values, 0 or 1. For example, X might be a worker's gender (= 1 if female, = 0 if male), whether a school district is urban or rural (= 1 if urban, = 0 if rural), or whether the district's class size is small or large (= 1 if small, = 0 if large). A binary variable is also called an **indicator variable** or sometimes a **dummy variable**.

Interpretation of the Regression Coefficients

The mechanics of regression with a binary regressor are the same as if it is continuous. The interpretation of β_1, however, is different, and it turns out that regression with a binary variable is equivalent to performing a difference of means analysis, as described in Section 3.4.

To see this, suppose you have a variable D_i that equals either 0 or 1, depending on whether the student–teacher ratio is less than 20:

$$D_i = \begin{cases} 1 \text{ if the student–teacher ratio in } i^{th} \text{ district} < 20 \\ 0 \text{ if the student–teacher ratio in } i^{th} \text{ district} \geq 20. \end{cases} \quad (5.14)$$

The population regression model with D_i as the regressor is

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \quad i = 1, \ldots, n. \quad (5.15)$$

This is the same as the regression model with the continuous regressor X_i, except that now the regressor is the binary variable D_i. Because D_i is not continuous, it is not useful to think of β_1 as a slope; indeed, because D_i can take on only two values, there is no "line" so it makes no sense to talk about a slope. Thus we will not refer to β_1 as the slope in Equation (5.15); instead we will simply refer to β_1 as the **coefficient multiplying D_i** in this regression or, more compactly, the **coefficient on D_i**.

If β_1 in Equation (5.15) is not a slope, then what is it? The best way to interpret β_0 and β_1 in a regression with a binary regressor is to consider, one at a time the two possible cases, $D_i = 0$ and $D_i = 1$. If the student–teacher ratio is high, then $D_i = 0$ and Equation (5.15) becomes

$$Y_i = \beta_0 + u_i \quad (D_i = 0). \quad (5.16)$$

Because $E(u_i|D_i) = 0$, the conditional expectation of Y_i when $D_i = 0$ is $E(Y_i|D_i = 0) = \beta_0$; that is, β_0 is the population mean value of test scores when the student–teacher ratio is high. Similarly, when $D_i = 1$,

$$Y_i = \beta_0 + \beta_1 + u_i \quad (D_i = 1). \tag{5.17}$$

Thus, when $D_i = 1$, $E(Y_i|D_i = 1) = \beta_0 + \beta_1$; that is, $\beta_0 + \beta_1$ is the population mean value of test scores when the student–teacher ratio is low.

Because $\beta_0 + \beta_1$ is the population mean of Y_i when $D_i = 1$ and β_0 is the population mean of Y_i when $D_i = 0$, the difference $(\beta_0 + \beta_1) - \beta_0 = \beta_1$ is the difference between these two means. In other words, β_1 is the difference between the conditional expectation of Y_i when $D_i = 1$ and when $D_i = 0$, or $\beta_1 = E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$. In the test score example, β_1 is the difference between mean test score in districts with low student–teacher ratios and the mean test score in districts with high student–teacher ratios.

Because β_1 is the difference in the population means, it makes sense that the OLS estimator β_1 is the difference between the sample averages of Y_i in the two groups, and in fact this is the case.

Hypothesis tests and confidence intervals. If the two population means are the same, then β_1 in Equation (5.15) is zero. Thus, the null hypothesis that the two population means are the same can be tested against the alternative hypothesis that they differ by testing the null hypothesis $\beta_1 = 0$ against the alternative $\beta_1 \neq 0$. This hypothesis can be tested using the procedure outlined in Section 5.1. Specifically, the null hypothesis can be rejected at the 5% level against the two-sided alternative when the OLS t-statistic $t = \hat{\beta}_1/SE(\hat{\beta}_1)$ exceeds 1.96 in absolute value. Similarly, a 95% confidence interval for β_1, constructed as $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$ as described in Section 5.2, provides a 95% confidence interval for the difference between the two population means.

Application to test scores. As an example, a regression of the test score against the student–teacher ratio binary variable D defined in Equation (5.14) estimated by OLS using the 420 observations in Figure 4.2, yields

$$\begin{aligned}\widehat{TestScore} &= \underset{(1.3)}{650.0} + \underset{(1.8)}{7.4}D,\ R^2 = 0.035,\ SER = 18.7,\end{aligned} \tag{5.18}$$

where the standard errors of the OLS estimates of the coefficients β_0 and β_1 are given in parentheses below the OLS estimates. Thus the average test score for the subsample with student–teacher ratios greater than or equal to 20 (that is, for which $D = 0$) is 650.0, and the average test score for the subsample with student–teacher ratios less than 20 (so $D = 1$) is $650.0 + 7.4 = 657.4$. The difference between the sample average test scores for the two groups is 7.4. This is the OLS estimate of β_1, the coefficient on the student–teacher ratio binary variable D.

Is the difference in the population mean test scores in the two groups statistically significantly different from zero at the 5% level? To find out, construct the t-statistic on β_1: $t = 7.4/1.8 = 4.04$. This exceeds 1.96 in absolute value, so the hypothesis that the population mean test scores in districts with high and low student–teacher ratios is the same can be rejected at the 5% significance level.

The OLS estimator and its standard error can be used to construct a 95% confidence interval for the true difference in means. This is $7.4 \pm 1.96 \times 1.8 = (3.9, 10.9)$. This confidence interval excludes $\beta_1 = 0$, so that (as we know from the previous paragraph) the hypothesis $\beta_1 = 0$ can be rejected at the 5% significance level.

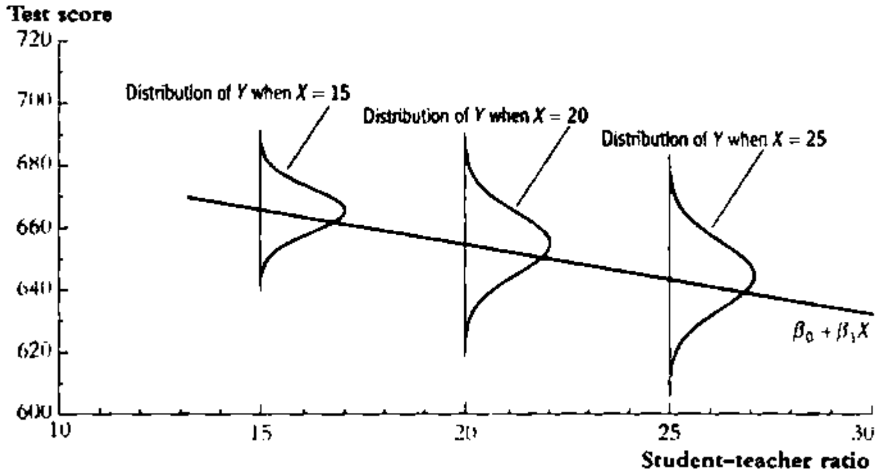
5.4 Heteroskedasticity and Homoskedasticity

Our only assumption about the distribution of u_i conditional on X_i is that it has a mean of zero (the first least squares assumption). If, furthermore, the *variance* of this conditional distribution does not depend on X_i, then the errors are said to be homoskedastic. This section discusses homoskedasticity, its theoretical implications, the simplified formulas for the standard errors of the OLS estimators that arise if the errors are homoskedastic, and the risks you run if you use these simplified formulas in practice.

What Are Heteroskedasticity and Homoskedasticity?

Definitions of heteroskedasticity and homoskedasticity. The error term u_i is **homoskedastic** if the variance of the conditional distribution of u_i given X_i is constant for $i = 1, \ldots, n$ and in particular does not depend on X_i. Otherwise, the error term is **heteroskedastic**.

FIGURE 5.2 An Example of Heteroskedasticity



Like Figure 4.4, this shows the conditional distribution of test scores for three different class sizes. Unlike Figure 4.4, these distributions become more spread out (have a larger variance) for larger class sizes. Because the variance of the distribution of *u* given *X*, var(*u*|*X*), depends on *X*, *u* is heteroskedastic.

As an illustration, return to Figure 4.4. The distribution of the errors u_i is shown for various values of x. Because this distribution applies specifically for the indicated value of x, this is the conditional distribution of u_i given $X_i = x$. As drawn in that figure, all these conditional distributions have the same spread; more precisely, the variance of these distributions is the same for the various values of x. That is, in Figure 4.4, the conditional variance of u_i given $X_i = x$ does not depend on x, so the errors illustrated in Figure 4.4 are homoskedastic.

In contrast, Figure 5.2 illustrates a case in which the conditional distribution of u_i spreads out as x increases. For small values of x, this distribution is tight, but for larger values of x, it has a greater spread. Thus, in Figure 5.2 the variance of u_i given $X_i = x$ increases with x, so that the errors in Figure 5.2 are heteroskedastic.

The definitions of heteroskedasticity and homoskedasticity are summarized in Key Concept 5.4.

KEY CONCEPT 5.4 Heteroskedasticity and Homoskedasticity

The error term u_i is homoskedastic if the variance of the conditional distribution of u_i given X_i, $\mathrm{var}(u_i|X_i = x)$, is constant for $i = 1, \ldots, n$, and in particular does not depend on x. Otherwise, the error term is heteroskedastic.

Example. These terms are a mouthful and the definitions might seem abstract. To help clarify them with an example, we digress from the student–teacher ratio/test score problem and instead return to the example of earnings of male versus female college graduates considered in the box in Chapter 3, "The Gender Gap in Earnings of College Graduates in the United States." Let $MALE_i$ be a binary variable that equals 1 for male college graduates and equals 0 for female graduates. The binary variable regression model relating someone's earnings to his or her gender is

$$Earnings_i = \beta_0 + \beta_1 MALE_i + u_i \tag{5.19}$$

for $i = 1, \ldots, n$. Because the regressor is binary, β_1 is the difference in the population means of the two groups—in this case, the difference in mean earnings between men and women who graduated from college.

The definition of homoskedasticity states that the variance of u_i does not depend on the regressor. Here the regressor is $MALE_i$, so at issue is whether the variance of the error term depends on $MALE_i$. In other words, is the variance of the error term the same for men and for women? If so, the error is homoskedastic; if not, it is heteroskedastic.

Deciding whether the variance of u_i depends on $MALE_i$ requires thinking hard about what the error term actually is. In this regard, it is useful to write Equation (5.19) as two separate equations, one for men and one for women:

$$Earnings_i = \beta_0 + u_i \quad \text{(women) and} \tag{5.20}$$

$$Earnings_i = \beta_0 + \beta_1 + u_i \quad \text{(men).} \tag{5.21}$$

Thus, for women, u_i is the deviation of the i^{th} woman's earnings from the population mean earnings for women (β_0), and for men, u_i is the deviation of the i^{th} man's earnings from the population mean earnings for men ($\beta_0 + \beta_1$). It follows that the statement, "the variance of u_i does not depend on $MALE$," is equivalent to the statement, "the variance of earnings is the same for men as it is for women." In other words, in this example, the error term is homoskedastic if the variance of the population distribution of earnings is the same for men and women; if these variances differ, the error term is heteroskedastic.

Mathematical Implications of Homoskedasticity

The OLS estimators remain unbiased and asymptotically normal. Because the least squares assumptions in Key Concept 4.3 place no restrictions on the conditional variance, they apply to both the general case of heteroskedasticity and the special case of homoskedasticity. Therefore, the OLS estimators remain unbiased and consistent even if the errors are homoskedastic. In addition, the OLS estimators have sampling distributions that are normal in large samples even if the errors are homoskedastic. Whether the errors are homoskedastic or heteroskedastic, the OLS estimator is unbiased, consistent, and asymptotically normal.

Efficiency of the OLS estimator when the errors are homoskedastic. If the least squares assumptions in Key Concept 4.3 hold and the errors are homoskedastic, then the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are efficient among all estimators that are linear in $Y_1, \ldots, Y_n$ and are unbiased, conditional on $X_1, \ldots, X_n$. This result, which is called the Gauss-Markov theorem, is discussed in Section 5.5.

Homoskedasticity-only variance formula. If the error term is homoskedastic, then the formulas for the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ in Key Concept 4.4 simplify. Consequently, if the errors are homoskedastic, then there is a specialized formula that can be used for the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. The **homoskedasticity-only standard error** of $\hat{\beta}_1$, derived in Appendix 5.1, is $SE(\hat{\beta}_1) = \sqrt{\tilde{\sigma}^2_{\hat{\beta}_1}}$, where $\tilde{\sigma}^2_{\hat{\beta}_1}$ is the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$:

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s_{\hat{u}}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{(homoskedasticity-only)}, \qquad (5.22)$$

where $s_{\hat{u}}^2$ is given in Equation (4.19). The homoskedasticity-only formula for the standard error of $\hat{\beta}_0$ is given in Appendix 5.1. In the special case that X is a binary variable, the estimator of the variance of $\hat{\beta}_1$ under homoskedasticity (that is, the square of the standard error of $\hat{\beta}_1$ under homoskedasticity) is the so-called pooled variance formula for the difference in means, given in Equation (3.23).

Because these alternative formulas are derived for the special case that the errors are homoskedastic and do not apply if the errors are heteroskedastic, they will be referred to as the "homoskedasticity-only" formulas for the variance and standard error of the OLS estimators. As the name suggests, if the errors are heteroskedastic, then the homoskedasticity-only standard errors are inappropriate. Specifically, if the errors are heteroskedastic, then the t-statistic computed using the homoskedasticity-only standard error does not have a standard normal distribution, even in large samples. In fact, the correct critical values to use for this homoskedasticity-only t-statistic depend on the precise nature of the heteroskedasticity, so those critical values cannot be tabulated. Similarly, if the errors are heteroskedastic but a confidence interval is constructed as ±1.96 homoskedasticity-only standard errors, in general the probability that this interval contains the true value of the coefficient is not 95%, even in large samples.

In contrast, because homoskedasticity is a special case of heteroskedasticity, the estimators $\hat{\sigma}^2_{\hat{\beta}_1}$ and $\hat{\sigma}^2_{\hat{\beta}_0}$ of the variances of $\hat{\beta}_1$ and $\hat{\beta}_0$ given in Equations (5.4) and (5.26) produce valid statistical inferences whether the errors are heteroskedastic or homoskedastic. Thus hypothesis tests and confidence intervals based on those standard errors are valid whether or not the errors are heteroskedastic. Because the standard errors we have used so far [i.e., those based on Equations (5.4) and (5.26)] lead to statistical inferences that are valid whether or not the errors are heteroskedastic, they are called **heteroskedasticity-robust standard errors**. Because such formulas were proposed by Eicker (1967), Huber (1967), and White (1980), they are also referred to as Eicker-Huber-White standard errors.

What Does This Mean in Practice?

Which is more realistic, heteroskedasticity or homoskedasticity? The answer to this question depends on the application. However, the issues can be clarified by returning to the example of the gender gap in earnings among college graduates. Familiarity with how people are paid in the world around us gives some clues as to which assumption is more sensible. For many years—and, to a lesser extent, today—women were not found in the top-paying jobs: There have always been poorly paid men, but there have rarely been highly paid women. This suggests that the distribution of earnings among women is tighter than among men (See the box in Chapter 3, "The Gender Gap in Earnings of College Graduates in the United States"). In other words, the variance of the error term in Equa-

The Economic Value of a Year of Education: Homoskedasticity or Heteroskedasticity?

On average, workers with more education have higher earnings than workers with less education. But if the best-paying jobs mainly go to the college educated, it might also be that the *spread* of the distribution of earnings is greater for workers with more education. Does the distribution of earnings spread out as education increases?

This is an empirical question, so answering it requires analyzing data. Figure 5.3 is a scatterplot of the hourly earnings and the number of years of education for a sample of 2950 full-time workers in the United States in 2004, ages 29 and 30, with between 6 and 18 years of education. The data come from the March 2005 Current Population Survey, which is described in Appendix 3.1.

Figure 5.3 has two striking features. The first is that the mean of the distribution of earnings increases with the number of years of education. This increase is summarized by the OLS regression line,

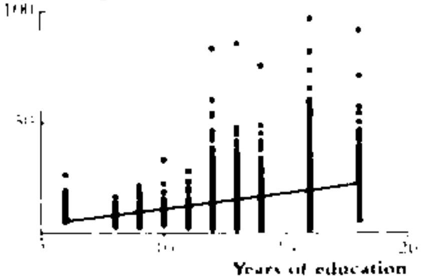
$$\begin{aligned}\widehat{Earnings} &= -3.13 + 1.47 Years\ Education,\\ &\quad\ (0.93)\quad (0.07)\\ R^2 &= 0.130,\ SER = 8.77.\end{aligned} \qquad (5.23)$$

This line is plotted in Figure 5.3. The coefficient of 1.47 in the OLS regression line means that, on average, hourly earnings increase by $1.47 for each additional year of education. The 95% confidence interval for this coefficient is 1.47 ± 1.96 × 0.07, or 1.33 to 1.61.

The second striking feature of Figure 5.3 is that the spread of the distribution of earnings increases with the years of education. While some workers with many years of education have low-paying jobs, very few workers with low levels of education have high-paying jobs. This can be stated more precisely by looking at the spread of the residuals around the OLS regression line. For workers with ten years of education, the standard deviation of the residuals is $5.46; for workers with a high school diploma, this standard deviation is $7.43; and for workers with a college degree, this standard deviation increases to $10.78. Because these standard deviations differ for different levels of education, the variance of the residuals in the regression of Equation (5.23) depends on the value of the regressor (the years of education): in other words, the regression errors are heteroskedastic. In real-world terms, not all college graduates will be earning $50/hour by the time they are 29, but some will, and workers with only ten years of education have no shot at those jobs.

FIGURE 5.3 **Scatterplot of Hourly Earnings and Years of Education for 29- to 30-Year Olds in the United States in 2004**

Hourly earnings are plotted against years of education for 2950 full time, 29- to 30-year-old workers. The spread around the regression line increases with the years of education, indicating that the regression errors are heteroskedastic.



tion (5.20) for women is plausibly less than the variance of the error term in Equation (5.21) for men. Thus, the presence of a "glass ceiling" for women's jobs and pay suggests that the error term in the binary variable regression model in Equation (5.19) is heteroskedastic. Unless there are compelling reasons to the contrary—and we can think of none—it makes sense to treat the error term in this example as heteroskedastic.

As this example of modeling earnings illustrates, heteroskedasticity arises in many econometric applications. At a general level, economic theory rarely gives any reason to believe that the errors are homoskedastic. It therefore is prudent to assume that the errors might be heteroskedastic unless you have compelling reasons to believe otherwise.

Practical implications. The main issue of practical relevance in this discussion is whether one should use heteroskedasticity-robust or homoskedasticity-only standard errors. In this regard, it is useful to imagine computing both, then choosing between them. If the homoskedasticity-only and heteroskedasticity-robust standard errors are the same, nothing is lost by using the heteroskedasticity-robust standard errors; if they differ, however, then you should use the more reliable ones that allow for heteroskedasticity. The simplest thing, then, is always to use the heteroskedasticity-robust standard errors.

For historical reasons, many software programs use the homoskedasticity-only standard errors as their default setting, so it is up to the user to specify the option of heteroskedasticity-robust standard errors. The details of how to implement heteroskedasticity-robust standard errors depend on the software package you use.

All of the empirical examples in this book employ heteroskedasticity-robust standard errors unless explicitly stated otherwise.[1]

*5.5 The Theoretical Foundations of Ordinary Least Squares

As discussed in Section 4.5, the OLS estimator is unbiased, is consistent, has a variance that is inversely proportional to n, and has a normal sampling distribution

[1] In case this book is used in conjunction with other texts, it might be helpful to note that some textbooks add homoskedasticity to the list of least squares assumptions. As just discussed, however, this additional assumption is not needed for the validity of OLS regression analysis as long as heteroskedasticity-robust standard errors are used.

*This section is optional and is not used in later chapters.

when the sample size is large. In addition, under certain conditions the OLS estimator is more efficient than some other candidate estimators. Specifically, if the least squares assumptions hold and if the errors are homoskedastic, then the OLS estimator has the smallest variance of all conditionally unbiased estimators that are linear functions of $Y_1, \dots, Y_n$. This section explains and discusses this result, which is a consequence of the Gauss-Markov theorem. The section concludes with a discussion of alternative estimators that are more efficient than OLS when the conditions of the Gauss-Markov theorem do not hold.

Linear Conditionally Unbiased Estimators and the Gauss-Markov Theorem

If the three least squares assumptions (Key Concept 4.3) hold and if the error is homoskedastic, then the OLS estimator has the smallest variance, conditional on $X_1, \dots, X_n$, among all estimators in the class of linear conditionally unbiased estimators. In other words, the OLS estimator is the **Best Linear conditionally Unbiased Estimator**—that is, it is BLUE. This result extends to regression the result, summarized in Key Concept 3.3, that the sample average $\overline{Y}$ is the most efficient estimator of the population mean among the class of all estimators that are unbiased and are linear functions (weighted averages) of $Y_1, \dots, Y_n$.

Linear conditionally unbiased estimators. The class of linear conditionally unbiased estimators consists of all estimators of β_1 that are linear functions of $Y_1, \dots, Y_n$ and that are unbiased, conditional on $X_1, \dots, X_n$. That is, if $\tilde{\beta}_1$ is a linear estimator, then it can be written as

$$\tilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i \quad (\tilde{\beta}_1 \text{ is linear}), \tag{5.24}$$

where the weights $a_1, \dots, a_n$ can depend on $X_1, \dots, X_n$ but *not* on $Y_1, \dots, Y_n$. The estimator $\tilde{\beta}_1$ is conditionally unbiased if the mean of its conditional sampling distribution, given $X_1, \dots, X_n$, is β_1. That is, the estimator $\tilde{\beta}_1$ is conditionally unbiased if

$$E(\tilde{\beta}_1 | X_1, \dots, X_n) = \beta_1 \quad (\tilde{\beta}_1 \text{ is conditionally unbiased}). \tag{5.25}$$

The estimator $\tilde{\beta}_1$ is a linear conditionally unbiased estimator if it can be written in the form of Equation (5.24) (it is linear) and if Equation (5.25) holds (it is

KEY CONCEPT 5.5

THE GAUSS-MARKOV THEOREM FOR $\hat{\beta}_1$

If the three least squares assumptions in Key Concept 4.3 hold *and* if errors are homoskedastic, then the OLS estimator $\hat{\beta}_1$ is the **B**est (most efficient) **L**inear conditionally **U**nbiased **E**stimator (is **BLUE**).

conditionally unbiased). It is shown in Appendix 5.2 that the OLS estimator is linear and conditionally unbiased.

The Gauss-Markov theorem. The Gauss-Markov theorem states that, under a set of conditions known as the Gauss-Markov conditions, the OLS estimator $\hat{\beta}_1$ has the smallest conditional variance, given $X_1, \ldots, X_n$, of all linear conditionally unbiased estimators of β_1; that is, the OLS estimator is BLUE. The Gauss-Markov conditions, which are stated in Appendix 5.2, are implied by the three least squares assumptions plus the assumption that the errors are homoskedastic. Consequently, if the three least squares assumptions hold and the errors are homoskedastic, then OLS is BLUE. The Gauss-Markov theorem is stated in Key Concept 5.5 and proven in Appendix 5.2.

Limitations of the Gauss-Markov theorem. The Gauss-Markov theorem provides a theoretical justification for using OLS. However, the theorem has two important limitations. First, its conditions might not hold in practice. In particular, if the error term is heteroskedastic—as it often is in economic applications—then the OLS estimator is no longer BLUE. As discussed in Section 5.4, the presence of heteroskedasticity does not pose a threat to inference based on heteroskedasticity-robust standard errors, but it does mean that OLS is no longer the efficient linear conditionally unbiased estimator. An alternative to OLS when there is heteroskedasticity of a known form, called the weighted least squares estimator, is discussed below.

The second limitation of the Gauss-Markov theorem is that even if the conditions of the theorem hold, there are other candidate estimators that are not linear and conditionally unbiased; under some conditions, these other estimators are more efficient than OLS.

Regression Estimators Other Than OLS

Under certain conditions, some regression estimators are more efficient than OLS.

The weighted least squares estimator. If the errors are heteroskedastic, then OLS is no longer BLUE. If the nature of the heteroskedastic is known—specifically, if the conditional variance of u_i given X_i is known up to a constant factor of proportionality—then it is possible to construct an estimator that has a smaller variance than the OLS estimator. This method, called **weighted least squares** (WLS), weights the i^{th} observation by the inverse of the square root of the conditional variance of u_i given X_i. Because of this weighting, the errors in this weighted regression are homoskedastic, so OLS, when applied to the weighted data, is BLUE. Although theoretically elegant, the practical problem with weighted least squares is that you must know how the conditional variance of u_i depends on X_i —something that is rarely known in applications.

The least absolute deviations estimator. As discussed in Section 4.3, the OLS estimator can be sensitive to outliers. If extreme outliers are not rare, then other estimators can be more efficient than OLS and can produce inferences that are more reliable. One such estimator is the least absolute deviations (LAD) estimator, in which the regression coefficients β_0 and β_1 are obtained by solving a minimization like that in Equation (4.6), except that the absolute value of the prediction "mistake" is used instead of its square. That is, the least absolute deviations estimators of β_0 and β_1 are the values of b_0 and b_1 that minimize $\sum_{i=1}^{n}|Y_i - b_0 - b_1X_i|$. In practice, this estimator is less sensitive to large outliers in u than is OLS.

In many economic data sets, severe outliers in u are rare, so use of the LAD estimator, or other estimators with reduced sensitivity to outliers, is uncommon in applications. Thus the treatment of linear regression throughout the remainder of this text focuses exclusively on least squares methods.

*5.6 Using the *t*-Statistic in Regression When the Sample Size Is Small

When the sample size is small, the exact distribution of the t-statistic is complicated and depends on the unknown population distribution of the data. If, however, the three least squares assumptions hold, the regression errors are homoskedastic, *and* the regression errors are normally distributed, then the OLS

*This section is optional and is not used in later chapters.

estimator is normally distributed and the homoskedasticity-only t-statistic has a Student t distribution. These five assumptions—the three least squares assumptions, that the errors are homoskedastic, and that the errors are normally distributed—are collectively called the **homoskedastic normal regression assumptions**.

The *t*-Statistic and the Student *t* Distribution

Recall from Section 2.4 that the Student t distribution with m degrees of freedom is defined to be the distribution of $Z/\sqrt{W/m}$, where Z is a random variable with a standard normal distribution, W is a random variable with a chi-squared distribution with m degrees of freedom, and Z and W are independent. Under the null hypothesis, the t-statistic computed using the homoskedasticity-only standard error can be written in this form.

The homoskedasticity-only t-statistic testing $\beta_1 = \beta_{1,0}$ is $\tilde{t} = (\hat{\beta}_1 - \beta_{1,0})/\tilde{\sigma}_{\hat{\beta}_1}$, where $\tilde{\sigma}^2_{\hat{\beta}_1}$ is defined in Equation (5.22). Under the homoskedastic normal regression assumptions, Y has a normal distribution, conditional on $X_1, \ldots, X_n$. As discussed in Section 5.5, the OLS estimator is a weighted average of $Y_1, \ldots, Y_n$, where the weights depend on $X_1, \ldots, X_n$ [see Equation (5.32) in Appendix 5.2]. Because a weighted average of independent normal random variables is normally distributed, $\hat{\beta}_1$ has a normal distribution, conditional on $X_1, \ldots, X_n$. Thus $(\hat{\beta}_1 - \beta_{1,0})$ has a normal distribution under the null hypothesis, conditional on $X_1, \ldots, X_n$. In addition, the (normalized) homoskedasticity-only variance estimator has a chi-squared distribution with $n - 2$ degrees of freedom, divided by $n - 2$, and $\tilde{\sigma}^2_{\hat{\beta}_1}$ and $\hat{\beta}_1$ are independently distributed. Consequently, the homoskedasticity-only t-statistic has a Student t distribution with $n - 2$ degrees of freedom.

This result is closely related to a result discussed in Section 3.5 in the context of testing for the equality of the means in two samples. In that problem, if the two population distributions are normal with the same variance and if the t-statistic is constructed using the pooled standard error formula [Equation (3.23)], then the (pooled) t-statistic has a Student t distribution. When X is binary, the homoskedasticity-only standard error for $\hat{\beta}_1$ simplifies to the pooled standard error formula for the difference of means. It follows that the result of Section 3.5 is a special case of the result that, if the homoskedastic normal regression assumptions hold, then the homoskedasticity-only regression t-statistic has a Student t distribution (see Exercise 5.10).

Use of the Student *t* Distribution in Practice

If the regression errors are homoskedastic and normally distributed and if the homoskedasticity-only t-statistic is used, then critical values should be taken from the Student *t* distribution (Appendix Table 2) instead of the standard normal distribution. Because the difference between the Student *t* distribution and the normal distribution is negligible if *n* is moderate or large, this distinction is relevant only if the sample size is small.

In econometric applications, there is rarely a reason to believe that the errors are homoskedastic and normally distributed. Because sample sizes typically are large, however, inference can proceed as described in Sections 5.1 and 5.2—that is, by first computing heteroskedasticity-robust standard errors, and then using the standard normal distribution to compute *p*-values, hypothesis tests, and confidence intervals.

5.7 Conclusion

Return for a moment to the problem that started Chapter 4: the superintendent who is considering hiring additional teachers to cut the student–teacher ratio. What have we learned that she might find useful?

Our regression analysis, based on the 420 observations for 1998 in the California test score data set, showed that there was a negative relationship between the student–teacher ratio and test scores: Districts with smaller classes have higher test scores. The coefficient is moderately large, in a practical sense: Districts with 2 fewer students per teacher have, on average, test scores that are 4.6 points higher. This corresponds to moving a district at the 50th percentile of the distribution of test scores to approximately the 60th percentile.

The coefficient on the student–teacher ratio is statistically significantly different from 0 at the 5% significance level. The population coefficient might be 0, and we might simply have estimated our negative coefficient by random sampling variation. However, the probability of doing so (and of obtaining a *t*-statistic on β_1 as large as we did) purely by random variation over potential samples is exceedingly small, approximately 0.001%. A 95% confidence interval for β_1 is $-3.30 \le \beta_1 \le -1.26$.

This represents considerable progress toward answering the superintendent's question. Yet, a nagging concern remains. There is a negative relationship between the student–teacher ratio and test scores, but is this relationship necessarily the *causal* one that the superintendent needs to make her decision? Districts with lower student–teacher ratios have, on average, higher test scores. But does this mean that reducing the student–teacher ratio will, in fact, increase scores?

There is, in fact, reason to worry that it might not. Hiring more teachers, after all, costs money, so wealthier school districts can better afford smaller classes. But students at wealthier schools also have other advantages over their poorer neighbors, including better facilities, newer books, and better-paid teachers. Moreover, students at wealthier schools tend themselves to come from more affluent families, and thus have other advantages not directly associated with their school. For example, California has a large immigrant community; these immigrants tend to be poorer than the overall population and, in many cases, their children are not native English speakers. It thus might be that our negative estimated relationship between test scores and the student–teacher ratio is a consequence of large classes being found in conjunction with many other factors that are, in fact, the real cause of the lower test scores.

These other factors, or "omitted variables," could mean that the OLS analysis done so far has little value to the superintendent. Indeed, it could be misleading: Changing the student–teacher ratio alone would not change these other factors that determine a child's performance at school. To address this problem, we need a method that will allow us to isolate the effect on test scores of changing the student–teacher ratio, ***holding these other factors constant.*** That method is multiple regression analysis, the topic of Chapter 7 and 8.

Summary

1. Hypothesis testing for regression coefficients is analogous to hypothesis testing for the population mean: Use the t-statistic to calculate the p-values and either accept or reject the null hypothesis. Like a confidence interval for the population mean, a 95% confidence interval for a regression coefficient is computed as the estimator ± 1.96 standard errors.
2. When X is binary, the regression model can be used to estimate and test hypotheses about the difference between the population means of the "$X = 0$" group and the "$X = 1$" group.

3. In general the error u_i is heteroskedastic—that is, the variance of u_i at a given value of X_i, $\text{var}(u_i|X_i = x)$ depends on x. A special case is when the error is homoskedastic, that is, $\text{var}(u_i|X_i = x)$ is constant. Homoskedasticity-only standard errors do not produce valid statistical inferences when the errors are heteroskedastic, but heteroskedasticity-robust standard errors do.
4. If the three least squares assumption hold *and* if the regression errors are homoskedastic, then, as a result of the Gauss-Markov theorem, the OLS estimator is BLUE.
5. If the three least squares assumptions hold, if the regression errors are homoskedastic, *and* if the regression errors are normally distributed, then the OLS t-statistic computed using homoskedasticity-only standard errors has a Student t distribution when the null hypothesis is true. The difference between the Student t distribution and the normal distribution is negligible if the sample size is moderate or large.

Key Terms

null hypothesis (150)
two-sided alternative hypothesis (150)
standard error of $\hat{\beta}_1$ (151)
t-statistic (151)
p-value (151)
confidence interval for β_1 (156)
confidence level (156)
indicator variable (158)
dummy variable (158)
coefficient multiplying variable D_i (158)
coefficient on D_i (158)
heteroskedasticity and homoskedasticity (160)
homoskedasticity-only standard errors (163)
heteroskedasticity-robust standard error (164)
best linear unbiased estimator (BLUE) (168)
Gauss-Markov theorem (168)
weighted least squares (169)
homoskedastic normal regression assumptions (170)
Gauss-Markov conditions (182)

Review the Concepts

5.1 Outline the procedures for computing the p-value of a two-sided test of $H_0: \mu_Y = 0$ using an i.i.d. set of observations $Y_i, i = 1, \ldots, n$. Outline the procedures for computing the p-value of a two-sided test of $H_0: \beta_1 = 0$ in a regression model using an i.i.d. set of observations $(Y_i, X_i), i = 1, \ldots, n$.

5.2 Explain how you could use a regression model to estimate the wage gender gap using the data on earnings of men and women. What are the dependent and independent variables?

5.3 Define ***homoskedasticity*** and ***heteroskedasticity***. Provide a hypothetical empirical example in which you think the errors would be heteroskedastic and explain you reasoning.

Exercises

5.1 Suppose that a researcher, using data on class size (CS) and average test scores from 100 third-grade classes, estimates the OLS regression.

$$\underset{}{\widehat{TestScore}} = \underset{(20.4)}{520.4} - \underset{(2.21)}{5.82} \times CS,\ R^2 = 0.08,\ SER = 11.5.$$

a. Construct a 95% confidence interval for β_1, the regression slope coefficient.

b. Calculate the p-value for the two-sided test of the null hypothesis $H_0: \beta_1 = 0$. Do you reject the null hypothesis at the 5% level? At the 1% level?

c. Calculate the p-value for the two-sided test of the null hypothesis $H_0: \beta_1 = -5.6$. Without doing any additional calculations, determine whether -5.6 is contained in the 95% confidence interval for β_1.

d. Construct a 99% confidence interval for β_0.

5.2 Suppose that a researcher, using wage data on 250 randomly selected male workers and 280 female workers, estimates the OLS regression.

$$\widehat{Wage} = \underset{(.23)}{12.52} + \underset{(0.36)}{2.12} \times Male,\ R^2 = 0.06,\ SER = 4.2.$$

where *Wage* is measured in \$/hour and *Male* is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. Define the wage gender gap as the difference in mean earnings between men and women.

a. What is the estimated gender gap?

b. Is the estimated gender gap significantly different from zero? (Compute the p-value for testing the null hypothesis that there is no gender gap.)

c. Construct a 95% confidence interval for the gender gap.

d. In the sample, what is the mean wage of women? Of men?

e. Another researcher uses these same data, but regresses *Wages* on *Female*, a variable that is equal to 1 if the person is female and 0 if the person a male. What are the regression estimates calculated from this regression?

$$\widehat{Wage} = \_\_\_\_\_\_ + \_\_\_\_\_\_ \times Female,\ R^2 = \_\_\_\_\_\_,\ SER = \_\_\_\_\_\_.$$

5.3 Suppose that a random sample of 200 twenty-year-old men is selected from a population and their heights and weights are recorded. A regression of weight on height yields

$$\underset{}{\widehat{Weight}} = \underset{(2.15)}{-99.41} + \underset{(0.31)}{3.94} \times Height,\ R^2 = 0.81,\ SER = 10.2.$$

where *Weight* is measured in pounds and *Height* is measured in inches. A man has a late growth spurt and grows 1.5 inches over the course of a year. Construct a 99% confidence interval for the person's weight gain.

5.4 Read the box "The Economic Value of a Year of Education: Heteroskedasticity or Homoskedasticity?" in Section 5.4. Use the regression reported in Equation (5.23) to answer the following.

a. A randomly selected 30-year-old worker reports an education level of 16 years. What is the worker's expected average hourly earnings?

b. A high school graduate (12 years of education) is contemplating going to a community college for a two-year degree. How much is this worker's average hourly earnings expected to increase?

c. A high school counselor tells a student that, on average, college graduates earn \$10 per hour more than high school graduates. Is this statement consistent with the regression evidence? What range of values is consistent with the regression evidence?

5.5 In the 1980s, Tennessee conducted an experiment in which kindergarten students were randomly assigned to "regular" and "small" classes, and given standardized tests at the end of the year. (Regular classes contained approximately 24 students and small classes contained approximately 15 students.) Suppose that, in the population, the standardized tests have a mean score of 925 points and a standard deviation of 75 points. Let *SmallClass* denote a binary variable equal to 1 if the student is assigned to a small class and equal to 0 otherwise. A regression of *Testscore* on *SmallClass* yields

$$\widehat{TestScore} = \underset{(1.6)}{918.0} + \underset{(2.5)}{13.9} \times SmallClass,\ R^2 = 0.01,\ SER = 74.6.$$

a. Do small classes improve test scores? By how much? Is the effect large? Explain.

b. Is the estimated effect of class size on test scores statistically significant? Carry out a test at the 5% level.

c. Construct a 99% confidence interval for the effect of *SmallClass* on test score.

5.6 Refer to the regression described in Exercise 5.5.

a. Do you think that the regression errors plausibly are homoskedastic? Explain.

b. $SE(\hat{\beta}_1)$ was computed using Equation (5.3). Suppose that the regression errors were homoskedastic: Would this affect the validity of the confidence interval constructed in Exercise 5.5(c)? Explain.

5.7 Suppose that (Y_i, X_i) satisfy the assumptions in Key Concept 4.3. A random sample of size $n = 250$ is drawn and yields

$$\hat{Y} = \underset{(3.1)}{5.4} + \underset{(1.5)}{3.2}X,\ R^2 = 0.26,\ SER = 6.2.$$

a. Test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$ at the 5% level.

b. Construct a 95% confidence interval for β_1.

c. Suppose you learned that Y_i and X_i were independent. Would you be surprised? Explain.

d. Suppose that Y_i and X_i are independent and many samples of size $n = 250$ are drawn, regressions estimated, and (a) and (b) answered. In what fraction of the samples would H_0 from (a) be rejected? In what fraction of samples would the value $\beta_1 = 0$ be included in the confidence interval from (b)?

5.8 Suppose that (Y_i, X_i) satisfy the assumptions in Key Concept 4.3 and, in addition, u_i is $N(0, \sigma_u^2)$ and is independent of X_i. A sample of size $n = 30$ yields

$$\underset{}{\hat{Y}} = \underset{(10.2)}{43.2} + \underset{(7.4)}{61.5}X,\ R^2 = 0.54,\ SER = 1.52,$$

where the numbers in parentheses are the homoskedastic-only standard errors for the regression coefficients.

a. Construct a 95% confidence interval for β_0.

b. Test $H_0: \beta_1 = 55$ vs. $H_1: \beta_1 \neq 55$ at the 5% level.

c. Test $H_0: \beta_1 = 55$ vs. $H_1: \beta_1 > 55$ at the 5% level.

5.9 Consider the regression model

$$Y_i = \beta X_i + u_i,$$

where u_i and X_i satisfy the assumptions in Key Concept 4.3. Let $\overline{\beta}$ denote an estimator of β that is constructed as $\overline{\beta} = \frac{\overline{Y}}{\overline{X}}$, where $\overline{Y}$ and $\overline{X}$ are the sample means of Y_i and X_i, respectively.

a. Show that $\overline{\beta}$ is a linear function of $Y_1, Y_2, \ldots, Y_n$.

b. Show that $\overline{\beta}$ is conditionally unbiased.

5.10 Let X_i denote a binary variable and consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$. Let $\overline{Y}_0$ denote the sample mean for observations with $X = 0$ and $\overline{Y}_1$

denote the sample mean for observations with $X = 1$. Show that $\hat{\beta}_0 = \overline{Y}_0$, $\hat{\beta}_0 + \hat{\beta}_1 = \overline{Y}_1$, and $\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$.

5.11 A random sample of workers contains $n_m = 120$ men and $n_w = 131$ women. The sample average of men's weekly earnings ($\overline{Y}_m = \frac{1}{n_m}\sum_{i=1}^{n_m} Y_{m,i}$) is \$523.1 and the sample standard deviation ($s_m = \sqrt{\frac{1}{n_m - 1}\sum_{i=1}^{n_m}(Y_{m,i} - \overline{Y}_m)^2}$) is \$68.1. The corresponding values for women are $\overline{Y}_w = \$485.10$ and $s_w = \$51.10$. Let *Women* denote an indicator variable that is equal to 1 for women and 0 for men, and suppose that all 251 observations are used in the regression $Y_i = \beta_0 + \beta_1 \, Women_i + u_i$ is run. Find the OLS estimates of β_0 and β_1 and their corresponding standard errors.

5.12 Starting from Equation (4.22), derive the variance of $\hat{\beta}_0$ under homoskedasticity given in Equation (5.28) in Appendix 5.1.

5.13 Suppose that (Y_i, X_i) satisfy the assumptions in Key Concept 4.3 and, in addition, u_i is $N(0, \sigma_u^2)$ and is independent of X_i.

a. Is $\hat{\beta}_1$ conditionally unbiased?

b. Is $\hat{\beta}_1$ the best linear conditionally unbiased estimator of β_1?

c. How would your answers to (a) and (b) change if you assumed only that (Y_i, X_i) satisfied the assumptions in Key Concept 4.3 and $\text{var}(u_i | X_i = x)$ is constant?

d. How would your answers to (a) and (b) change if you assumed only that (Y_i, X_i) satisfied the assumptions in Key Concept 4.3?

5.14 Suppose that $Y_i = \beta X_i + u_i$, where (u_i, X_i) satisfy the Gauss-Markov conditions given in Equation (5.31).

a. Derive the least squares estimator of β and show that it is a linear function of $Y_1, \ldots, Y_n$.

b. Show that the estimator is conditionally unbiased.

c. Derive the conditional variance of the estimator.

d. Prove that the estimator is BLUE.

5.15 A researcher has two independent samples of observations on (Y_i, X_i). To be specific, suppose that Y_i denotes earnings, X_i denotes years of schooling, and the independent samples are for men and women. Write the regression for men as $Y_{m,i} = \beta_{m,0} + \beta_{m,1}X_{m,i} + u_{m,i}$, and the regression for women as $Y_{w,i} = \beta_{w,0} + \beta_{w,1}X_{w,i} + u_{w,i}$. Let $\hat{\beta}_{m,1}$ denote the OLS estimator constructed using the sample of men, $\hat{\beta}_{w,1}$ denote the OLS estimator constructed from the sample of women, and $SE(\hat{\beta}_{m,1})$ and $SE(\hat{\beta}_{w,1})$ denote the corresponding standard errors. Show that the standard error of $\hat{\beta}_{m,1} - \hat{\beta}_{w,1}$ is given by $SE(\hat{\beta}_{m,1} - \hat{\beta}_{w,1}) = \sqrt{[SE(\hat{\beta}_{m,1})]^2 + [SE(\hat{\beta}_{w,1})]^2}$.

Empirical Exercises

E5.1 Using the data set **CPS04** described in Empirical Exercise 4.1, run a regression of average hourly earnings (*AHE*) on *Age* and carry out the following exercises.

- **a.** Is the estimated regression slope coefficient statistically significant? That is, can you reject the null hypothesis $H_0: \beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?
- **b.** Construct a 95% confidence interval for the slope coefficient.
- **c.** Repeat (a) using only the data for high school graduates.
- **d.** Repeat (a) using only the data for college graduates.
- **e.** Is the effect of age on earnings different for high school graduates than for college graduates? Explain. (*Hint:* See Exercise 5.15.)

E5.2 Using the data set **TeachingRatings** described in Empirical Exercise 4.2, run a regression of *Course_Eval* on *Beauty*. Is the estimated regression slope coefficient statistically significant? That is, can you reject the null hypothesis $H_0: \beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?

E5.3 Using the data set **CollegeDistance** described in Empirical Exercise 4.3, run a regression of years of completed education (ED) on distance to the nearest college (*Dist*) and carry out the following exercises.

- **a.** Is the estimated regression slope coefficient statistically significant? That is, can you reject the null hypothesis $H_0: \beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?
- **b.** Construct a 95% confidence interval for the slope coefficient.
- **c.** Run the regression using data only on females and repeat (b).

d. Run the regression using data only on males and repeat (b).

e. Is the effect of distance on completed years of education different for men than for women? (*Hint:* See Exercise 5.15.)

APPENDIX 5.1 Formulas for OLS Standard Errors

This appendix discusses the formulas for OLS standard errors. These are first presented under the least squares assumptions in Key Concept 4.3, which allow for heteroskedasticity; these are the "heteroskedasticity-robust" standard errors. Formulas for the variance of the OLS estimators and the associated standard errors are then given for the special case of homoskedasticity.

Heteroskedasticity-Robust Standard Errors

The estimator $\hat{\sigma}^2_{\hat{\beta}_1}$ defined in Equation (5.4) is obtained by replacing the population variances in Equation (4.21) by the corresponding sample variances, with a modification. The variance in the numerator of Equation (4.21) is estimated by $\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2$, where the divisor $n - 2$ (instead of n) incorporates a degrees-of-freedom adjustment to correct for downward bias, analogously to the degrees-of-freedom adjustment used in the definition of the *SER* in Section 4.3. The variance in the denominator is estimated by $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Replacing $\text{var}[(X_i - \mu_X)u_i]$ and $\text{var}(X_i)$ in Equation (4.21) by these two estimators yields $\hat{\sigma}^2_{\hat{\beta}_1}$ in Equation (5.4). The consistency of heteroskedasticity-robust standard errors is discussed in Section 17.3.

The estimator of the variance of $\hat{\beta}_0$ is

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{H}_i^2\hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{H}_i^2\right)^2}, \qquad (5.26)$$

where $\hat{H}_i = 1 - [\bar{X} / \frac{1}{n}\sum_{i=1}^{n} X_i^2] X_i$. The standard error of $\hat{\beta}_0$ is $SE(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_0}}$. The reasoning behind the estimator $\hat{\sigma}^2_{\hat{\beta}_0}$ is the same as behind $\hat{\sigma}^2_{\hat{\beta}_1}$ and stems from replacing population expectations with sample averages.

Homoskedasticity-Only Variances

Under homoskedasticity, the conditional variance of u_i given X_i is a constant: $\mathrm{var}(u_i|X_i) = \sigma_u^2$. If the errors are homoskedastic, the formulas in Key Concept 4.4 simplify to

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma_u^2}{n\sigma_X^2} \text{ and} \tag{5.27}$$

$$\sigma^2_{\hat{\beta}_0} = \frac{E(X_i^2)}{n\sigma_X^2}\sigma_u^2. \tag{5.28}$$

To derive Equation (5.27), write the numerator in Equation (4.21) as $\mathrm{var}[(X_i - \mu_X)u_i] = E(\{(X_i - \mu_X)u_i - E[(X_i - \mu_X)u_i]\}^2) = E\{[(X_i - \mu_X)u_i]^2\} = E[(X_i - \mu_X)^2u_i^2] = E[(X_i - \mu_X)^2\mathrm{var}(u_i|X_i)]$, where the second equality follows because $E[(X_i - \mu_X)u_i] = 0$ (by the first least squares assumption) and where the final equality follows from the law of iterated expectations (Section 2.3). If u_i is homoskedastic, then $\mathrm{var}(u_i|X_i) = \sigma_u^2$ so $E[(X_i - \mu_X)^2\mathrm{var}(u_i|X_i)] = \sigma_u^2 E[(X_i - \mu_X)^2] = \sigma_u^2\sigma_X^2$. The result in Equation (5.27) follows by substituting this expression into the numerator of Equation (4.21) and simplifying. A similar calculation yields Equation (5.28).

Homoskedasticity-Only Standard Errors

The homoskedasticity-only standard errors are obtained by substituting sample means and variances for the population means and variances in Equations (5.27) and (5.28), and by estimating the variance of u_i by the square of the *SER*. The homoskedasticity-only estimators of these variances are

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \frac{s_{\hat{u}}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{(homoskedasticity-only) and} \tag{5.29}$$

$$\tilde{\sigma}^2_{\hat{\beta}_0} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)s_{\hat{u}}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{(homoskedasticity-only),} \tag{5.30}$$

where s_u^2 is given in Equation (4.19). The homoskedasticity-only standard errors are the square roots of $\tilde{\sigma}^2_{\hat{\beta}_0}$ and $\tilde{\sigma}^2_{\hat{\beta}_1}$.

APPENDIX 5.2 The Gauss-Markov Conditions and a Proof of the Gauss-Markov Theorem

As discussed in Section 5.5, the Gauss-Markov theorem states that if the Gauss-Markov conditions hold, then the OLS estimator is the best (most efficient) conditionally linear unbiased estimator (is BLUE). This appendix begins by stating the Gauss-Markov conditions and showing that they are implied by the three least squares condition plus homoskedasticity. We next show that the OLS estimator is a linear conditionally unbiased estimator. Finally, we turn to the proof of the theorem.

The Gauss-Markov Conditions

The three **Gauss-Markov conditions** are

$$
\begin{aligned}
&\text{(i)}\ E(u_i \mid X_1, \ldots, X_n) = 0 \\
&\text{(ii)}\ \mathrm{var}(u_i \mid X_1, \ldots, X_n) = \sigma_u^2,\ 0 < \sigma_u^2 < \infty \\
&\text{(iii)}\ E(u_i u_j \mid X_1, \ldots, X_n) = 0,\ i \neq j
\end{aligned}
\tag{5.31}
$$

where the conditions hold for $i, j = 1, \ldots, n$. The three conditions, respectively, state that u_i has mean zero, that u_i has a constant variance, and that the errors are uncorrelated for different observations, where all these statements hold conditionally on all observed X's $(X_1, \ldots, X_n)$.

The Gauss-Markov conditions are implied by the three least squares assumptions (Key Concept 4.3), plus the additional assumptions that the errors are homoskedastic. Because the observations are i.i.d. (Assumption 2), $E(u_i \mid X_1, \ldots, X_n) = E(u_i \mid X_i)$, and by Assumption 1, $E(u_i \mid X_i) = 0$; thus condition (i) holds. Similarly, by Assumption 2, $\mathrm{var}(u_i \mid X_1, \ldots, X_n) = \mathrm{var}(u_i \mid X_i)$, and because the errors are assumed to be homoskedastic, $\mathrm{var}(u_i \mid X_i) = \sigma_u^2$, which is constant. Assumption 3 (nonzero finite fourth moments) ensures that $0 < \sigma_u^2 < \infty$, so condition (ii) holds. To show that condition (iii) is implied by the least squares assumptions, note that $E(u_i u_j \mid X_1, \ldots, X_n) = E(u_i u_j \mid X_i, X_j)$ because (X_i, Y_i) are i.i.d. by Assumption 2. Assumption 2 also implies that $E(u_i u_j \mid X_i, X_j) = E(u_i \mid X_i)\, E(u_j \mid X_j)$ for $i \neq j$; because $E(u_i \mid X_i) = 0$ for all i, it follows that $E(u_i u_j \mid X_1, \ldots, X_n) = 0$ for all $i \neq j$, so condition (iii) holds. Thus, the least squares assumptions in Key Concept 4.3, plus homoskedasticity of the errors, imply the Gauss-Markov conditions in Equation (5.31).

The OLS Estimator $\hat{\beta}_1$ Is a Linear Conditionally Unbiased Estimator

To show that $\hat{\beta}_1$ is linear, first note that, because $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$ (by the definition of $\overline{X}$), $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_i - \overline{X})Y_i - \overline{Y}\sum_{i=1}^{n}(X_i - X) = \sum_{i=1}^{n}(X_i - \overline{X})Y_i$. Substituting this result into the formula for $\hat{\beta}_1$ in Equation (4.7) yields

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sum_{i=1}^{n}\hat{a}_iY_i, \text{ where } \hat{a}_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{5.32}$$

Because the weights $\hat{a}_i$, $i = 1, \ldots, n$ in Equation (5.32) depend on $X_1, \ldots, X_n$ but not on $Y_1, \ldots, Y_n$, the OLS estimator $\hat{\beta}_1$ is a linear estimator.

Under the Gauss-Markov conditions, $\hat{\beta}_1$ is conditionally unbiased, and the variance of the conditional distribution of $\hat{\beta}_1$, given $X_1, \ldots, X_n$, is

$$\text{var}(\hat{\beta}_1 | X_1, \ldots, X_n) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - X)^2}. \tag{5.33}$$

The result that $\hat{\beta}_1$ is conditionally unbiased was previously shown in Appendix 4.3.

Proof of the Gauss-Markov Theorem

We start by deriving some facts that hold for all linear conditionallly unbiased estimators—that is, for all estimators $\tilde{\beta}_1$ satisfying Equations (5.24) and (5.25). Substituting $Y_i = \beta_0 + \beta_1X_i + u_i$ into $\tilde{\beta}_1 = \sum_{i=1}^{n}a_iY_i$ and collecting terms, we have that

$$\tilde{\beta}_1 = \beta_0\left(\sum_{i=1}^{n}a_i\right) + \beta_1\left(\sum_{i=1}^{n}a_iX_i\right) + \sum_{i=1}^{n}a_iu_i. \tag{5.34}$$

By the first Gauss-Markov condition, $E(\sum_{i=1}^{n}a_iu_i | X_1, \ldots, X_n) = \sum_{i=1}^{n}a_iE(u_i | X_1, \ldots, X_n) = 0$; thus, taking conditional expectations of both sides of Equation (5.34) yields $E(\tilde{\beta}_1 | X_1, \ldots, X_n) = \beta_0(\sum_{i=1}^{n}a_i) + \beta_1(\sum_{i=1}^{n}a_iX_i)$. Because $\tilde{\beta}_1$ is conditionally unbiased by assumption, it must be that $\beta_0(\sum_{i=1}^{n}a_i) + \beta_1(\sum_{i=1}^{n}u_iX_i) = \beta_1$, but for this equality to hold for all values of β_0 and β_1 it must be the case that, for $\tilde{\beta}_1$ to be conditionally unbiased,

$$\sum_{i=1}^{n}a_i = 0 \text{ and } \sum_{i=1}^{n}a_iX_i = 1. \tag{5.35}$$

Under the Gauss-Markov conditions, the variance of $\tilde{\beta}_1$, conditional on $X_1, \ldots, X_n$, has a simple form. Substituting Equation (5.35) into Equation (5.34) yields $\tilde{\beta}_1 - \beta_1 = \sum_{i=1}^{n} a_i u_i$. Thus $\text{var}(\tilde{\beta}_1 | X_1, \ldots, X_n) = \text{var}(\sum_{i=1}^{n} a_i u_i | X_1, \ldots, X_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{cov}(u_i, u_j | X_1, \ldots, X_n)$ applying the second and third Gauss-Markov conditions, the cross terms in the double summation vanish and the expression for the conditional variance simplifies to

$$\text{var}(\tilde{\beta}_1 | X_1, \ldots, X_n) = \sigma_u^2 \sum_{i=1}^{n} a_i^2. \tag{5.36}$$

Note that Equations (5.35) and (5.36) apply to $\hat{\beta}_1$ with weights $a_i = \hat{a}_i$, given in Equation (5.32).

We now show that the two restrictions in Equation (5.35) and the expression for the conditional variance in Equation (5.36) imply that the conditional variance of $\tilde{\beta}_1$ exceeds the conditional variance of $\hat{\beta}_1$ unless $\tilde{\beta}_1 = \hat{\beta}_1$. Let $a_i = \hat{a}_i + d_i$ so $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} (\hat{a}_i + d_i)^2 = \sum_{i=1}^{n} \hat{a}_i^2 + 2\sum_{i=1}^{n} \hat{a}_i d_i + \sum_{i=1}^{n} d_i^2$.

Using the definition of $\hat{a}_i$, we have that

$$\begin{aligned}\sum_{i=1}^{n} \hat{a}_i d_i &= \sum_{i=1}^{n} (X_i - \overline{X}) d_i \Big/ \sum_{j=1}^{n} (X_j - \overline{X})^2 = \left(\sum_{i=1}^{n} d_i X_i - \overline{X} \sum_{i=1}^{n} d_i\right) \Big/ \sum_{j=1}^{n} (X_j - \overline{X})^2 \\ &= \left[\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} \hat{a}_i X_i\right) - \overline{X}\left(\sum_{i=1}^{n} a_i - \sum_{i=1}^{n} \hat{a}_i\right)\right] \Big/ \sum_{j=1}^{n} (X_j - \overline{X})^2 = 0,\end{aligned}$$

where the final equality follows from Equation (5.35) (which holds for both a_i and $\hat{a}_i$). Thus $\sigma_u^2 \sum_{i=1}^{n} a_i^2 = \sigma_u^2 \sum_{i=1}^{n} \hat{a}_i + \sigma_u^2 \sum_{i=1}^{n} d_i^2 = \text{var}(\hat{\beta}_1 | X_1, \ldots, X_n) + \sigma_u^2 \sum_{i=1}^{n} d_i^2$; substituting this result into Equation (5.36) yields

$$\text{var}(\tilde{\beta}_1 | X_1, \ldots, X_n) - \text{var}(\hat{\beta}_1 | X_1, \ldots, X_n) = \sigma_u^2 \sum_{i=1}^{n} d_i^2. \tag{5.37}$$

Thus $\tilde{\beta}_1$ has a greater conditional variance than $\hat{\beta}_1$ if d_i is nonzero for any $i = 1, \ldots, n$. But if $d_i = 0$ for all i, then $a_i = \hat{a}_i$ and $\tilde{\beta}_1 = \hat{\beta}_1$, which proves that OLS is BLUE.

The Gauss-Markov Theorem When X Is Nonrandom

With a minor change in interpretation, the Gauss-Markov theorem also applies to nonrandom regressors; that is, it applies to regressors that do not change their values over repeated samples. Specifically, if the second least squares assumption is replaced by the assumption that $X_1, \ldots, X_n$ are nonrandom (fixed over repeated samples) and $u_1, \ldots, u_n$ are i.i.d., then the foregoing statement and proof of the Gauss-Markov theorem apply directly, except that all of the "conditional on $X_1, \dots, X_n$" statements are unnecessary because $X_1, \dots, X_n$ take on the same values from one sample to the next.

The Sample Average is the Efficient Linear Estimator of $E(Y)$

An implication of the Gauss-Markov theorem is that the sample average, $\overline{Y}$, is the most efficient linear estimator of $E(Y_i)$ when $Y_1, \dots, Y_n$ are i.i.d. To see this, consider the case of regression without an "X," so that the only regressor is the constant regressor $X_{0i} = 1$. Then the OLS estimator $\hat{\beta}_0 = \overline{Y}$. It follows that, under the Gauss-Markov assumptions, $\overline{Y}$ is BLUE. Note that the Gauss-Markov requirement that the error be homoskedastic is irrelevant in this case because there is no regressor, so it follows that $\overline{Y}$ is BLUE if $Y_1, \dots, Y_n$ are i.i.d. This result was stated previously in Key Concept 3.3.