# Chapter 21

# *Maximum Likelihood Estimation*

*"If it walks like a duck, and quacks like a duck, then it is reasonable to guess it's . . ."*

—UNKNOWN

Linear estimators, such as the least squares and instrumental variables estimators, are only one weapon in the econometrician's arsenal. When the relationships of interest to us are not linear in their parameters, attractive linear estimators are difficult, or even impossible, to come by. This chapter expands our horizons by adding a nonlinear estimation strategy to the linear estimation strategy of earlier chapters. This chapter offers *maximum likelihood estimation* as a strategy for obtaining asymptotically efficient estimators and concludes by examining hypothesis testing from a large-sample perspective.

There are two pitfalls to avoid as we read this chapter. The chapter shows that if the Gauss–Markov Assumptions hold and the disturbances are normally distributed, then ordinary least squares (OLS) is also the maximum likelihood estimator. If the new method just returns us to the estimators we've already settled upon, why bother with them, we might think. However, in many settings more complex than the data-generating processes (DGPs) we have analyzed thus far, best linear unbiased estimators (BLUE) don't exist, and econometricians need alternative estimation tools.

The second pitfall is to think that maximum likelihood estimators are always going to have the fine small sample properties that OLS has under the Gauss–Markov Assumptions. Although maximum likelihood estimators are valuable tools when our data-generating processes grow more complex and finite sample tools become unmanageable, the value of maximum likelihood is sometimes limited to large samples, because their small sample properties are sometimes quite unattractive.

# 21.1 The Maximum Likelihood Estimator

**How Do We Create an Estimator?**

In earlier chapters, we used mathematics and statistics to determine BLUE estimators for various data-generating processes. Is there also a formal method for finding asymptotically efficient estimators? Yes, there is. The method is called maximum likelihood estimation (MLE). MLE asks, "What values of the unknown parameters make the data we see least surprising?" In practice, we always obtain one specific sample. This sample had a probability of appearing that depends on the true values of the parameters in the DGP. For some values of those parameters, this sample might almost never appear; for others, it might be altogether unsurprising. Because we did get this sample, it seems plausible to guess that the parameters are of the latter variety, rather than of the former variety. Maximum likelihood estimation formalizes this intuitive approach to estimation, using calculus to determine the parameters that make the sample in hand as unsurprising as possible. Under quite general conditions, maximum likelihood estimators are consistent and asymptotically efficient.

## An Intuitive Example of Maximum Likelihood Estimation

A six-sided die is the most common die, but some games use dice with more sides than six. For example, there are 20-sided dice readily available in game stores. Suppose we have one 6-sided die and one 20-sided die, and we flip a coin to choose one of the two of the dice. The chosen die is rolled and comes up with a four. Which die do you guess we rolled? Many students guess that the 6-sided die was rolled if a four comes up. Their reason is that a four is more likely to come up on a 6-sided die (the probability is one in six) than on a 20-sided die (the probability is one in 20). This reasoning follows the basic principle of maximum likelihood estimation—make the guess for which the observed data are least surprising.

## The Informational Requirements of Maximum Likelihood

Students guessing which die came up relied on knowing that one die had more faces than the other; they relied on knowledge about the probabilities of various outcomes. The Gauss–Markov Assumptions do not provide such detailed probability information about the outcomes in regression models. Maximum likelihood estimation generally requires more detailed information about the DGP than does BLUE estimation. For maximum likelihood estimation, we need to know not only the means, variances, and covariances of the observations (as we needed for deriving the BLUE estimator in earlier chapters), but we also need to know their specific probabilistic distribution. For example, if we begin with the Gauss–Markov Assumptions, we do not yet have sufficient information to find maxi-

mum likelihood estimates. To determine the maximum likelihood estimator, we need also to assume the specific statistical distribution of the disturbances. The next example adds the assumption of normally distributed disturbances to our usual Gauss–Markov Assumptions. Thus the DGP we study is

$$Y_i = \beta X_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$
$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j$$
$$X\text{'s fixed across samples.}$$

## The Likelihood of a Sample

How do we describe the likelihood of observing any one specific sample of data, given our DGP? Suppose we have three observations, $Y_1$, $Y_2$, and $Y_3$, corresponding to the three fixed $X$-values, $X_1$, $X_2$, and $X_3$ (recall that the $X_i$-values are fixed across samples). Define the joint likelihood of $Y_1$, $Y_2$, and $Y_3$ occurring to be $f(Y_1,Y_2,Y_3)$. Probability theory (as described in the Statistical Appendix) tells us that an alternative expression for this joint probability, $f(Y_1,Y_2,Y_3)$, is

$$f_1(Y_1)f_2(Y_2|Y_1)f_3(Y_3|Y_1,Y_2),$$

where $f_i(Y_i)$ is the simple probability density function for $Y_i$ (here $i = 1$); $f_2(Y_2|Y_1)$ is the conditional probability density function for $Y_2$, given $Y_1$; and $f_3(Y_3|Y_1,Y_2)$ is the conditional probability density function for $Y_3$, given $Y_1$ and $Y_2$. However, uncorrelated, normally distributed disturbances are statistically independent, so the joint probability of $Y_1$, $Y_2$, and $Y_3$, $f(Y_1,Y_2,Y_3)$, simplifies to

$$f_1(Y_1)f_2(Y_2)f_3(Y_3).$$

Because we assume in this example that the observations are normally distributed with mean $\beta X_i$, and variance $\sigma^2$,

$$f_i(Y_i) = \left(e^{-(Y_i-\beta X_i)^2/2\sigma^2}\right)/\left(\sqrt{2\sigma^2\pi}\right)$$

so that

$$f(Y_1,Y_2,Y_3) = \left(e^{-(Y_1-\beta X_1)^2/2\sigma^2}\right)\left(e^{-(Y_2-\beta X_2)^2/2\sigma^2}\right)\left(e^{-(Y_3-\beta X_3)^2/2\sigma^2}\right)/\left(\sqrt{2\sigma^2\pi}\right)^3$$

With an explicit expression for the likelihood of the sample data expressed in terms of the observed $X$'s and the parameters of the DGP, we can ask what parameter values would make a given sample least surprising. Mathematically, the problem becomes to choose the estimate of $\beta$ that minimize a particular mathematical function, as we shall now see.

## The Maximum Likelihood Estimator

To emphasize that the joint probability of $Y_1$, $Y_2$, and $Y_3$ depends on the value of $\beta$, we write $f(Y_1,Y_2,Y_3)$ as $f(Y_1,Y_2,Y_3;\beta)$. Maximum likelihood estimation of $\beta$ estimates $\beta$ to be the value of $\beta$ that maximizes $f(Y_1,Y_2,Y_3;\beta)$. Notice the subtle, but important, shift in focus that occurs here. The function $f(Y_1,Y_2,Y_3;\beta)$ is a function of $Y_1$, $Y_2$, and $Y_3$ in which $\beta$ is a parameter. The maximum likelihood strategy inverts these roles, treating $Y_1$, $Y_2$, and $Y_3$ as parameters and $\beta$ as the argument of the function. In effect, we say

$$\left(e^{-(Y_1-\beta X_1)^2/2\sigma^2}\right)\left(e^{-(Y_2-\beta X_2)^2/2\sigma^2}\right)\left(e^{-(Y_3-\beta X_3)^2/2\sigma^2}\right)/\left(\sqrt{2\sigma^2\pi}\right)^3 = g(\beta;Y_1,Y_2,Y_3)$$

and maximize $g(\beta;Y_1,Y_2,Y_3)$ with respect to $\beta$. We compute the value of $\beta$ that makes the particular sample in hand least surprising and call that value the **maximum likelihood estimate**.

As a practical matter, econometricians usually maximize not $g(\beta;Y_1,Y_2,Y_3)$, but the natural logarithm of $g(\beta;Y_1,Y_2,Y_3)$. Working with the log of the likelihood function, $g(\beta;Y_1,Y_2,Y_3)$, which we call $L(\beta)$, doesn't alter the solution, but it does simplify the calculus because it enables us to take the derivative of a sum rather than the derivative of a product when solving for the maximum.

Figure 21.1, Panel A, pictures the log of the likelihood function for a particular sample of data. On the horizontal axis are the values of possible guesses, $\tilde{\beta}$. On the vertical axis is the log of the likelihood of observing the sample in hand if the guess we make is the true parameter, $z = L(\tilde{\beta})$. The particular sample in question is most likely to arise when $\beta = \beta^{mle}$; $\beta^{mle}$ is the maximum likelihood estimate of $\beta$ given this sample. In Panel A, $\beta^{mle} = 15.7$; the true value of $\beta$ in this DGP is 12.

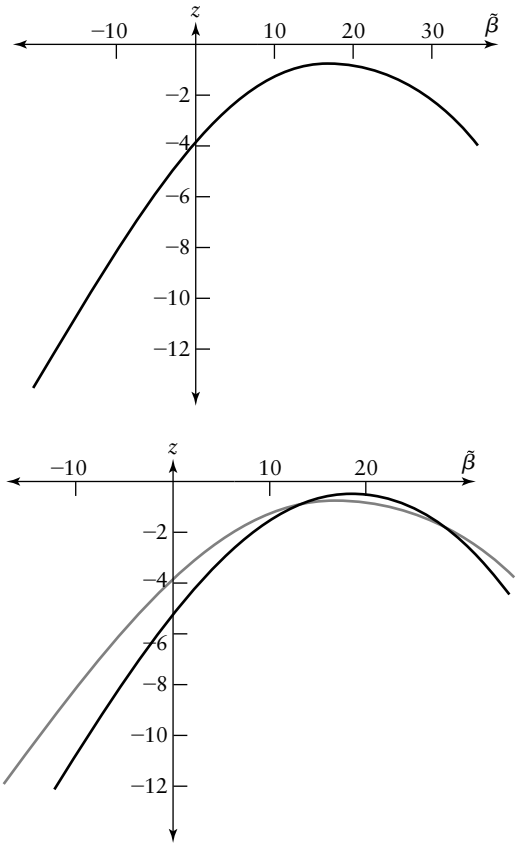
Values of $\tilde{\beta}$ close to $\beta^{mle}$ are almost as likely to give rise to this particular sample as $\beta^{mle}$ is, but values further from $\beta^{mle}$ are increasingly less likely to give rise to this sample of data because the likelihood function declines as we move away from $\beta^{mle}$.

Panel B of Figure 21.1 shows the log of the likelihood function for the same sample as in Panel A, superimposed over the log of the likelihood function for a different sample from the same DGP. The maximum likelihood estimate given the second sample is 12.1, which maximizes the likelihood function for that second observed sample.

Next, we use calculus to obtain the maximum likelihood estimator for $\beta$. We maximize

$$\begin{aligned} L(\beta) &= \ln[g(\beta;Y_1,Y_2,Y_3)] \\ &= \sum_{i=1}^{3}\ln\left(e^{-(Y_i-\beta X_i)^2/2\sigma^2}\right) - \ln\left(\sqrt{2\sigma^2\pi}\right)^3 \\ &= \sum_{i=1}^{3}\ln\left(-(Y_i - \beta X_i)^2/2\sigma^2\right) - \ln\left(\sqrt{2\sigma^2\pi}\right)^3 \end{aligned} \qquad \textbf{21.1}$$

**Figure 21.1**
Log of the Likelihood Function




with respect to $\beta$. Notice that any logarithmic transformation would have turned the products into sums, but it is the natural logarithm that made the ($e$) terms disappear, because $\ln(e) = 1$.

To obtain the maximum likelihood estimator for $\beta$, $\beta^{mle}$, we take the derivative of $L$ with respect to $\beta$, $dL/d\beta$, and set it to zero:

$$dL/d\beta = \sum_{i=1}^{3}(-2X_i(Y_i - \beta^{mle}X_i)/2\sigma^2) = 0$$

so

$$\sum_{i=1}^{3}(-X_i(Y_i - \beta^{mle}X_i)/\sigma^2) = 0$$

thus

$$\sum_{i=1}^{3}(X_iY_i - \beta^{mle}X_i^2) = 0$$

**Table 21.1 The Cobb–Douglas Production Function**

Dependent Variable: LOGOUTCP
Method: Least Squares
Date: 06/18/02 Time: 22:04
Sample: 1899 1922
Included observations: 24

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| **C** | **0.014545** | **0.019979** | **0.727985** | **0.4743** |
| **LOGLABCP** | **0.745866** | **0.041224** | **18.09318** | **0.0000** |

| | | | |
|---|---|---|---|
| R-squared | 0.937028 | Mean dependent var | −0.279148 |
| Adjusted R-squared | 0.934166 | S.D. dependent var | 0.222407 |
| S.E. of regression | 0.057066 | Akaike info criterion | −2.809574 |
| Sum squared resid | 0.071643 | Schwarz criterion | −2.711403 |
| Log likelihood | 35.71489 | F-statistic | 327.3633 |
| Durbin–Watson stat | 1.616390 | Prob(F-statistic) | 0.000000 |

which yields

$$\sum_{i=1}^{3} X_i Y_i - \sum_{i=1}^{3} \beta^{mle} X_i^2 = 0$$

so that

$$\beta^{mle} = \sum_{i=1}^{3} X_i Y_i / \sum_{i=1}^{3} X_i^2.$$

Voila! We discover that $\beta_{g4}/(= \beta^{mle})$ is not only the BLUE estimator for $\beta$, it is also the maximum likelihood estimator if the disturbances are normally distributed! Conveniently, the maximum likelihood estimate of $\beta$ does not depend on the variance of the disturbances, $\sigma^2$, even though the likelihood function does depend on $\sigma^2$. We are often, but not always, so fortunate.

Table 21.1 shows the regression output from Table 4.3. Notice that this table reports the log of the likelihood function for a regression in a row titled "log likelihood." The reported regression included a constant term.

Normally distributed disturbances are not the only ones for which $\beta_{g4}$ is the maximum likelihood estimator of the slope of a line through the origin. It is also the maximum likelihood for a broad statistical family called the exponential distribution. Because maximum likelihood estimation generally yields consistent and

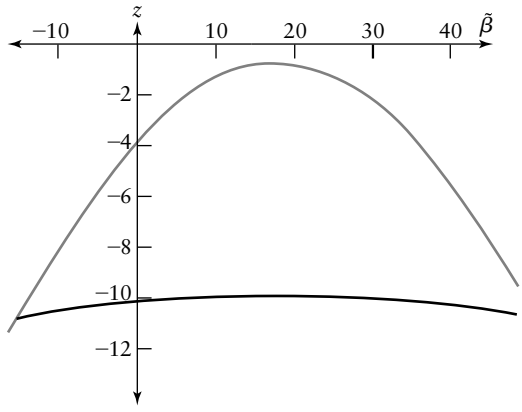



**Figure 21.2**
The Log of the Likelihood Function for One Sample from Two DGPs

asymptotically efficient estimators, $\beta_{g4}$ is asymptotically efficient among all estimators if the disturbances in the DGP follow an exponential distribution. There are, however, distributions for the disturbances for which $\beta_{g4}$ is *not* the maximum likelihood estimator of $\beta$.[1]

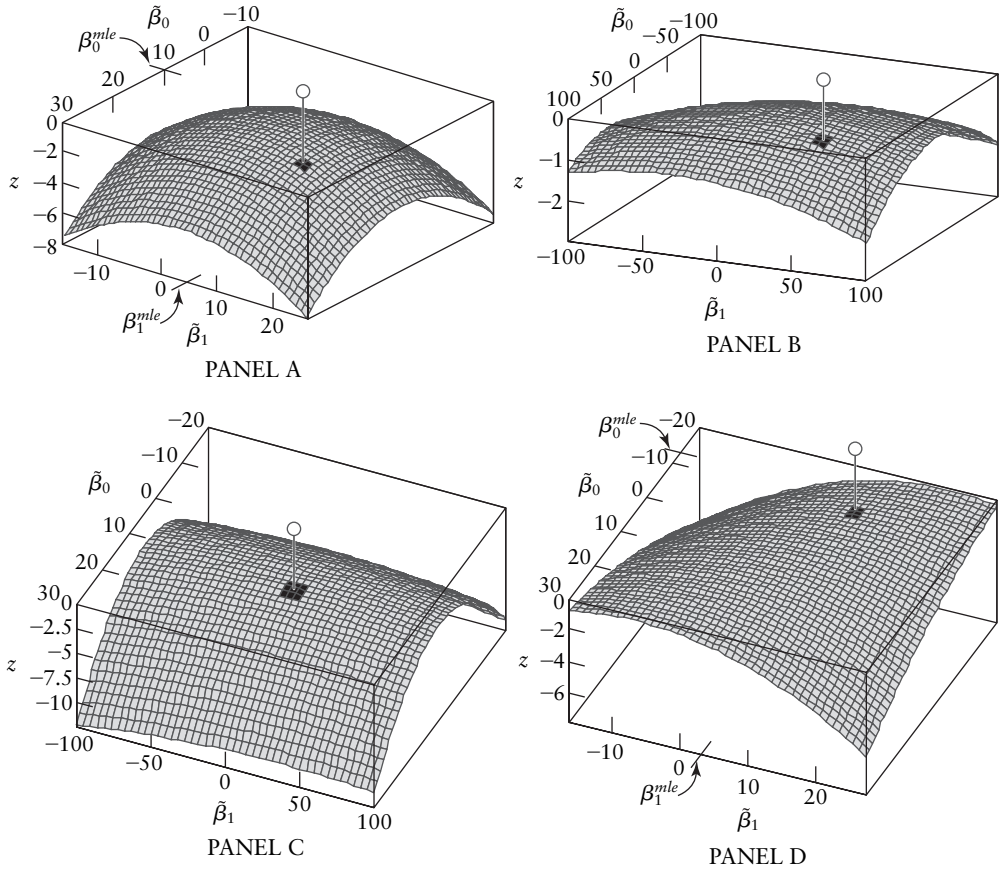
Note carefully the precise meaning of maximum likelihood. $\beta^{mle}$ is *not* necessarily the most likely value of $\beta$. Rather $\beta^{mle}$ is the value of $\beta$ for which this sample that we obtain is least surprising; that is, for any other value of $\beta$, this particular sample would be less likely to arise than it was if $\beta = \beta^{mle}$.

## Efficiency and the Shape of the Likelihood Function

The shape of the log of the likelihood function reflects the efficiency with which we can estimate $\beta$. Figure 21.2 shows two likelihood functions for two DGPs in which $\beta = 12$. What differentiates the two DGPs is the variance of the disturbances. The flatter function in Figure 21.2 reflects a larger variance of the disturbances in that sample's DGP; the other function, taken from Figure 21.1A, corresponds to a smaller variance of the disturbances. *The flatter the likelihood function, the less precisely we estimate* $\beta$. Indeed, the maximum likelihood estimae of $\beta$ using the sample from the DGP with the larger variance is 26.9, much farther from 12 than the maximum likelihood estimate from the other DGPs sample, 15.7.

When several parameters are being estimated, the connection between efficiency and the shape of the likelihood function becomes more complicated. Figure 21.3 shows four likelihood functions, each for a sample from a different DGP in which both the slope (= 5) and the intercept (= 10) are being estimated. In each case, the maximum likelihood estimates of $\beta_0$ and $\beta_1$ occur at the highest point on the likelihood surface, ($\beta_0^{mle}$,$\beta_1^{mle}$).

**Figure 21.3**
Log of the Likelihood Functions for Several DGPs




The likelihood function in Panel A that falls off rapidly in all directions indicates that both $\beta_0$ and $\beta_1$ are estimated relatively precisely by maximum likelihood in this instance—we cannot change either guess much from the maximum likelihood values without making the observed data much more surprising. The maximum likelihood estimates from this one sample are, in fact, very close to 10 and 5. The much flatter likelihood function in Panel B (note the ranges of $\tilde{\beta}_0$ and $\tilde{\beta}_1$ in this figure compared to those in Panel A), indicates that both $\beta_0$ and $\beta_1$ are relatively imprecisely estimated by maximum likelihood in this instance—many values other than the maximum likelihood estimates would make the observed data only a little more surprising. The maximum likelihood estimates from this one sample are, in fact, 21.4 and 1.1. The more tunnel-like surface in Panel C indicates that $\beta_0$ is relatively precisely estimated, and $\beta_1$ is relatively imprecisely estimated by maximum likelihood. Changing the guess of $\beta_0$ would make the data much more surprising, but changing the guess of $\beta_1$ would not.

The shape of a likelihood function also tells us about the covariance between maximum likelihood estimators. If we were to cut a slice of the surface in Panel A of Figure 12.3 along the axis of a particular value of $\tilde{\beta}_1$, the maximum of the likelihood with respect to $\tilde{\beta}_0$, given $\tilde{\beta}_1$, would be just about the same, no matter where along the $\tilde{\beta}_1$ axis the slice were cut. This implies that the maximum likelihood estimators in Panel A are almost completely uncorrelated. In contrast, in Panel D, slicing the surface at a different value of $\tilde{\beta}_1$ would yield a different $\tilde{\beta}_0$ that maximized the likelihood, given $\tilde{\beta}_1$. This implies that in Panel D the maximum likelihood estimators of the slope and intercept are quite correlated.

## Estimating $\sigma^2$

We can use the maximum likelihood approach to estimate $\sigma^2$ just as we use it to estimate $\beta$. To estimate $\sigma^2$, we would look at $f(Y_1, Y_2, Y_3)$ as a function of $\sigma^2$, and would maximize that function with respect to $\sigma^2$. The maximum likelihood estimator for $\sigma^2$ is the variance of the residuals, $\frac{1}{n}\sum e_i^2$. We learned in Chapter 5 that this is a biased estimator of the variance of the disturbances. Maximum likelihood estimators are frequently biased estimators, even though they are generally consistent estimators.

## Quasi-maximum Likelihood Estimation

Maximum likelihood estimation requires more information about the distribution of random variables in the DGP than does GMM or the sampling distribution estimators of earlier chapters. But the gain from the added information is considerable, as maximum likelihood estimators are often not difficult to implement and their asymptotic efficiency makes them fine estimators when we have large samples of data. Maximum likelihood estimation is commonplace in applied econometrics. Indeed, calculating maximum likelihood estimates is frequently so simple that some econometricians, when they lack better options, settle for maximum likelihood procedures even when they know the actual distribution of disturbances differs from that assumed in the maximum likelihood procedure. Using maximum likelihood procedures based upon an incorrect distribution for the errors is called **quasi-maximum likelihood estimation**. Quasi-maximum likelihood estimation risks serious bias, both in small samples and in large samples, but bias is not inevitable. For example, the maximum likelihood estimator when normal disturbances are embedded in the Gauss–Markov Assumptions is OLS. OLS provides unbiased estimates even when the true disturbances are heteroskedastic or serially correlated and the disturbances are not normally distributed. What is biased when maximum likelihood is applied to a DGP that does not satisfy the Gauss–Markov Assumptions are the standard estimates of the variances and covariances of the OLS estimators. However, we have seen in Chapters

10 and 11 that robust standard error estimators are available for such cases. Similar care needs to be taken when conducting quasi-maximum likelihood estimation in more general settings.

## 21.2 Hypothesis Testing Using Asymptotic Distributions

**How Do We Test Hypotheses?**

In Chapters 7 and 9 we encountered test statistics for hypotheses about a single coefficient (a *t*-test), about a single linear combination of coefficients (another *t*-test), about multiple linear combinations of coefficients (an *F*-test), and about coefficients across regression regimes (another *F*-test). Chapters 10 and 11 developed additional tests for detecting heteroskedasticity and serial correlation. It is not surprising that we need different tests for different hypotheses. More surprising is that econometricians often have several tests for a single hypothesis. This section explains why econometricians want multiple tests for a single use, and introduces some new tests suitable to large samples.

### An Example: Multiple Tests for a Single Hypothesis

Suppose you have been invited to your roommate's home for spring break. As you are being shown through the house, you pass through the room of Jeff, your roommate's brother. Your friends have told you that Jeff is a big fellow, but they have pulled your leg so often that you want to test this hypothesis, rather than take it on faith. You don't have much time, and you don't want to be too obvious, but you wonder how you might test the hypothesis that Jeff is large. Two options are open to you.

First, you see some shoes in Jeff's closet. If you were to look more closely at the shoes, you could use their size as your test statistic. If the shoes are sufficiently small, you would reject the claim that Jeff is large. (Of course a type I error might arise if Jeff has unusually small feet for his large stature! That is, you might wrongly reject the true null hypothesis, that Jeff is large, when he is.)

Alternatively, you could pick up the sweater that's lying folded on the bed "to admire it" and use its size as your test statistic; if the sweater is large, you might fail to reject the null hypothesis that Jeff is also large. (Of course, a type II error might result from this not being Jeff's sweater! That is, you might wrongly fail to reject the false null hypothesis, that Jeff is large, when he is not large.)

The test statistic you use will depend on both the convenience and the power of the two tests. If the closet is out of your path, it might be embarrassingly difficult to get a good look at the shoes. If you know that Jeff has several football players visiting him over break, you might think it fairly likely that if Jeff is small,

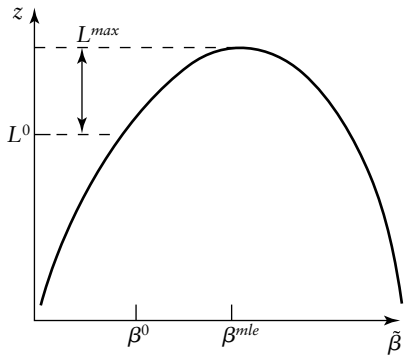



**Figure 21.4**
The Log of the Likelihood for One Specific Sample: Testing $H_0$: $\beta = \beta^0$

there might be a large friend's sweater on the bed, making the sweater a weak, and therefore unappealing, test.

In econometrics, we often choose among alternative statistics to test a single hypothesis. Which specific statistic provides the best test varies with circumstances. Sample size, the alternative hypotheses at hand, ease of computation, and the sensitivity of a statistic's distribution to violations of the assumptions of our DGP all play a role in determining the best test statistic in a given circumstance.

## The Likelihood Ratio Test for a Single Coefficient

In this section we develop an important alternative test statistic for hypotheses about $\beta$ based on the maximum likelihood property of $\beta_{gb}$. The strategy of this test generalizes to testing many more complex hypotheses. The new test statistic is based on the likelihoods of the data arising for various coefficient values. The intuition underlying tests based on likelihoods is that if the null hypothesis implies that the observed data are substantially less likely to arise than would be the case were $\beta = \beta^{mle}$, we have reason to reject the null hypothesis.

Figure 21.4 shows the log of the likelihood function for a specific sample from a DGP with a straight line through the origin. The maximum likelihood estimate of $\beta$ in Figure 21.4 is the guess $\tilde{\beta}$ at which the log of the likelihood, $z$, reaches its peak. The maximum value of the log of the likelihood function in Figure 21.4, $L^{max}$, occurs when $\tilde{\beta} = \beta^{mle}$.

In our example, the null hypothesis is $H_0$: $\beta = \beta^0$. In Figure 21.4, the value of the log of the likelihood if $\beta^0$ were the true parameter is $L^0$. The **likelihood ratio test** compares the log of the likelihood of the observed data under the null hypothesis, $L^0$, with the log of the likelihood of the observed data if the maximum likelihood estimator were correct, $L^{mle}$.

For likelihood ratio tests, econometricians use the test statistic

$$LR = 2[L(\beta^{mle}) - L(\beta^0)],$$

where $L(\beta)$ is the log-likelihood function defined in Equation 21.1. (Recall that the log of a ratio is a difference.) In large samples, $LR$ has approximately the chi-square distribution with as many degrees of freedom as there are constraints ($r$ is the number of constraints; it equals one in our example). We obtain the critical value for any significance level from a chi-square table. When $LR$ exceeds the critical value, we reject the null hypothesis at the selected level of significance. In Figure 21.4, if the difference between $L^{max}$ and $L^0$ is large enough, we reject the null hypothesis that $\beta = \beta^0$.

When might the likelihood ratio test be superior to the $t$-test? Not for a small sample from a DGP with normally distributed disturbances, because in that case the $t$-distribution holds exactly, whereas the likelihood ratio may not yet have converged to its asymptotic chi-square distribution. In practice, econometricians almost always rely on the $t$-test to test hypotheses about a single slope coefficient. But in small samples from a DGP with a known, sharply non-normal distribution of the disturbances, the likelihood ratio test may prove better than the $t$-test. In large samples, the two estimators tend toward yielding the same answer.

Likelihood ratio tests are used extensively in studying nonlinear models and nonlinear constraints, cases in which maximum likelihood estimation is common.

## An Application: The Fisher Hypothesis

In Chapter 9 we studied a model of long-term interest rates. One explanator in that model is the expected rate of inflation. Irving Fisher hypothesized that (nominal) long-term interest rates would rise percent for percent with the expected rate of inflation. Here we test Fisher's claim with a likelihood ratio test.

As in Chapter 9, we specify that long-term interest rates depend on four explanators in all, plus an intercept term:

$$\begin{aligned}(\textit{Long-term interest rate}) = {} & \beta_0 + \beta_1(\textit{expected inflation}) \\ & + \beta_2(\textit{real short-term interest rate}) \\ & + \beta_3(\textit{change in real per capita income}) \\ & + \beta_4(\textit{real per capita deficit}) + \varepsilon.\end{aligned}$$

Table 21.2 reports the regression results for this model using 1982–2000 data. The log likelihood for the equation is –12.85. Table 21.3 reports the constrained version of this model, with $\beta_1$ constrained to equal one; the dependent variable in this constrained specification is therefore the long-term interest rate minus the expected rate of inflation, and the expected rate of inflation does not appear among the explanators. The log likelihood for the constrained regression is $-13.35$.

The likelihood ratio statistic for the hypothesis that the rate of inflation has a coefficient of one is $2[-12.85 - (-13.35)] = 1.0$. The critical value for the

**Table 21.2 An Unconstrained Model of Long-term Interest Rates**

Dependent Variable: Long Term Interest Rate
Method: Least Squares
Date: 08/02/03 Time: 21:20
Sample: 1982 2000
Included observations: 19

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| **C** | **0.901969** | **0.618455** | **1.458423** | **0.1668** |
| **Real Short Rate** | **0.695469** | **0.111246** | **6.251646** | **0.0000** |
| **Expected Inflation** | **1.224125** | **0.258913** | **4.727937** | **0.0003** |
| **Change in Income** | **−0.045950** | **0.310465** | **−0.148002** | **0.8845** |
| **US Deficit** | **0.002830** | **0.001237** | **2.287053** | **0.0383** |

| | | | |
|---|---|---|---|
| R-squared | 0.953810 | Mean dependent var | 8.065877 |
| Adjusted R-squared | 0.940613 | S.D. dependent var | 2.275029 |
| S.E. of regression | 0.554410 | Akaike info criterion | 1.879110 |
| Sum squared resid | 4.303188 | Schwarz criterion | 2.127646 |
| Log likelihood | −12.85154 | F-statistic | 72.27469 |
| Durbin–Watson stat | 1.456774 | Prob(F-statistic) | 0.000000 |

chi-square distribution with one degree of freedom is 3.84 , so we fail to reject the null hypothesis. The *t*-distribution applied to the *t*-statistic for $\hat{\beta}_1$ in Table 21.2 leads to the same conclusion, though it need not do so. For example, in the special case of the Gauss–Markov Assumptions and normally distributed disturbances, the likelihood ratio statistic equals the square of the *t*-statistic. Because the disturbances are normally distributed in such cases, the *t*-distribution applies to the *t*-statistic exactly for any sample size. In small samples, the likelihood ratio statistic does not follow its asymptotic distribution—just as the *t*-statistic does not follow its asymptotic distribution (which is the normal distribution) in small samples. When the disturbances are normally distributed and the sample size is small, the *t*-test, using the exact *t*-distribution, is better than the likelihood ratio test.

## Wald Tests

Notice that the *t*-test and the likelihood ratio test call for quite different computations. We can compute the *t*-test knowing only the results of a single regression—the unconstrained regression. That regression suffices to form the *t*-statistic: $(\beta_{g4} - \beta^0)/s_{\beta_{g4}}$. The likelihood ratio test, in contrast, requires that we compute

**Table 21.3 A Constrained Model of Long-term Interest Rates**

Dependent Variable: Long Term Interest Rate − Expected Inflation
Method: Least Squares
Date: 08/02/03 Time: 21:22
Sample: 1982 2000
Included observations: 19

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| **C** | **1.343051** | **0.347559** | **3.864232** | **0.0015** |
| **Real Short Rate** | **0.763398** | **0.078190** | **9.763390** | **0.0000** |
| **Change in Income** | **−0.032644** | **0.307482** | **−0.106165** | **0.9169** |
| **US Deficit** | **0.003623** | **0.000824** | **4.394629** | **0.0005** |

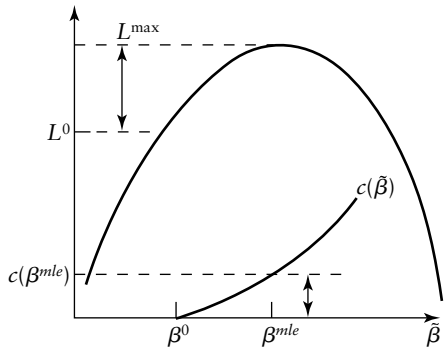
| | | | |
|---|---|---|---|
| R-squared | 0.887699 | Mean dependent var | 4.450698 |
| Adjusted R-squared | 0.865238 | S.D. dependent var | 1.497574 |
| S.E. of regression | 0.549758 | Akaike info criterion | 1.825987 |
| Sum squared resid | 4.533510 | Schwarz criterion | 2.024816 |
| Log likelihood | −13.34688 | F-statistic | 39.52302 |
| Durbin–Watson stat | 1.498800 | Prob(F-statistic) | 0.000000 |

the log of the likelihood for *both* the unconstrained model ($L^{max}$) and the constrained model ($L^0$). In general, likelihood ratio tests require computing both the unconstrained and the constrained model, whereas tests based on how hypothesized values of the parameters differ from the unconstrained estimates only require computing the unconstrained model.

We call test statistics based on how much the hypothesized parameter values differ from the unconstrained estimates **Wald tests.** If computing the constrained model is particularly difficult (as sometimes happens when the constraints are highly nonlinear or when the alternative model is only broadly defined), Wald tests may be preferable to likelihood ratio tests because Wald tests side-step computing the constrained model.

Figure 21.5 depicts Wald and maximum likelihood tests in a single picture. Our possible guesses, $\tilde{\beta}$, are on the horizontal axis. On the vertical axis, we measure two things—the log of the likelihood for this sample were the true parameter value equal to $\tilde{\beta}$, and the value of the Wald statistic for a particular guess, $c(\tilde{\beta})$. In our example, the Wald statistic is the square of the *t*-statistic, so the function $c(\tilde{\beta})$ shown in the figure is parabolic. As noted earlier, the likelihood ratio test looks at

**Figure 21.5**
The Log-likelihood Function and the Wald Statistic




the difference between $L^{max}$ and $L^0$. The Wald test looks at the difference between the Wald function evaluated at $\beta^{mle}$, $c(\tilde{\beta})$, and 0. The comparison is to 0, because if the estimated value equals the hypothesized value exactly, $\beta_0$, the *t*-statistic, and therefore the Wald function, will be zero.

## Wald Tests or Likelihood Ratio Tests

Computational advantages are not the only reason econometricians sometimes turn to Wald tests instead of likelihood ratio tests. A Wald test sometimes requires weaker assumptions about the DGP than does the likelihood ratio test; in such cases, the Wald test can be preferable. Here we examine the choice between Wald and likelihood ratio tests in the context of testing for a change in regimes between two time periods.

Recall from Chapter 9 that we use the *F*-statistic to test multiple linear restrictions on coefficients, including hypotheses that a model's regression coefficients are the same in two time periods. The *F*-test is close kin to the likelihood ratio test in that both require computing both the constrained and the unconstrained model. The *F*-statistic is based on the sum of squared residuals in the constrained and unconstrained regressions:

$$F = \frac{\dfrac{SSR^{c} - SSR^{u}}{r}}{\dfrac{SSR^{u}}{(n - k - 1)}}.$$

Like the *t*-statistic, the *F*-statistic follows its *F*-distribution [with $r$ and $(n - k - 1)$ degrees of freedom] only if the underlying disturbances are normally distributed and only if the constraints among the parameters are linear in form. However, in

large samples, $rF$ is asymptotically distributed chi-square (with $r$ degrees of freedom) under the Gauss–Markov Assumptions, even if the underlying disturbances are not normally distributed and even if the constraints are nonlinear.

The Wald statistic for testing several linear hypotheses about the slope and intercept of a line is a weighted sum of the squares and cross products of the deviations of the unconstrained estimates from their hypothesized values. For example, when testing the equality of coefficients between two regression regimes, A and B, the Wald statistic, $W$, is a weighted sum of $(\beta_0^{mleA} - \beta_0^{mleB})^2$, $(\beta_1^{mleA} - \beta_1^{mleB})^2$, and $(\beta_0^{mleA} - \beta_0^{mleB})(\beta_1^{mleA} - \beta_1^{mleB})$, with the weights depending on the variances and covariances of the estimators. The Wald statistic is asymptotically distributed chi-square with $r$ degrees of freedom, where $r$ is the number of constraints (in this example, $r = 2$).

Which estimator should we use to test the hypothesis that a regression model's coefficients are the same in two time periods? The $F$-statistic (or its chi-square cousin, $rF$), or the Wald statistic?[2] The $F$-statistic is not computationally difficult to obtain, despite needing both the constrained and unconstrained sums of squared residuals, so there is little computational reason for preferring the Wald statistic to the $F$-statistic. And there may be sound reason for preferring the $F$-statistic over the Wald statistic. In small samples, the Wald statistic will not follow its asymptotic distribution, whereas the $F$-statistic will exactly follow the $F$-distribution if the underlying disturbances are normally distributed. Thus, the $F$-statistic frequently performs better than the Wald statistic in small samples. Sometimes, though, the Wald statistic is, indeed, preferable to the $F$-statistic. For example, when testing the hypothesis that regression coefficients are the same in two time periods, the distribution of the $F$-statistic and its chi-square cousin depend on the Gauss–Markov Assumption of homoskedastic disturbances between the two regimes, $(\sigma_A^2 = \sigma_B^2)$. The Wald statistic's asymptotic distribution does not depend on such homoskedasticity. Thus, when we have substantial doubts about the homoskedasticity of the disturbances across regimes, the Wald statistic is preferable to the $F$-statistic.

This choice between the Wald and $F$ when testing for changes in regime captures the elements econometricians generally face when choosing among competing tests. Which tests are easiest to perform? Which perform best in finite samples? Which are most robust to any likely violations of our maintained hypotheses?

## Lagrange Multiplier Tests

Wald tests offer the advantage of only requiring estimates of the unconstrained model, whereas likelihood ratio tests require estimates of both the unconstrained and the constrained models. A third class of tests, *Lagrange multiplier tests*, only require estimates of the constrained model. The maximum likelihood estimator

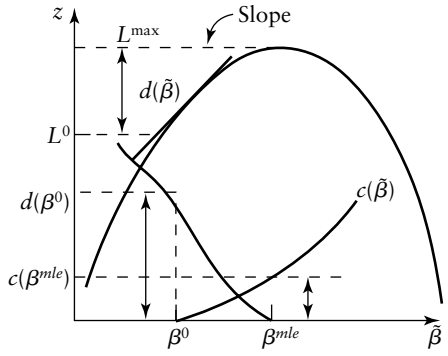



**Figure 21.6**
Likelihood Ratio, Wald, and Lagrange Multiplier Tests

maximizes the likelihood function. At the maximum, calculus tells us, the slope of the likelihood function is zero—the log of the likelihood function is flat at its peak. One measure of being "far" from the unconstrained maximum likelihood estimate, then, is the slope of the likelihood function when we impose the null hypothesis. If there are multiple constraints, there is a slope of the likelihood function with respect to each constraint. **Lagrange multiplier tests** combine the slopes of the likelihood function with respect to each constraint in an appropriate fashion to obtain a test statistic that is asymptotically distributed chi-square with $r$ degrees of freedom, where $r$ is the number of constraints.[3]

In Chapter 10 we encountered the White test for heteroskedasticity. The White test is a Lagrange multiplier test. Because a Lagrange multiplier test only requires that we estimate the constrained (homoskedastic) model, the White test spares us specifying the actual form of heteroskedasticity that might plague us.

## Picturing the Three Test Strategies

Building on Figure 21.5, Figure 21.6 provides a unifying picture of all three testing strategies, likelihood ratio tests, Wald tests, and Lagrange multiplier tests. On the horizontal axis are the possible estimates, $\tilde{\beta}$. On the vertical axis we measure, in turn, the likelihood of a sample, $z$, if the true parameter equals $\tilde{\beta}$, the constraint expression of the Wald statistic, $c(\tilde{\beta})$, and a function, $d(\tilde{\beta})$, that shows the slope of the likelihood function at each value of $\tilde{\beta}$. As in Figure 21.5, the likelihood ratio test depends on the difference between $L^{max}$ and $L^0$, and the Wald test depends on the difference between $c(\beta^{mle})$ and 0. The Lagrange multiplier test depends on the difference between $d(\beta_0)$ and 0.

Figure 21.6 highlights that likelihood ratio tests require computing both the constrained and unconstrained models (to obtain $L^{max}$ and $L^0$), whereas Wald tests require only the unconstrained model [to obtain ($\beta^{mle}$)], and Lagrange multiplier tests require only the constrained model [to obtain $d(\beta^0)$].


# An Econometric Top 40—A Classical Favorite

## The Translog Production Frontier

Economists make simplifying assumptions to avoid analytical tasks that are impossible or impractical. But what is possible and what is practical changes over time. Sometimes, new mathematical techniques make simpler what was previously too difficult. Sometimes, it is technological advances in computation. In this hit, we revisit relatively simple production functions that appeared in Golden Oldies in earlier chapters and see how the introduction of inexpensive high-speed computing altered the technology of representing technology. In the 1960s, faster computers made a richer specification of technology feasible. With a new, more sophisticated specification of technology, economists were able to test more sophisticated questions than previously, such as "Do firms maximize profits?" and "Can technological change be accurately summarized in a single measure?"

The Cobb–Douglas production function appeared in a Golden Oldie in Chapter 3. The Cobb–Douglas form for a production function is

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(L_i) + \beta_2 \ln(K_i) + \beta_3 \ln(M_i),$$

where $Y$ is output, $L$ is labor, $K$ is capital, and $M$ is materials. In another Golden Oldie, in Chapter 4, we encountered the constant elasticity of substitution (CES) production function:

$$Y^{\rho} = \tau_1 L^{\rho} + \tau_2 K^{\rho} + \tau_3 M^{\rho},$$

which was introduced in 1961 using just two inputs, capital and labor, rather than the three inputs shown here.[4]

Both the Cobb–Douglas and the CES production functions simplify the general form $Y = F(L,K,M)$ by assuming that $Y$, $L$, $K$, and $M$ can each be separated from the others in its own subfunction, and $F$ can be reformed by adding those separate functions together: $g(Y) = h(L) + q(K) + r(M)$. Because this additive form is easier to manipulate than other forms, economic theorists have made extensive use of both the Cobb–Douglas and CES production functions. Early econometric applications also relied heavily on these forms, largely because estimating their parameters was much simpler than was estimating more complicated production functions.

Improvements in computer technology during the 1960s eased the computational constraints faced by econometricians and opened the door to more complicated specifications. In 1973, Laurits Christiansen of the University of Wisconsin, Dale Jorgenson of Harvard University, and Lawrence Lau of Stanford University introduced a richer specification of technology, called the translog production function:

$$\begin{aligned}\ln(Y_i) = {} & \alpha_0 + \alpha_L \ln(L) + \alpha_k \ln(K_i) + \alpha_M \ln(M_i) \\ & + \alpha_{LK} \ln(L_i)\ln(K_i) + \alpha_{LM} \ln(L_i)\ln(M_i) \\ & + \alpha_{KM} \ln(K_i)\ln(M_i) + \alpha_{LL} \ln(L_i)^2 \\ & + \alpha_{KK} \ln(K_i)^2 + \alpha_{MM} \ln(M_i)^2,\end{aligned}$$

a quadratic in logarithms that has since become a standard specification in many empirical production studies. The translog allows us to test whether the interaction terms, such as $\ln(L_i)\ln(M_i)$, are required in the production function specification and whether either of the two special cases, the Cobb–Douglas and the CES, apply. Although studies of some industries or firms find that the data do not reject the simpler specifications, numerous studies of other industries and firms do reject the simpler specifications in favor of the translog form.

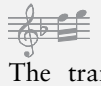
Christiansen, Jorgenson, and Lau's original use of the translog was actually more ambi-

tious than the translog production function, with its single output and several inputs. The authors use the quadratic in logs form to describe the production possibilities frontier of the United States, with two outputs, investment goods and consumption goods, and two inputs, capital and labor. They have four ambitious objectives that we review here. First, they test the neoclassical theory of production that claims that firms maximize profits. If markets are competitive, the neoclassical theory says that the ratios of goods' prices equal the ratios of those goods' marginal rates of transformation. Second, they test the hypothesis that technological change can be measured by a single variable, an index called "total factor productivity" that appears like another input in the translog expression. Third, they test whether interaction terms like $\ln(L_i)\ln(M_i)$ belong in the model. Finally, they test the Cobb–Douglas and CES specifications of the production frontier.

Christiansen, Jorgenson, and Lau test all these hypotheses by examining the coefficients of the translog production function and two other equations derived from the translog specification and economic theory. The first additional equation expresses the value of investment goods production relative to the value of capital inputs as a linear function of the logs of the two inputs, the two outputs, and the technology index. The second additional equation expresses the value of labor inputs relative to the value of capital inputs, also as a linear function of the logs of the two inputs, the two outputs, and the technology index. The data are annual observations on the U.S. economy from 1929 to 1969.

Christiansen, Jorgenson, and Lau demonstrate that if the production frontier has the translog shape, then each of the hypotheses they offer implies a particular relationship among the coefficients of the two relative value regressions and the translog production frontier. Thus, the authors arrive at a complex variant on the hypothesis tests we have already devised. They estimate not one equation, but three, and their constraints on the coefficients constrain not just the coefficients within one equation, but also the coefficients across the three equations. Christiansen, Jorgenson, and Lau estimate their three equations by maximum likelihood, and their various test statistics are likelihood ratio tests that are asymptotically distributed chi-square with as many degrees of freedom as there are constraints.

What do the authors conclude? They fail to reject the neoclassical theory of production and the appropriateness of total factor productivity as a one-dimensional measure of technology. But they do reject the exclusion of interaction terms like $\ln(L_i)\ln(M_i)$ and consequently both the Cobb–Douglas and CES forms for the U.S. production frontier. They find the technological patterns of substitution between capital and labor are more complex than allowed for by either the Cobb–Douglas or the CES production function.



## Final Notes

The translog production function has been used in hundreds of papers since it was introduced by Christiansen, Jorgenson, and Lau. Maximum likelihood estimation offers one attractive strategy for taming the complex of equations and hypotheses that the translog presents. To obtain their maximum likelihood estimates, Christiansen, Jorgenson, and Lau did not directly maximize the likelihood function. Instead, they followed an iterative computational algorithm that did not refer to the likelihood function, but that converged to the same solution as if they had maximized the likelihood directly. Contemporary econometric software packages, many of which were developed during the 1970s, usually avoid such indirect routes to maximum likelihood estimates.

Advances in computer technology made the computations undertaken by Christiansen,

Jorgenson, and Lau possible. Since their work in 1973, even greater strides have been made in computing technology—and in the writing of econometric software programs. Students of econometrics today have access to computers far more powerful and to software far more friendly than were available to even the most sophisticated econometricians in 1973. These advances allow us to rely on more complex estimation methods, to conduct more complex tests, and to use much larger data sets than we could in the past. Computer scientists, econometric theorists, and data gatherers in both the public and the private sectors have combined to give economists empirical tools that were undreamed of a short while ago. ■

When several different tests are available for a single hypothesis, econometricians will balance computational ease, the burden of required assumptions, and the performance of each estimator in finite samples, to settle on a best choice. In large samples, if the null hypothesis is true, the likelihood ratio, Wald, and Lagrange multiplier tests all tend to the same answer. However, even in large samples, the three estimators can differ in their power against various alternatives.

***An Organizational Structure for the Study of Econometrics***

1. **What Is the DGP?**
   Maximum likelihood requires knowing the the joint distribution of the disturbances and *X*'s.
2. **What Makes a Good Estimator?**
   Consistency and asymptotic efficiency.
3. **How Do We Create an Estimator?**
   Maximum likelihood.
4. **What Are an Estimator's Properties?**
   Maximum likelihood usually yields asymptotic efficiency.
5. **How Do We Test Hypotheses?**
   Likelihood ratio, Wald, and Lagrange multiplier tests.

## Summary

The chapter began by introducing a strategy, maximum likelihood estimation, for constructing asymptotically efficient estimators. Maximum likelihood estimation

requires a complete specification of the distribution of variables and disturbances in the DGP, which is often an unfillable order. But when such detailed distributional information is available, maximum likelihood provides asymptotically normally distributed, asymptotically efficient estimators under quite general conditions.

The chapter then introduced three classes of hypothesis tests: likelihood ratio tests, Wald tests, and Lagrange multiplier tests. The three differ in their computational requirements, maintained hypotheses, small sample performance, and power against various alternative hypotheses. Likelihood ratio tests require estimates of both the unconstrained and the constrained model. Wald tests only require estimates of the unconstrained model. Lagrange multiplier tests only require estimates of the constrained model. When the null hypothesis is true, all three tests tend toward the same conclusion in large samples, but in small samples, the statistics may lead to conflicting conclusions, even when the null hypothesis is true.

## Concepts for Review

**Lagrange multiplier tests**
**Likelihood ratio test**
**Maximum likelihood estimate**
**Quasi-maximum likelihood estimation**
**Wald tests**

## Questions for Discussion

1. "Maximum likelihood estimation provides the coefficient estimates that are most likely to be true." Agree or disagree, and discuss.

## Problems for Analysis

1. Show that under the Gauss–Markov Assumptions, the maximum likelihood estimator of $\sigma^2$ is the mean squared residual.
2. When our null hypothesis is sharply defined, but the potential alternatives are only vaguely known, which family of test statistics is most apt to yield us a useful way to test our null hypothesis? Briefly explain.
3. When the DGP includes complete information about the distributions of all random variables appearing in the DGP, and we can precisely state what the world looks like under both the null and alternative hypotheses, which family of test statistics is most apt to yield us a useful way to test our null hypothesis?

## Endnotes

1. For example, if the disturbances follow the double exponential distribution, the maximum likelihood estimator for $\beta$ is the estimator $\tilde{\beta}$ such that $(Y_i - X_i\tilde{\beta})$ has a median of zero.
2. In the special case of linear constraints on a linear regression with a DGP that satisfies the Gauss-Markov Assumptions, the Wald test statistic is exactly *rF*. Thus, in this special (but frequently encountered) case, the *F*-statistic is close kin (essentially an identical twin) to the Wald statistic, as well as to the likelihood ratio statistic. For this reason, some statistical packages refer to the *F*-test as a Wald test, and contrast it with likelihood ratio tests. However, in general, the Wald statistic differs substantively from the *F*-statistic. For this reason, other statistical packages report both an *F*-statistic and a Wald test statistic when testing constraints on coefficients.
3. When the Gauss-Markov Assumptions hold and disturbances are normally distributed, the Lagrange multiplier statistic is the same as *rF*, except that in the denominator we replace the sum of squared residuals from the unconstrained regression, with those from the constrained regression.
4. K. A. Arrow, Hollis Chenery, B. S. Minhas, and Robert Solow, "Capital-Labor Substitution and Economic Efficiency," *Review of Economics and Statistics* (August 1961): 225–250.