Hk

Note that the population mean of $Y_i - \mu$ is zero here by construction; remarkably, this turns out to be the key necessary assumption for a version of the Central Limit Theorem to apply, leading to a normal distribution in large samples. In fact, the IID assumption (and hence the equal weighting) is substantially relaxed in other versions of the Central Limit Theorem.[21]

How large does N need to be in order for the Central Limit Theorem to imply that a sum like $\left(1/\sqrt{N}\right)\sum_{i=1}^{N}(Y_i - \mu)$ is essentially normally distributed? It depends, basically, on two things. First, it depends on just how non-normal the distribution of the Y_i actually is. After all, if the Y_i are normally distributed, then $\left(1/\sqrt{N}\right)\sum_{i=1}^{N}(Y_i - \mu)$ is exactly normally distributed – even for $N = 1$ – because the second property of normally distributed random variables states that any weighted sum of normally distributed random variables is itself normally distributed. Consequently, one should expect that even small values of N will suffice if the density function for the Y_i is vaguely bell-shaped.

The other crucial issue hinges on exactly what one means by "essentially normal." If one seeks to broadly characterize the shape of the distribution of $\left(1/\sqrt{N}\right)\sum_{i=1}^{N}(Y_i - \mu)$ over the values which it is most likely to take on, then small values of N (such as 20) will usually suffice, even if the density function for the individual Y_i is not particularly bell-shaped. Unfortunately, for the statistical inference applications in which the Central Limit Theorem is used below, what is usually important is the shape of the *tails* of the distribution of random variables like $\left(1/\sqrt{N}\right)\sum_{i=1}^{N}(Y_i - \mu)$, and this part of the distribution converges to normality relatively slowly. With values of N greater than 100, however, people generally feel comfortable using the Central Limit Theorem even for the tails of the distribution.[22]

2.12 DISTRIBUTION OF A LINEAR COMBINATION OF NORMALLY DISTRIBUTED RANDOM VARIABLES

Many of the quantities dealt with below (parameter estimators, predictions, and the like) will turn out to be linear combinations (weighted sums) of normally distributed random variables, so it is useful to derive the distribution of such a linear combination here. In this way, the mechanics of the derivation will be familiar when the distribution of this kind of variable is considered later on. To make plain exactly what role the normality assumption plays in the derivation, at the outset the underlying random variables will not be assumed to be normally distributed. Instead, it will only be assumed that these random variables – here called $U_1 \ldots U_N$ to correspond with the notation used in Chapter 5 – all have the same mean and variance.

Why is the distribution of a parameter estimator all that important? Foreshadowing some of th results in the chapters to come, the basic motivation is this: When one estimates a parameter in model using the sample data, all one gets is a number – a single realization of the estimator of tha parameter. But we want more. We want to be able to say something about whether or not this is *good* estimate of the parameter. For that, we need to know whether the distribution of this estimatc

[21] For example, a more general form of the Central Limit Theorem is invoked in Appendix 12-1 (Chapter 12), where tl asymptotic distribution of the parameter estimator which is the topic of this chapter – and which involves a sum of zero-mea random variables with possibly differing variances – is derived. Also, ~~while~~ it is not an essential issue at this point, but t $1/\sqrt{N}$ is needed so that the variance of the random variable will tend toward a positive, finite consta

where the factor of $\frac{1}{N}$ can come inside the sum (because it is a constant) and where the Linearity Property is used to express this expectation of a weighted sum of random variables as the weighted sum of the expectations of the random variables.

Note that averaged over its sampling distribution, $\overline{Y}$ is equal to the population mean it is trying to estimate. This implies that $\overline{Y}$ has the property of unbiasedness or, equivalently, that $\overline{Y}$ is an unbiased estimator of μ. For a general estimator $\hat{\theta}$ which is being used to estimate a population parameter θ,

Bias and the Unbiasedness Property

Bias: If $\hat{\theta}$ is an estimator of θ, $\text{bias}(\hat{\theta}) \equiv E[\hat{\theta}] - \theta$ (3-16)

Unbiasedness: If $\hat{\theta}$ is an estimator of θ, then $\hat{\theta}$ is an unbiased estimator of θ if and only if $\text{bias}(\hat{\theta}) = 0$ or, equivalently, $E[\hat{\theta}] = \theta$

Note that most of the content of the assumption that $Y_i \sim \text{NIID}[\mu, \sigma^2]$ is never actually used in the proof given above that $\overline{Y}$ is an unbiased estimator of μ. In fact, this property for $\overline{Y}$ depends only on the assumption that each observation on Y has the same population mean, μ. Thus, one of the key advantages to understanding the derivation of the unbiasedness of $\overline{Y}$ is the realization that this result is very insensitive to failures of the actual data to conform to our assumptions about it. In particular, the derivation makes it plain that $\overline{Y}$ will still be unbiased for μ even if the observations are not normally distributed, have differing variances, and are not independent of one another!

The issue of how much value should be placed on unbiasedness – how "good" this property really is – is taken up shortly. In the remainder of this section, the variance of the sampling distribution of $\overline{Y}$ – the "sampling variance" of $\overline{Y}$ – is explicitly derived for this specific case. The purpose for doing this is to make it more clear where the result $\text{var}(\overline{Y}) = \frac{\sigma^2}{N}$ comes from and, as in the unbiasedness derivation above, to clarify precisely how this result rests on the validity of the model assumptions.

The derivation begins by re-expressing the variance of $\overline{Y}$ using its definition in terms of the expectations operator; then both $\overline{Y}$ and $E[\overline{Y}]$ are rewritten as sums over the N observation so that $E\left[\left[\overline{Y} - E[\overline{Y}]\right]^2\right]$ can be expressed as the expectation of the square of a single sum:

$$
\begin{aligned}
\text{var}\{\overline{Y}\} = E\left[\left[\overline{Y} - E[\overline{Y}]\right]^2\right] &= E\left[\left[\sum_{i=1}^{N}\frac{1}{N}Y_i - \sum_{i=1}^{N}\frac{1}{N}\mu\right]^2\right] \\
&= E\left[\left[\sum_{i=1}^{N}\left(\frac{1}{N}Y_i - \frac{1}{N}\mu\right)\right]^2\right] \\
&= E\left[\left[\sum_{i=1}^{N}\frac{1}{N}(Y_i - \mu)\right]^2\right] \\
&= E\left[\left[\sum_{i=1}^{N}\frac{1}{N}(Y_i - \mu)\right]\left[\sum_{j=1}^{N}\frac{1}{N}(Y_j - \mu)\right]\right]
\end{aligned}
\qquad (3\text{-}17)
$$

They look rather similar, but the MSE($\hat{\mu}$) and the var($\hat{\mu}$) are identical only for an unbiased estimator:

$$\boxed{\begin{aligned}\text{MSE}(\hat{\mu}) &= \text{dispersion of } \hat{\mu} \text{ around true value of } \mu = E\left[(\hat{\mu}-\mu)^2\right]\\ \text{var}(\hat{\mu}) &= \text{dispersion of } \hat{\mu} \text{ around its expected value} = E\left[(\hat{\mu}-E[\hat{\mu}])^2\right]\end{aligned}} \tag{3-22}$$

There is, however, a simple relationship between the MSE($\hat{\mu}$) and the var($\hat{\mu}$): the mean square error of any estimator equals its sampling variance plus the square of its bias. We will use this result in a number of contexts below, so it is worth proving in general – i.e., not just for the estimator of the population mean of a normally distributed variable. Let $\hat{\mu}$ be an estimator of a parameter μ, then:[8]

$+[\text{bias}(\hat{\mu})]^2$

$$\begin{aligned}\text{MSE}(\hat{\mu}) &= E\left[(\hat{\mu}-\mu)^2\right]\\ &= E\left[(\hat{\mu}-E[\hat{\mu}]+E[\hat{\mu}]-\mu)^2\right]\\ &= E\left[([\hat{\mu}-E[\hat{\mu}]]+[E[\hat{\mu}]-\mu])^2\right]\\ &= E\left[[\hat{\mu}-E[\hat{\mu}]]^2+2[E[\hat{\mu}]-\mu][\hat{\mu}-E[\hat{\mu}]]+[E[\hat{\mu}]-\mu]^2\right]\\ &= E\left[\left([\hat{\mu}-E[\hat{\mu}]]^2+2[\text{bias}(\hat{\mu})][\hat{\mu}-E[\hat{\mu}]]+[\text{bias}(\hat{\mu})]\right)^2\right]\\ &= E\left[[\hat{\mu}-E[\hat{\mu}]]^2\right]+2\text{bias}(\hat{\mu})E[\hat{\mu}-E[\hat{\mu}]]+[\text{bias}(\hat{\mu})]^2\\ &= E\left[[\hat{\mu}-E[\hat{\mu}]]^2\right]+2\text{bias}(\hat{\mu})\,0+[\text{bias}(\hat{\mu})]^2\\ &= \text{var}(\hat{\mu})+[\text{bias}(\hat{\mu})]^2\end{aligned} \tag{3-23}$$

Thus, because the expected loss of $\hat{\mu}$ (due to the errors $\hat{\mu}$ makes in estimating μ) is proportional to MSE($\hat{\mu}$) under a squared error loss function, small sampling variance and small squared bias in $\hat{\mu}$ are both good in the sense of yielding lower expected losses.

What, then, is the optimal ("best") estimator under the squared error loss function assumption? Returning to the simple case of estimating the population mean of a normally distributed variable, suppose that we assume that the optimal estimator of μ is a constant (k) times $\hat{\mu}_{LS} = \overline{Y}$ and calculate the value of k which minimizes the expected losses incurred by this estimator. This is the value of k which minimizes:

$$\begin{aligned}E\left[(k\overline{Y}-\mu)^2\right] = \text{MSE}(k\overline{Y}) &= \text{var}(k\overline{Y})+\left[\text{bias}(k\overline{Y})\right]^2\\ &= k^2\text{var}(\overline{Y})+\left[E\left[k\overline{Y}\right]-\mu\right]^2\\ &= k^2\left(\frac{\sigma^2}{N}\right)+[k\mu-\mu]^2\\ &= k^2\left(\frac{\sigma^2}{N}\right)+[k-1]^2\mu^2\end{aligned} \tag{3-24}$$

Thus, k^* – the optimal value of k – satisfies

$$\begin{aligned}0 = \frac{\text{dMSE}(k\overline{Y})}{\text{d}k} &= 2k^*\left(\frac{\sigma^2}{N}\right)+2[k^*-1]\mu^2\\ &= 2k^*\left(\frac{\sigma^2}{N}+\mu^2\right)-2\mu^2\end{aligned} \tag{3-25}$$

[8] Note that in the following derivation $E[\hat{\mu}]$ is added and subtracted because the sampling variance of $\hat{\mu}$ is the expected value of $(\hat{\mu}-E[\hat{\mu}])^2$. Also note that the bias $(\hat{\mu}) \equiv E[\hat{\mu}]-\mu$ is a fixed (not a random) quantity.

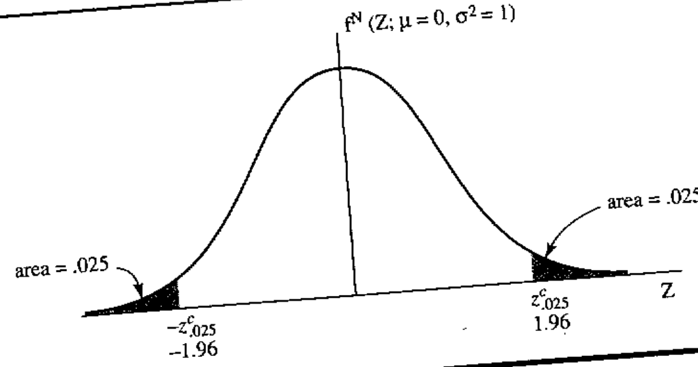

Figure 4-1 The 2½% Critical Point of the Unit Normal Distribution

This implies that

$$-z^c_{.025} \le \frac{\bar{Y}-\mu}{\sqrt{\frac{\sigma^2}{N}}} \le z^c_{.025} \quad \text{with probability .95} \tag{4-6}$$

Multiplying both sides of each of these inequalities by $\sqrt{\sigma^2/N}$ yields

$$\sqrt{\frac{\sigma^2}{N}} \longleftarrow -z^c_{.025} \le \bar{Y}-\mu \le z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \quad \text{with probability .95} \qquad \text{(4-7}$$

And subtracting $\bar{Y}$ from both sides of each inequality yields

$$-\bar{Y} - z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \le -\mu \le -\bar{Y} + z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \quad \text{with probability .95} \qquad \text{(4-}$$

Finally, multiplying both sides of each inequality by –1 (and therefore flipping the sense the inequalities from "≤" to "≥") yields a 95% confidence interval for **μ**:

$$\bar{Y} + z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \ge \mu \ge \bar{Y} - z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \quad \text{with probability .95} \qquad \text{(4}$$

or

$$\bar{Y} + 1.96\sqrt{\frac{\sigma^2}{N}} \ge \mu \ge \bar{Y} - 1.96\sqrt{\frac{\sigma^2}{N}} \quad \text{with probability .95} \qquad \text{(4-}$$

Because σ^2 is known, realizations of the endpoints of this interval can be estimated from the sar data – i.e., from the observed sample mean, $\bar{y}$.

This estimated confidence interval conveys very useful information as to how precisely $\bar{Y}$ estim μ. Note that this confidence interval does *not* indicate that there is any uncertainty in the value of μ What the interval *does* do is quantify how the sampling variability 10,000 indeper

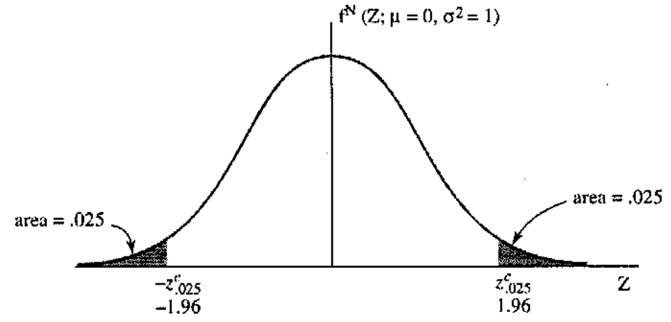

Figure 4-1 The 2½% Critical Point of the Unit Normal Distribution

This implies that

$$-z^c_{.025} \le \frac{\overline{Y} - \mu}{\sqrt{\frac{\sigma^2}{N}}} \le z^c_{.025} \quad \text{with probability .95} \tag{4-6}$$

Multiplying both sides of each of these inequalities by $\sqrt{\sigma^2/N}$ yields

$$-z^c_{.025}\underset{\wedge}{\overset{\sqrt{\frac{\sigma^2}{N}}}{}} \le \overline{Y} - \mu \le z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \quad \text{with probability .95} \tag{4-7}$$

And subtracting $\overline{Y}$ from both sides of each inequality yields

$$-\overline{Y} - z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \le -\mu \le -\overline{Y} + z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \quad \text{with probability .95} \tag{4-8}$$

Finally, multiplying both sides of each inequality by −1 (and therefore flipping the sense of the inequalities from "≤" to "≥") yields a 95% confidence interval for μ:

$$\overline{Y} + z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \ge \mu \ge \overline{Y} - z^c_{.025}\sqrt{\frac{\sigma^2}{N}} \quad \text{with probability .95} \tag{4-9}$$

or

$$\overline{Y} + 1.96\sqrt{\frac{\sigma^2}{N}} \ge \mu \ge \overline{Y} - 1.96\sqrt{\frac{\sigma^2}{N}} \quad \text{with probability .95} \tag{4-10}$$

Because σ^2 is known, realizations of the endpoints of this interval can be estimated from the sample data – i.e., from the observed sample mean, $\bar{y}$.

This estimated confidence interval conveys very useful information as to how precisely $\overline{Y}$ estimates μ. Note that this confidence interval does *not* indicate that there is any uncertainty in the value of μ: μ's value is unknown but fixed. What the interval *does* do is quantify how the sampling variability in $\overline{Y}$ impacts our ability to "know" μ using a single realization of $\overline{Y}$. Imagine drawing 10,000 independent samples of $Y_1 \ldots Y_N$, yielding 10,000 realizations of $\overline{Y}$ and hence 10,000 different 95% confidence intervals for μ. We can expect that around 9,500 of these intervals will overlap the true value of μ and that about 500 of them will not.

4.5 USING S^2 TO ESTIMATE σ^2 (AND INTRODUCING THE CHI-SQUARED DISTRIBUTION)

The population variance (σ^2) is never known in practical circumstances, so the results on confidence intervals and hypothesis testing given above are not yet complete. To make these results complete, σ^2 must be estimated and the fact that this estimator is a random variable with its own sampling distribution must be explicitly taken into account. In particular, it will be shown explicitly in the next section that it suffices to find an estimator of σ^2 which is distributed as a chi-squared variate and is independent of $\overline{Y}$; identifying and analyzing such an estimator is the topic of this section.

If μ were known, finding an estimator of σ^2 and obtaining its sampling distribution would be quite easy. In that case, the natural choice for an estimator of σ^2 would be the sample variance,

$$\hat{\sigma}^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2 \tag{4-18}$$

Note that (using the Linearity Property and the definition of σ^2) it is simple to show that $\hat{\sigma}^2$ is unbiased:

$$E\left[\hat{\sigma}^2\right] = E\left[\sum_{i=1}^{N}\frac{1}{N}(Y_i - \mu)^2\right] = \sum_{i=1}^{N}\frac{1}{N}E\left[(Y_i - \mu)^2\right] = \sum_{i=1}^{N}\frac{1}{N}\sigma^2 = \sigma^2 \tag{4-19}$$

This estimator of σ^2 also has a simple sampling distribution: $N\hat{\sigma}^2/\sigma^2$ is chi-squared with N degrees of freedom. The chi-squared distribution is most conveniently defined as follows:

Chi-Squared Distribution

If

$$Z_i \sim \text{NIID}\overset{[}{\{}0, 1] \quad \text{for} \quad i = 1 \ldots m,$$

then

$$Q \equiv \sum_{i=1}^{m} Z_i^2 \sim \chi^2(m) \tag{4-20}$$

"Q is chi-squared with m degrees of freedom"

So any random variable which can be written as the sum of the squares of m independently distributed unit normals is distributed χ^2 with m degrees of freedom.

Multiplying $\hat{\sigma}^2$ by N and dividing by σ^2 yields

$$\frac{N\hat{\sigma}^2}{\sigma^2} = \frac{N}{\sigma^2}\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2\right) = \sum_{i=1}^{N}\frac{1}{\sigma^2}(Y_i - \mu)^2 = \sum_{i=1}^{N}\left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2(N) \tag{4-21}$$

because the $(Y_i - \mu)/\sigma$ are independent unit normal variables when Y_i is distributed NIID $[\mu, \sigma^2]$.

But μ is not known, so $\hat{\sigma}^2$ is not usable as an estimator of σ^2. Instead, it is necessary to replace μ by an estimator $(\overline{Y})$, yielding the unbiased estimator

$$S^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 \tag{4-22}$$

It is by no means obvious that S^2 is unbiased. This and several additional key results on S^2 are proven in the remainder of this section using the following intermediate results:

$\bar{Y}$

Using the intermediate results of Equation 4-23 to evaluate these three expectations,

$$\begin{aligned} E[S^2] &\equiv \frac{1}{N-1}\sum_{i=1}^{N}\left[E[Y_i^2] - 2E[Y_i\boxed{\bar{Y}}] + E[\bar{Y}^2]\right] \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left[(\mu^2+\sigma^2) - 2\left(\mu^2+\frac{\sigma^2}{N}\right) + \left(\mu^2+\frac{\sigma^2}{N}\right)\right] \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left[\sigma^2 - 2\frac{\sigma^2}{N} + \frac{\sigma^2}{N}\right] \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left[(N-2+1)\frac{\sigma^2}{N}\right] = \sigma^2 \end{aligned} \tag{4-26}$$

Thus, using the factor $1/(N-1)$ instead of $1/N$ in the definition of S^2 makes it an unbiased estimator of σ^2. Referring back to the derivation earlier in this section showing that $E[\hat{\sigma}^2] = \sigma^2$, it is plain that the factor $1/(N-1)$ is needed in the definition of S^2 because the observable estimator $\bar{Y}$ is being substituted for the unknown population mean μ. What is going on here is that the observed deviations $y_1 - \bar{y} \ldots y_N - \bar{y}$ are (on average) a bit smaller in magnitude because these deviations are being calculated from the same $y_1 \ldots y_N$ data used in computing $\bar{y}$; therefore, the sum of the squared deviations in the estimator S^2 must be "blown up" a bit by the $1/(N-1)$ factor to compensate.

In order to derive confidence intervals and hypothesis tests for μ when σ^2 is unknown, it is necessary to show that $(N-1)S^2/\sigma^2$ is distributed $\chi^2(N-1)$ independently from $\bar{Y}$. The required independence is easy to show (and is derived later in this section), but the derivation that $(N-1)$ S^2/σ^2 is distributed $\chi^2(N-1)$ requires matrix algebra which is beyond the scope of this book.[8] Consequently, this result is instead motivated here by showing that $(N-1)$ S^2/σ^2 is approximately distributed $\chi^2(N)$ for large N; this turns out to be quite easy.

First, multiply the expression for S^2 by $N-1$ and divide it by σ^2 to yield

$$\frac{(N-1)S^2}{\sigma^2} = \frac{N-1}{\sigma^2}\left(\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2\right) = \sum_{i=1}^{N}\frac{1}{\sigma^2}(Y_i-\bar{Y})^2 = \sum_{i=1}^{N}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 \tag{4-27}$$

Now note that $(Y_i - \bar{Y})/\sigma$ is normally distributed because it is the weighted sum of two normally distributed variables; and it has mean zero because $E[Y_i]$ and $E[\bar{Y}]$ both equal μ. But $(Y_i - \bar{Y})$ is not a unit normal because – referring to the "intermediate results" box – its variance is $1 - \frac{1}{N}$ rather than one. For sufficiently large N, however, the variance of $(Y_i - \bar{Y})/\sigma$ is arbitrarily close to one. Thus, for large N, $(Y_i - \bar{Y})/\sigma$ is essentially a unit normal variable.

Again using the intermediate results of Equation 4-23, the covariance of $(Y_i - \bar{Y})/\sigma$ with $(Y_j - \bar{Y})/\sigma$ is $-1/N$ for $i \neq j$ – i.e., deviations from the sample mean are somewhat negatively correlated with each other. This negative correlation arises because both deviations are using the same $\bar{Y}$, which is calculated using both Y_i and Y_j. Consequently, $(Y_i - \bar{Y})/\sigma$ and $(Y_j - \bar{Y})/\sigma$ cannot be independent of one another. But this correlation disappears as N grows, so that the $(Y_i - \bar{Y})/\sigma$ terms in the expression for $(N-1)$ S^2/σ^2 become arbitrarily close to being independent as N becomes arbitrarily large. Thus, for sufficiently large N, $(N-1)S^2/\sigma^2$ becomes arbitrarily close to being the sum of the squares of N independently distributed unit normals – i.e., for large N, $(N-1)$ S^2/σ^2 is approximately distributed $\chi^2(N)$.

Remarkably – as noted above – a more sophisticated derivation shows that $(N-1)$ S^2/σ^2 is exactly distributed $\chi^2(N-1)$. As N becomes large, of course, the distinction between these two results becomes negligible.

The estimator S^2 has one other crucial property: it is distributed independently from $\bar{Y}$. This follows from the result (quoted in Equation 4-23) that the $\text{cov}(Y_i - \bar{Y}, \bar{Y})$ is zero for all i, which implies that

[8] See Johnston, J., 1984, *Econometric Methods*, McGraw-Hill: New York, pp. 165–7 and 180–2, for a proof.

1

the sample mean of the observations is uncorrelated with each individual observation's deviation from this sample mean. Both of these random variables are normally distributed; consequently, their uncorrelatedness implies that they are independent – i.e., completely unrelated – even though (because both terms in the covariance involve $\overline{Y}$) they look like they ought to be related. Because $\overline{Y}$ is unrelated to each of the deviations $Y_1 - \overline{Y}, \ldots, Y_N - \overline{Y}$, it must be unrelated to their squared values, and to the sum of their squared values, and (hence) to S^2.

It took some effort, but now we have an estimator of σ^2 which is distributed as a chi-squared variable independently from our estimator of μ. These results are used in the next section to obtain confidence intervals and hypothesis tests for μ when, as is typically the case, σ^2 is unknown.

4.6 INFERENCE RESULTS ON μ WHEN σ^2 IS UNKNOWN (AND INTRODUCING THE STUDENT'S t DISTRIBUTION)

In summary, we have now seen that assumption that the Y_i are distributed NIID $[\mu, \sigma^2]$ implies that

$$\frac{\overline{Y} - \mu}{\sqrt{\frac{\sigma^2}{N}}} \sim N[0,\,1] \qquad (4\text{-}28)$$

and that

$$\frac{(N-1)\,S^2}{\sigma^2} \sim \chi^2(N-1) \qquad \text{independently from } \overline{Y} \qquad (4\text{-}29)$$

As shown in earlier sections of this chapter, the unit normal statistic in Equation 4-28 can be used to construct a confidence interval for μ or to test a null hypothesis about μ when σ^2 is given. But those results are not very useful when σ^2 is unknown, because the resulting confidence interval endpoints and hypothesis testing statistics depend explicitly on σ^2. Now that we have an unbiased estimator of σ^2, however, it is natural to consider simply substituting S^2 for σ^2 and making the approximation

$$\frac{\overline{Y} - \mu}{\sqrt{\frac{S^2}{N}}} \approx \frac{\overline{Y} - \mu}{\sqrt{\frac{\sigma^2}{N}}} \qquad (4\text{-}30)$$

But this new random variable is no longer a unit normal because the estimator S^2 now in its denominator is itself a random variable. Consequently, one might expect the density function for this new random variable to have thicker tails than a unit normal – particularly for small samples, where S^2 will be a fairly "noisy" estimate of σ^2. It is impossible to proceed, however, without knowing the tail areas for the distribution of this new random variable.

Fortunately, because S^2 is a chi-squared variable independent of $\overline{Y}$, this new statistic has a simple and well-known distribution:

Student's *t* Distribution

If

$Z \sim N(0,\,1)$ and $Q \sim \chi^2(df)$ independently from Z

then

$$T \equiv \frac{Z}{\sqrt{\frac{Q}{df}}} \sim t(df) \qquad (4\text{-}31)$$

and

"T is distributed as Student's t with df degrees of freedom"

Letting

p 79

$$\frac{\dfrac{\overline{Y}-\mu}{\sqrt{\dfrac{\sigma^2}{N}}}}{\sqrt{\dfrac{\dfrac{(N-1)S^2}{\sigma^2}}{N-1}}} = \frac{\dfrac{\overline{Y}-\mu}{\sqrt{\dfrac{\sigma^2}{N}}}}{\sqrt{\dfrac{S^2}{\sigma^2}}} = \left[\frac{\overline{Y}-\mu}{\sqrt{\dfrac{\sigma^2}{N}}}\right]\sqrt{\frac{\sigma^2}{S^2}} = \frac{\overline{Y}-\mu}{\sqrt{\dfrac{\sigma^2}{N}\dfrac{S^2}{\sigma^2}}} = \frac{\overline{Y}-\mu}{\sqrt{\dfrac{S^2}{N}}} \sim t(N-1) \qquad (4\text{-}32)$$

The shape of the density function for a Student's t variable is very similar to that of a unit normal, except that (as expected) it has thicker tails when the number of degrees of freedom is small. For example, plotting both the unit normal density function and the density function for the Student's t distribution with five degrees of freedom yields Figure 4-4:

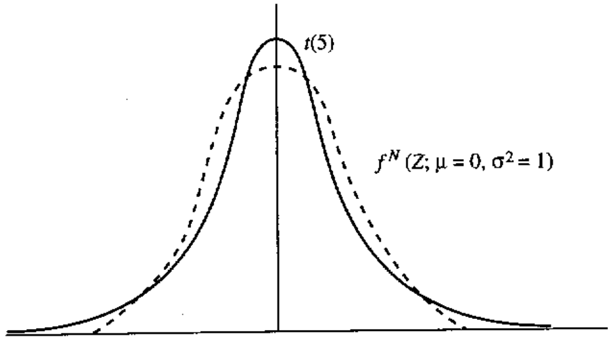

Figure 4-4 Comparison of Student's t with Unit Normal Distribution

Suppose now that there are $N = 26$ sample observations available. In that case, $(\overline{Y} - \mu)/\sqrt{S^2/26}$ is a pick from the Student's t distribution with 25 degrees of freedom. In the known-σ^2 case considered earlier, the analogous statistic – $(\overline{Y} - \mu)/\sqrt{\sigma^2/26}$ – was a unit normal random variable and we constructed confidence intervals for μ and tests of hypotheses concerning μ using critical points and tail areas of the unit normal distribution. To proceed in a a similar fashion here, we need to quantify the shape of the distribution of $(\overline{Y} - \mu)/\sqrt{S^2/26}$ by evaluating critical points and tail areas of the Student's t distribution with 25 degrees of freedom. As with the unit normal distribution considered in Chapter 2, these tail areas cannot be evaluated analytically, but they have been extensively tabulated and accurate numerical approximations for them are widely available.[9] For example, it is known that a random variable which is distributed $t(25)$ will exceed 2.06 with probability .025 – i.e., the area under the $t(25)$ density function to the right of 2.06 is .025 as in Figure 4-5:

[9] In particular, many spreadsheet programs provide functions for computing tail areas for the Student's t distribution – i.e., in Excel this function is called "tdist." Or you can program an excellent numerical approximation to this function yourself in just a few lines of code – see Press et al. (1986, p. 168) *Numerical Recipes*, Cambridge University Press. Or you can copy the program "ttail.exe" (available at www.wiley.com/college/ashley) and use it on any Windows-based computer.

distributed, and which all share the same population means

ut which all share the same population mean,

This set of assumptions is very similar to the assumption – that Y_i is distributed NIID[μ, σ^2] – made regarding Y_i in Chapter 3, so the discussion in this section will in some ways be quite similar. These assumptions with regard to the U_i are crucial to the derivation, in Chapter 6, of the properties of the least squares estimators of the model parameters α and β and to the construction, in Chapter 7, of hypothesis tests and confidence intervals for these parameters.[7] Moreover, because one portion or another of these assumptions about U_i is frequently violated by real economic data, we will revisit these assumptions in Chapters 10 through 13, when we deal with the practical issues of how to recognize and cope with such violations.[8] The task of the present section, in contrast, is only to expand a bit on what these assumptions with regard to U_i imply.

As in Chapter 3, the normality assumption on U_i is motivated in large part by appealing to the Central Limit Theorem (and its cousins), which imply that roughly equally-weighted sums of large numbers of random variables that are not too far from being identically and independently ~~distributed~~ will be approximately normally distributed. Recalling that U_i quantifies the impact on Y_i of all of the explanatory variables which have been omitted from the model, it is often reasonable to think of U_i as being such a weighted sum of a large number of more or less independent influences on Y_i. Alternatively, as you discovered in working Exercise 2-40, when it is more reasonable to think of U_i as being the *product* of a large number of such influences, then the distribution of its *logarithm* will be approximately normal. In such cases, these considerations suggest that reframing the model in terms of the logarithm of the original dependent variable is more likely to yield a model error term which is at least roughly normal. Either way, the actual validity of this normality assumption is essentially an empirical issue. For example, we will see in Chapter 10 that a histogram of the observed model fitting errors $\left(u_1^{\text{fit}} \ldots u_N^{\text{fit}}\right)$ will typically shed substantial light on the plausibility of this assumption for a particular data set. (These model fitting errors are obtained and discussed later in this chapter; in Chapter 10 they are shown to be reasonable estimates of the model errors, $U_1 \ldots U_N$, for large samples.)

We saw in Chapter 2 that the density function for a normally distributed random variable is completely determined by its mean and its variance. Consequently, the assumption that the U_i are identically distributed reduces in this case to an assumption that both the mean and variance of U_i are the same for every observation – in other words, for all values of i. This constant mean can be taken to be 0 at no loss of generality because any such mean value that U_i might otherwise have had can be incorporated into the parameter α of the Bivariate Regression Model. (Any sample variation in the mean of U_i which we are unable to model is considered to simply be part of U_i's random variation.) In contrast, any variation in the mean of U_i which **can** be modeled – as being due to sample variation in an observable variable, z_i, say – implies that the (supposedly constant) intercept parameter α is really a function of z_i; in that case, this variation should be modeled explicitly by including z_i in the model as an additional explanatory variable. This leads to the Multiple Regression Model taken up in Chapter 9.

Similarly, the assumption that the U_i are identically distributed also implies that the population variance of U_i – here denoted σ^2 – is the same for all N observations. Model errors which all have the same variance are said to be "homoscedastic." Contrastingly, in an obvious notation, the failure of this assumption is called "heteroscedasticity," in which case the variance of U_i varies with i: in other words, $\text{var}(U_i) = \sigma_i^2$. By observing where this assumption is used in the next few chapters, it will become apparent that the failure of the homoscedasticity assumption is consequential for both the efficiency of least squares parameter estimation and for the validity of the usual statistical machinery we use for constructing hypothesis tests and confidence intervals on α and β. Moreover – as we already saw in the

[7] For expositional simplicity, attention in Chapter 6 and 7 will mainly focus on the parameter β.

[8] At that point we will need to confront an unfortunate *dis*similarity between the assumptions on Y_i made in Chapter 3 and the assumptions on the model errors (U_i) in the Bivariate Regression Model made here: sample realizations of Y_i are directly observable, whereas sample realizations of U_i are not. In fact, as the "parallel universes" example shows, we would need to know the values of α and β in order to use our sample data (y_i and x_i) to calculate (observe) u_i, a sample realization of U_i.

$SSE(\hat{\alpha}^{guess}, \hat{\beta}^{guess})$:

Thus, our object is to choose $\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$ so as to minimize what is called the "sum of squared fitting errors" function, $\text{SSE}(\hat{\alpha}, \hat{\beta})$:

$$\text{SSE}(\hat{\alpha}^{\text{guess}}, \hat{\beta}^{\text{guess}}) \equiv \sum_{i=1}^{N} \left(u_i^{\text{fit}}\right)^2 = \sum_{i=1}^{N} \left(y_i - \hat{\alpha}^{\text{guess}} - \hat{\beta}^{\text{guess}} x_i\right)^2 \tag{5-22}$$

The function $\text{SSE}(\hat{\alpha}^{\text{guess}}, \hat{\beta}^{\text{guess}})$ clearly depends not only on $\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$ but also on the sample data, $y_1 \dots y_N$ and $x_1{}' \dots x_N$. This dependence on the sample data is suppressed in the notation so as to focus on how sum of the squared fitting errors depends on the parameter estimates $\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$.[12]

Suppose that we fix the value of $\hat{\beta}^{\text{guess}}$ at some given value, $\hat{\beta}_o^{\text{guess}}$, and examine how SSE $(\hat{\alpha}^{\text{guess}}, \hat{\beta}_o^{\text{guess}})$ depends on $\hat{\alpha}^{\text{guess}}$. If $\hat{\alpha}^{\text{guess}}$ is very small, then $\hat{\alpha}^{\text{guess}} + \hat{\beta}_o^{\text{guess}} x_i$ – the height of the fitted line – will lie far below all of the y_i values. This implies that $u_1^{\text{fit}} \dots u_N^{\text{fit}}$ will all be positive and large, leading to a large value for $\text{SSE}(\hat{\alpha}^{\text{guess}}, \hat{\beta}_o^{\text{guess}})$. Similarly, if $\hat{\alpha}^{\text{guess}}$ is very large, then the fitted line $\hat{\alpha}^{\text{guess}} + \hat{\beta}_o^{\text{guess}} x_i$ will lie far *above* all of the y_i values, implying that $u_1^{\text{fit}} \dots u_N^{\text{fit}}$ will all be very negative, and again leading to a large value for $\text{SSE}(\hat{\alpha}^{\text{guess}}, \hat{\beta}_o^{\text{guess}})$. Somewhere in between – where $\hat{\alpha}^{\text{guess}}$ is such that the fitted line is reasonably close to the dots in the scatterplot – some of $u_1^{\text{fit}} \dots u_N^{\text{fit}}$ values will be negative, some will be positive, and most will be small in magnitude, leading to a relatively small value for $\text{SSE}(\hat{\alpha}^{\text{guess}}, \hat{\beta}_o^{\text{guess}})$. Thus, a plot of $\text{SSE}(\hat{\alpha}^{\text{guess}}, \hat{\beta}_o^{\text{guess}})$ versus $\hat{\alpha}^{\text{guess}}$ will look like Figure 5-8:

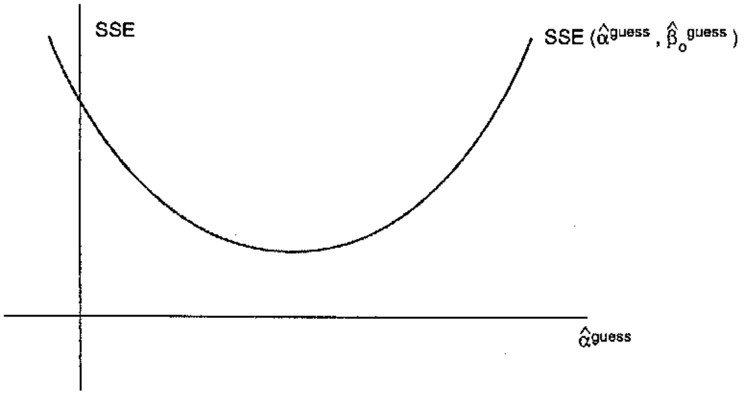

Figure 5-8 Sum of Squared Fitting Errors with $\hat{\beta}^{\text{guess}}$ Fixed

The least squares value for $\hat{\alpha}$ based on this particular value for $\hat{\beta}^{\text{guess}}$ is the value of $\hat{\alpha}^{\text{guess}}$ such that $\text{SSE}(\hat{\alpha}^{\text{guess}}, \hat{\beta}_o^{\text{guess}})$ is smallest. It is apparent from this diagram that the slope of the SSE $(\hat{\alpha}^{\text{guess}}, \hat{\beta}_o^{\text{guess}})$ curve is negative for values of $\hat{\alpha}^{\text{guess}}$ smaller than this and positive for values of

[12] It is preferable to call this function "the sum of squared fitting errors" so as to keep in mind that this function is summing up the observed squared *fitting* errors rather than squared realizations of the model errors, $U_1 \dots U_N$. Thus, "SSFE" would be better notation for this function, but "SSE" is the standard nomenclature.

$\hat{\alpha}^{guess}$ larger than this. Therefore we can characterize the least squares value for $\hat{\alpha}$ as that value of $\hat{\alpha}^{guess}$ at which the derivative of SSE $(\hat{\alpha}^{guess}, \hat{\beta}_o^{guess})$ with respect to $\hat{\alpha}^{guess}$ is zero.

Similarly, if we fix the value of $\hat{\alpha}^{guess}$ at some particular value $\hat{\alpha}_o^{guess}$ and examine how SSE $(\hat{\alpha}_o^{guess}, \hat{\beta}^{guess})$ depends on $\hat{\beta}^{guess}$, we find that the value of $\text{SSE}(\hat{\alpha}_o^{guess}, \hat{\beta}^{guess})$ will be large for extreme values of $\hat{\beta}^{guess}$ because these lead to fitted lines whose slopes are so extreme that the fitted line lies near the observed scatter of data only for a few values of the explanatory variable. Thus, a plot of $\text{SSE}(\hat{\alpha}_o^{guess}, \hat{\beta}^{guess})$ versus $\hat{\beta}^{guess}$ will look like Figure 5-9:

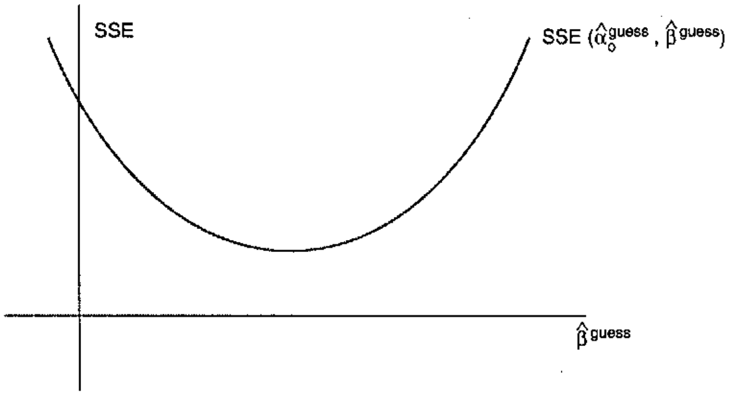

Figure 5-9 Sum of Squared Fitting Errors with $\hat{\alpha}^{guess}$ Fixed

Consequently, we can similarly characterize the least squares value for $\hat{\beta}$ as that value of $\hat{\beta}^{guess}$ at which the derivative of $\text{SSE}(\hat{\alpha}_o^{guess}, \hat{\beta}^{guess})$ with respect to $\hat{\beta}^{guess}$ is zero.

Thus, the least squares estimates of α and β must jointly satisfy the conditions

should be → $\frac{\partial SSE(\hat{\alpha}^{guess}, \hat{\beta}^{guess})}{\partial \hat{\alpha}^{guess}}$

$$\partial \frac{SSE(\hat{\alpha}^{guess}, \hat{\beta}^{guess})}{\partial \hat{\alpha}^{guess}} = 0$$
$$\frac{\partial SSE(\hat{\alpha}^{guess}, \hat{\beta}^{guess})}{\partial \hat{\beta}^{guess}} = 0 \qquad (5\text{-}23)$$

As noted earlier in this section, these estimates are customarily called the "ordinary least squares" estimates of α and β, or $\hat{\alpha}^*_{ols}$ and $\hat{\beta}^*_{ols}$, where the term "ordinary" merely refers to the fact that the Bivariate Regression Model is linear in the unknown parameters, α and β.[13]

The asterisks on $\hat{\alpha}^*_{ols}$ and $\hat{\beta}^*_{ols}$ are used to notationally reflect the fact that these estimates are fixed numbers determined by the observed sample data: $y_1 \dots y_N$ and $x_1 \dots x_N$. Later on (in

[13] This terminology using the word "ordinary" is mentioned and used here only because it is the standard nomenclature. In fact, to unclutter the notation, the "ols" subscript will be dropped altogether in Chapters 6 through 8, where only least squares estimators are considered.

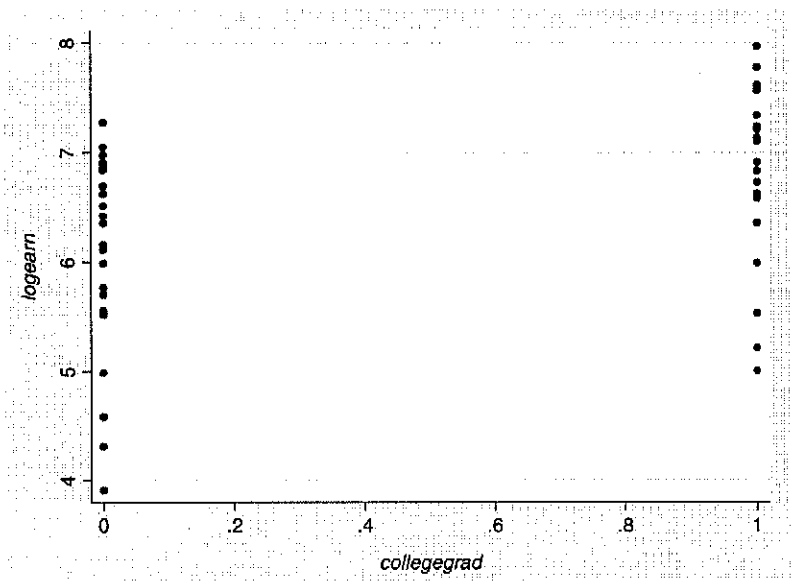

Figure 5-14 Scatterplot of the Logarithm of Weekly Income versus the College Graduation Dummy Variable

Note that the apparent variance of the model error term is now substantially more similar for the two values of *collegegrad*$_i$ than in the previous scatterplots.[19]

Note also, however, that the case for the proposition that expected earnings are higher for college graduates now seems a bit less compelling. Comparing these two scatterplots in this way is misleading, however, because the slope in the previous diagram corresponds to the expected increment in earnings (*PTERNWA*) for having graduated from college whereas the slope in this diagram corresponds to the expected increment in the *logarithm* of earnings for having graduated from college.

Estimating the parameters α_1 and β_1 in the model for $LOGEARN_i$ using the formulas for the least squares estimates derived earlier in this chapter yields

$$\frac{\sum_{i=1}^{50}\left(collegegrad_i - \overline{collegegrad}\right)\left(logearn_i - \overline{logearn}\right)}{\sum_{i=1}^{50}\left(collegegrad_i - \overline{collegegrad}\right)^2} = 0.873 \qquad (5\text{-}45)$$

and

$$\overline{logearn} - .873\,\overline{collegegrad} = 6.022 \qquad (5\text{-}46)$$

[19] Using the F test given at the end of Chapter 2, which assumes that each sample is normally identically and independently distributed, the p-value for rejecting the null hypothesis of equal variances is .57, so this null hypothesis cannot be rejected.

Y_i

In Chapter 6 we found that the least squares estimator, $\hat{\beta} = \sum_{i=1}^{N} w_i^{ols} y_i$ is distributed

$$\hat{\beta} \sim N\left[\beta, \frac{\sigma^2}{\sum_{j=1}^{N}(x_j - \bar{x})^2}\right] \tag{7-3}$$

under these assumptions.

Clearly, our uncertainty as the actual value of β depends on the sampling variance of $\hat{\beta}$, or $\sigma^2 / \sum_{j=1}^{N}(x_j - \bar{x})^2$. For expositional clarity it is initially assumed here that σ^2, the variance of the model errors ($U_1 \ldots U_N$), is given. Once confidence intervals and hypothesis tests for β are obtained for this special case, the results are extended to where, as is always the case in practice, σ^2 must be estimated from the sample data.

The basic ideas underlying the development of practical confidence intervals and hypothesis tests for β in the Bivariate Regression Model are essentially identical to those used in Chapter 4 to obtain analogous results for μ where $Y_i \sim \text{NIID}(\mu, \sigma^2)$. Consequently, the exposition given here begins (as it did in Chapter 4) by using the sampling distribution of the relevant estimator to obtain a statistic with a known distribution.

7.2 A STATISTIC FOR β WITH A KNOWN DISTRIBUTION

Here, the relevant estimator, of course, is $\hat{\beta}$. The expression for its sampling distribution given above proved very useful in Chapter 6 for deriving the properties of $\hat{\beta}$, but it is more convenient to work with the unit normal distribution for inference purposes. Subtracting the mean of $\hat{\beta}$ and dividing by the square root of its variance standardizes $\hat{\beta}$ to zero mean and unit variance, yielding

$$\frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / \sum_{j=1}^{N}(x_j - \bar{x})^2}} \sim N[0, 1] \tag{7-4}$$

7.3 A 95% CONFIDENCE INTERVAL FOR β WITH σ^2 GIVEN

Recall from Chapter 2 that by definition, a unit normal variate exceeds its 2½% critical point ($z^c_{.025}$, or 1.96) with probability .025 and is also less than its 97½% critical point ($z^c_{.975}$, or −1.96) with probability 1 − .975 or .025. Thus, a unit normal variate lies in between these two critical points with probability .95, as illustrated in Figure 7-2.

Therefore, with probability .95,

$$z^c_{.975} \le \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / \sum_{j=1}^{N}(x_j - \bar{x})^2}} \le z^c_{.025} \tag{7-5}$$

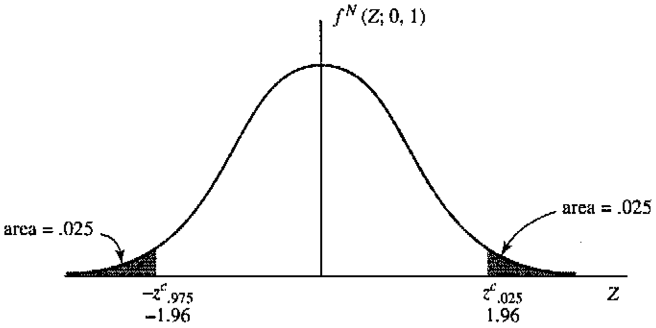

Figure 7-2 Unit Normal Density Function with Critical Points Illustrated

or, equivalently,

$$-1.96 \;\leq\; \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / \sum\limits_{j=1}^{N} (x_j - \bar{x})^2}} \;\leq\; 1.96 \tag{7-6}$$

It follows that, also with probability .95,

$$
\begin{gathered}
-1.96\sqrt{\sigma^2 / \sum_{j=1}^{N} (x_j - \bar{x})^2} \;\leq\; \hat{\beta} - \beta \;\leq\; 1.96\sqrt{\sigma^2 / \sum_{j=1}^{N} (xX_j - \bar{x})^2} \\
-\hat{\beta} - 1.96\sqrt{\sigma^2 / \sum_{j=1}^{N} (x_j - \bar{x})^2} \;\leq\; -\beta \;\leq\; -\hat{\beta} + 1.96\sqrt{\sigma^2 / \sum_{j=1}^{N} (x_j - \bar{x})^2} \\
\hat{\beta} + 1.96\sqrt{\sigma^2 / \sum_{j=1}^{N} (x_j - \bar{x})^2} \;\geq\; \beta \;\geq\; \hat{\beta} - 1.96\sqrt{\sigma^2 / \sum_{j=1}^{N} (x_j - \bar{x})^2}
\end{gathered}
\tag{7-7}
$$

x_j

like

so that the interval

$$\left[\hat{\beta} - 1.96\sqrt{\sigma^2 / \sum_{j=1}^{N} (x_j - \bar{x})^2},\;\; \hat{\beta} + 1.96\sqrt{\sigma^2 / \sum_{j=1}^{N} (x_j - \bar{x})^2}\right] \tag{7-8}$$

contains β with probability .95 – i.e., it is a 95% confidence interval for β. Of course, if any of the assumptions of the Bivariate Regression Model are invalid – i.e., the x_i are not fixed or the U_i are not distributed NIID[0, σ^2] – then this interval will *not* in general contain β with probability .95.

The derivation of a 99% confidence interval is essentially identical, only using the ½% critical point, $z^c_{.005}$, or 2.57, instead of $z^c_{.025}$. Clearly, this interval must be wider – by a factor of 2.57/1.96 – so as to contain β with the specified higher probability.

Note also, that the unbiasedness of $\hat{\beta}$ was necessary in order to standardize $\hat{\beta}$ to zero mean in Section 7.2, and that the derivation given above makes it plain that the width of any confidence interval for β based on $\hat{\beta}$ is proportional to the square root of the sampling variance of $\hat{\beta}$. Indeed, this put minus sign here

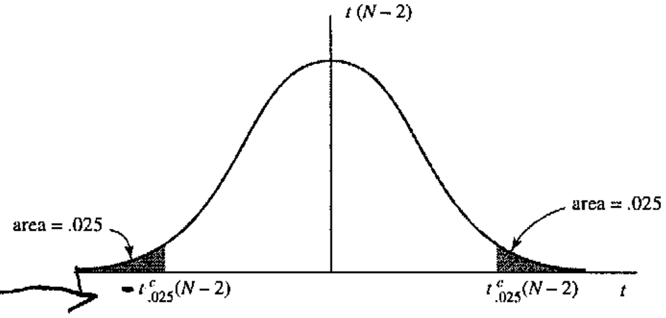

Figure 7-4 Student's *t* Density Function with Critical Points Illustrated

with probability .95. Like the unit normal distribution – which it resembles, except for having a sharper peak and thicker tails – the Student's *t* distribution is symmetric around zero; consequently, $t^c_{.975}(N-2) = -t^c_{.025}(N-2)$, so that

$$-t^c_{.025}(N-2) \le \frac{\hat{\beta}-\beta}{\sqrt{S^2/\sum\limits_{j=1}^{N}(x_j-\bar{x})^2}} \le t^c_{.025}(N-2) \tag{7-29}$$

with probability .95. Consequently, also with probability .95,

$$\begin{aligned}
-t^c_{.025}(N-2)\sqrt{S^2/\sum_{j=1}^{N}(x_j-\bar{x})^2} &\le \hat{\beta}-\beta \le t^c_{.025}(N-2)\sqrt{S^2/\sum_{j=1}^{N}(x_j-\bar{x})^2} \\
-\hat{\beta}-t^c_{.025}(N-2)\sqrt{S^2/\sum_{j=1}^{N}(x_j-\bar{x})^2} &\le -\beta \le -\hat{\beta}+t^c_{.025}(N-2)\sqrt{S^2/\sum_{j=1}^{N}(x_j-\bar{x})^2} \\
\hat{\beta}+t^c_{.025}(N-2)\sqrt{S^2/\sum_{j=1}^{N}(x_j-\bar{x})^2} &\ge \beta \ge \hat{\beta}-t^c_{.025}(N-2)\sqrt{S^2/\sum_{j=1}^{N}(x_j-\bar{x})^2}
\end{aligned} \tag{7-30}$$

so that

$$\beta = \hat{\beta} \pm t^c_{.025}(N-2)\sqrt{S^2/\sum_{j=1}^{N}(x_j-\bar{x})^2} \tag{7-31}$$

with probability .95. A sample realization of this confidence interval is thus

$$\beta = \hat{\beta}^* \pm t^c_{.025}(N-2)\sqrt{s^2/\sum_{j=1}^{N}(x_j-\bar{x})^2} \tag{7-32}$$

/lodel:

ate Regression Model using
ıe sampling distributions of
, show that the least squares
able properties of unbiased-
thing really practical with an
ion of $\hat{\beta}$ to derive confidence

ate Regression Model in this

: variation in x_i "explain" the

ıting the model? How can the
'als for predicted values of Y_i?

ial model validation issue: "Is

uoted goodness-of-fit measure
er job of measuring the overall
Multiple Regression Model is

ıomic time-series data or as in
:ions of Y_i is a highly important
netric models is to make such
ıe or more currently observed
ıarily devoted to this topic. Even
play an important role in model
ediction in regression modeling.
in the context of macroeconomic
economic sales forecasting). The
ion needs to be made. Or x_i might

be unknown, because (for example) it is a policy variable (such as a target interest rate) whose value is still under consideration. In that case, Y_i forecasts conditional on various likely choices for x_i often form whatever rational basis there can be for the analysis of policy. In either case, it is essential to be able to place error bars around whatever predictions are made.

Each of these topics is now considered in turn.

8.2 QUANTIFYING HOW WELL THE MODEL FITS THE DATA

In view of the fact that the estimates of the parameters α and β in the Bivariate Model are chosen so as to best fit a straight line to the sample data, it is clearly relevant to quantify how well the resulting estimated model actually does fit the data. In principle, such a measure will be useful both in assessing the value of a particular model and in comparing one proposed model for the data to another.

The most common goodness-of-fit measure is called R^2. This nomenclature derives from the result (obtained below) that this measure is numerically equal to the square of the sample correlation between the data on the explanatory and the dependent variables in the Bivariate Regression Model; this goodness-of-fit measure bears several other, probably more useful, interpretations as well, however.

On the other hand, it is fair to say that people pay far too much attention to R^2. For one thing, R^2 turns out to not be very useful for comparing regression models with differing numbers of explanatory variables; consequently, we will need to return to this topic in Chapter 9. Beyond that – as will become apparent below – the interpretation of R^2 is both subjective and problematic. Still, some form of R^2 is quoted for almost every estimated regression model. For example, the estimated model for the application analyzed in the previous chapters would ordinarily have been expressed

$$\begin{array}{lll} logearn_i = \underset{(.166)}{6.022} + \underset{(.234)}{0.873}\ collegegrad_i + w_i^{\text{fit}} & \begin{array}{l} R^2 = .224 \\ s^2 = .686 \end{array} & \text{(8-1)} \end{array}$$

Consequently, it is important to understand how R^2 is calculated and what it does (and doesn't) mean.

Fundamentally, we seek to model the variation in $y_1 \ldots y_N$, the realized values of the dependent variable, across the sample. R^2 quantifies the proportion of this sample variation in Y_i that is "captured" by the $\hat{\alpha} + \hat{\beta} x_i$ part of the fitted Bivariate Regression Model. This portion of the fitted model is an estimate of the mean of Y_i, conditional on the observed value of x_i:

$$Y_i = \alpha + \beta x_i + U_i = E[Y_i|x_i] + U_i \tag{8-2}$$

because the presence of the intercept term (α) implies that $E[U_i] = 0$. The sample variation in the realizations $y_1 \ldots y_N$ thus arises from two sources, corresponding to the two terms on the right-hand side of this equation. More explicitly, the y_i observations vary across the sample because:

1. The x_i vary with i, causing $E[Y_i|x_i] = \alpha + \beta x_i$ to vary.
2. The U_i are random variables, so that Y_i is not actually equal to $E[Y_i|x_i]$.

The amount of sample variation in the y_i can be measured by

$$\text{SST} = \text{sum of squares total} \equiv \sum_{i=1}^{N} (y_i - \bar{y})^2 = \sum_{i=1}^{N} \left(Y_i - \hat{\alpha}^* - \hat{\beta}^* \bar{x}\right)^2 \tag{8-3}$$

SST *looks* like a sample realization of $(N-1)$ times an estimator of the population variance of Y_i. But it really isn't, because the Y_i are not identically distributed: $E[Y_i] = \alpha + \beta x_i$ is different for each different value of x_i, whereas this estimator is taking $E[Y_i]$ to be a constant which could reasonably be estimated by the estimator $\bar{Y}$. SST can still, however, be sensibly used as a *descriptive statistic* summarizing how unequal to one another the $y_1 \ldots y_N$ are.

If we quantify the sample variation in the y_i using SST and estimate β using the least squares estimator, then the sample variation in the y_i splits neatly into two parts, one corresponding to each of the two sources listed above. To see this, first recall from Chapter 5 that the least squares estimators $\hat{\alpha}^*$ and $\hat{\beta}^*$ must satisfy the equations

$$\sum_{i=1}^{N}\left\{y_i - \hat{\alpha}^* - \hat{\beta}^* x_i\right\} = \sum_{i=1}^{N} u_i^{\text{fit}} = N\overline{y} - N\hat{\alpha}^* - N\hat{\beta}^*\overline{x} = 0 \tag{8-4}$$

Y_i is lower case →

$$\sum_{i=1}^{N} x_i\left\{\overset{Y_i}{\cancel{y_i}} - \hat{\alpha}^* - \hat{\beta}^* x_i\right\} = \sum_{i=1}^{N} x_i u_i^{\text{fit}} = 0 \tag{8-5}$$

in order to make the partial derivatives of SSE ($\hat{\alpha}^{\text{guess}}$, $\hat{\beta}^{\text{guess}}$) with respect to both $\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$ equal to zero. The first of these two equations implies that $\overline{y} - \hat{\alpha}^* - \hat{\beta}^*\overline{x} = 0$; substituting this and the equation for the fitted model $(y_i = \hat{\alpha}^* + \hat{\beta}^* x_i + u_i^{\text{fit}})$ into the definition of SST given above yields

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{N}(y_i - \overline{y})^2 \\
&= \sum_{i=1}^{N}\left\{\left(\hat{\alpha}^* + \hat{\beta}^* x_i + u_i^{\text{fit}}\right) - \left(\hat{\alpha}^* + \hat{\beta}^*\overline{x}\right)\right\}^2 \\
&= \sum_{i=1}^{N}\left\{\hat{\beta}^*[x_i - \overline{x}] + u_i^{\text{fit}}\right\}^2 \\
&= \sum_{i=1}^{N}\left\{\left(\hat{\beta}^*\right)^2[x_i - \overline{x}]^2 + 2\hat{\beta}^*[x_i - \overline{x}]u_i^{\text{fit}} + \left(u_i^{\text{fit}}\right)^2\right\}
\end{aligned} \tag{8-6}$$

so that

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{N}\left\{\left(\hat{\beta}^*\right)^2[x_i - \overline{x}]^2 + 2\hat{\beta}^*[x_i - \overline{x}]u_i^{\text{fit}} + \left(u_i^{\text{fit}}\right)^2\right\} \\
&= \sum_{i=1}^{N}\left(\hat{\beta}^*\right)^2[x_i - \overline{x}]^2 + 2\hat{\beta}^*\sum_{i=1}^{N}x_i u_i^{\text{fit}} - 2\hat{\beta}^*\overline{x}\sum_{i=1}^{N}u_i^{\text{fit}} + \sum_{i=1}^{N}\left(u_i^{\text{fit}}\right)^2 \\
&= \left(\hat{\beta}^*\right)^2\sum_{i=1}^{N}[x_i - \overline{x}]^2 + 0 + 0 + \text{SSE}(\hat{\alpha}^*, \hat{\beta}^*)
\end{aligned} \tag{8-7}$$

$\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$ as minimizing SSE($\hat{\alpha}^{\text{guess}}$, $\hat{\beta}^{\text{guess}}$)

If we quantify the sample variation in the y_i using SST and estimate β using the least squares estimator, then the sample variation in the y_i splits neatly into two parts, one corresponding to each of the two sources listed above. To see this, first recall from Chapter 5 that the least squares estimators $\hat{\alpha}^*$ and $\hat{\beta}^*$ must satisfy the equations

↑ Equations 5-24 and 5-27 in

$$\sum_{i=1}^{N}\left\{y_i - \hat{\alpha}^* - \hat{\beta}^* x_i\right\} = \sum_{i=1}^{N} u_i^{\text{fit}} = N\bar{y} - N\hat{\alpha}^* - N\hat{\beta}^*\bar{x} = 0 \tag{8-4}$$

$$\sum_{i=1}^{N} x_i\left\{u_i - \hat{\alpha}^* - \hat{\beta}^* x_i\right\} = \sum_{i=1}^{N} x_i u_i^{\text{fit}} = 0 \tag{8-5}$$

in order to make the partial derivatives of SSE $(\hat{\alpha}^{\text{guess}}, \hat{\beta}^{\text{guess}})$ with respect to both $\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$ equal to zero. The first of these two equations implies that $\bar{y} - \hat{\alpha}^* - \hat{\beta}^*\bar{x} = 0$; substituting this and the equation for the fitted model $\left(y_i = \hat{\alpha}^* + \hat{\beta}^* x_i + u_i^{\text{fit}}\right)$ into the definition of SST given above yields

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{N}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{N}\left\{\left(\hat{\alpha}^* + \hat{\beta}^* x_i + u_i^{\text{fit}}\right) - \left(\hat{\alpha}^* + \hat{\beta}^*\bar{x}\right)\right\}^2 \\
&= \sum_{i=1}^{N}\left\{\hat{\beta}^*[x_i - \bar{x}] + u_i^{\text{fit}}\right\}^2 \\
&= \sum_{i=1}^{N}\left\{\left(\hat{\beta}^*\right)^2[x_i - \bar{x}]^2 + 2\hat{\beta}^*[x_i - \bar{x}]u_i^{\text{fit}} + \left(u_i^{\text{fit}}\right)^2\right\}
\end{aligned} \tag{8-6}$$

so that

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{N}\left\{\left(\hat{\beta}^*\right)^2[x_i - \bar{x}]^2 + 2\hat{\beta}^*[x_i - \bar{x}]u_i^{\text{fit}} + \left(u_i^{\text{fit}}\right)^2\right\} \\
&= \sum_{i=1}^{N}\left(\hat{\beta}^*\right)^2[x_i - \bar{x}]^2 + 2\hat{\beta}^*\sum_{i=1}^{N}x_i u_i^{\text{fit}} - 2\hat{\beta}^*\bar{x}\sum_{i=1}^{N}u_i^{\text{fit}} + \sum_{i=1}^{N}\left(u_i^{\text{fit}}\right)^2 \\
&= \left(\hat{\beta}^*\right)^2\sum_{i=1}^{N}[x_i - \bar{x}]^2 + 0 + 0 + \text{SSE}(\hat{\alpha}^*, \hat{\beta}^*)
\end{aligned} \tag{8-7}$$

Thus, using the two conditions characterizing $\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$ as minimizing SSE$(\hat{\alpha}^{\text{guess}}, \hat{\beta}^{\text{guess}})$, SST splits cleanly into a part – SSE$(\hat{\alpha}^*, \hat{\beta}^*)$ – which is clearly due to the imperfect fit of the model to the sample data and a part – $\left(\hat{\beta}^*\right)^2 \sum_{i=1}^{N} [x_i - \bar{x}]^2$ – which is clearly due to the size of $\hat{\beta}^*$ and to the degree to which x_i varies across the sample. Because SSE$(\hat{\alpha}^*, \hat{\beta}^*)$ is the portion of SST which the

mate β using the least squares
arts, one corresponding to each
hapter 5 that the least squares

$$- N\hat{\beta}^*\bar{x} = 0 \qquad (8\text{-}4)$$

$$= 0 \qquad (8\text{-}5)$$

th respect to both $\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$
$- \hat{\beta}^*\bar{x} = 0$; substituting this and
he definition of SST given above

$$\cdot\, \hat{\beta}^*\bar{x}\Big)\Big\}^2$$

$$(8\text{-}6)$$

$$\bar{x}]u_i^{\text{fit}} + \left(u_i^{\text{fit}}\right)^2\Big\}$$

$$+ \left(u_i^{\text{fit}}\right)^2\Big\}$$

$$2\hat{\beta}^*\bar{x}\sum_{i=1}^{N} u_i^{\text{fit}} + \sum_{i=1}^{N}\left(u_i^{\text{fit}}\right)^2 \qquad (8\text{-}7)$$

$$\hat{\alpha}^*, \hat{\beta}^*)$$

${}^{\text{guess}}$ as minimizing $\text{SSE}(\hat{\alpha}^{\text{guess}}, \hat{\beta}^{\text{guess}})$,
ly due to the imperfect fit of the model to
; clearly due to the size of $\hat{\beta}^*$ and to the
$(\hat{\alpha}^*, \hat{\beta}^*)$ is the portion of SST which the

p181

sample variation in $\hat{\alpha}^* - \hat{\beta}^* x_i$ does *not* reproduce, $\left(\hat{\beta}^*\right)^2 \sum_{i=1}^{N} [x_i - \bar{x}]^2$ must evidently be the po
of SST which the sample variation in $\hat{\alpha}^* - \hat{\beta}^* x_i$ *does* reproduce.[1]

Therefore R^2 is defined as

$$R^2 \equiv \frac{\left(\hat{\beta}^*\right)^2 \sum_{i=1}^{N} [x_i - \bar{x}]^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = \frac{\left(\hat{\beta}^*\right)^2 \sum_{i=1}^{N} [x_i - \bar{x}]^2}{\text{SST}}$$

and interpreted as the fraction of the sample variation in the dependent variable which is "explai
by the fitted model. Viewing SST/$(N-1)$ as the sample variance of $y_1 \dots y_N$ – i.e., ignoring the
(noted above) that this statistic is really characterizing the dispersion of $y_1 \dots y_N$ around $\bar{y}$ rather
around an estimate (e.g., $\hat{\alpha}^* + \hat{\beta}^* x_i$) of their actual means – this interpretation is often verl
expressed by identifying R^2 as "the fraction of the variance of y_i explained by the fitted mo

Next, note that the expression for R^2 in Equation 8-8 is identical to the square of the expressio
r_{xy}, the sample correlation between the data on the explanatory variable ($x_1 \dots x_N$) and the obse
realizations of the dependent variable ($y_1 \dots y_N$), obtained in Section 5.6. Thus, R^2 also equal
square of the sample correlation between y_i and any linear function of x_i, such as $\hat{\alpha}^*_{\text{OLS}} + \hat{\beta}^*_{\text{O}}$
Hence R^2 bears a second interpretation as a consistent estimate of the squared correlation betv
the dependent variable in the regression model and the model's predicted value for Y_i, as de
and analyzed in Section 8.4. And this, of course, is the motivation for calling this statistic "R^2" i
first place.

R^2 has a third useful interpretation as a goodness-of-fit measure. Substituting SST – SSE($\hat{\alpha}^*$,
for $\left(\hat{\beta}^*\right)^2 \sum_{i=1}^{N} [x_i - \bar{x}]^2$ in the above expression yields (from Equation 8-7)

$$R^2 = \frac{\left(\hat{\beta}^*\right)^2 \sum_{i=1}^{N} [x_i - \bar{x}]^2}{\text{SST}} = \frac{\text{SST} - \text{SSE}(\hat{\alpha}^*, \hat{\beta}^*)}{\text{SST}} = 1 - \frac{\text{SSE}(\hat{\alpha}^*, \hat{\beta}^*)}{\text{SST}}$$

What can this set of equations tell us about the size of R^2? If $(\hat{\beta}^*)^2$ is close to zero, then the estin
model is a poor fit to the data – the sample variation in x_i is apparently irrelevant to the sample vari
in y_i – and R^2 is close to zero. In contrast, a model which fits the data very well (and hence for w
SSE($\hat{\alpha}^*$, $\hat{\beta}^*$) is very small) will have R^2 close to one. Thus R^2 lies in the interval [0, 1] and
reasonable to interpret it as a goodness-of-fit measure.

At this point, however, a good deal of subjectiveness creeps into the discourse – how high doe
need to be in order to characterize the fit as "good"? In practice, this depends on both the contex
on one's tastes. Generally speaking, household survey data contains so much noise in
observation that most analysts are quite happy with an R^2 value of around .20. In cont
most people would consider a regression equation involving aggregated (e.g., macroeconomic)
to be a poor fit to the data with an R^2 less than, say, .50.

It must also be pointed out that it is not always a good idea to use R^2 to quantify the degre
relationship between two time-series variables. For example, suppose that Y_t and x_t are
unrelated time-series, each of whose sample behavior is dominated by an upward time tre
aggregate annual U.S. consumption spending and the population of Madagascar. An estin

[1] that the assumptions about the model error term played no role here; this decomposition of the sample variat
is just a consequence of the fact that $\hat{\beta}$ is the least squares estimator of β.

9

The Multiple Regression Model

9.1 INTRODUCTION

Our coverage of the basic Bivariate Model is now complete. Analysis of this model allowed us to examine regression parameter estimation and inference in the simplest possible setting. But the modeling of real-world economic data sets – and even, as we saw in Chapter 8, beginning to check the assumptions of the Bivariate Model – requires that we broaden the modeling framework to include multiple explanatory variables in the regression model. This broader framework is called the Multiple Regression Model and is the topic of this chapter.

None of our work on the Bivariate Model will go to waste, however. In particular, because a complete analysis of multiple regression requires a background in matrix algebra which is beyond the scope of this book, many of the results on the Multiple Regression Model – for example, the BLUness of the least squares parameter estimates – will merely be stated here and motivated by reference to the analogous results for the Bivariate Regression Model. And you will find that a number of other derivations in this chapter – those, for example, of the sampling distributions of the parameter estimates and of the inference results based on these estimates – are quite similar to the analogous derivations in the context of the Bivariate Regression Model.

Other issues arise here which could not occur in the context of the Bivariate Regression Model. We must now deal with the issues associated with overelaborate and underelaborate models. Also unique to the Multiple Regression Model is the problem of multicollinearity, where the sample variation in one explanatory variable is uncomfortably similar to that of another.

The Multiple Regression Model is vastly more powerful than the Bivariate Regression Model; this chapter closes with several examples illustrating its applicability.

9.2 THE MULTIPLE REGRESSION MODEL

The Multiple Regression Model is a straightforward generalization of the Bivariate Model to include multiple explanatory variables:

The Multiple Regression Model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \cdots + \beta_k x_{i,k} + U_i \qquad i = 1 \ldots N$$
$$x_{i,1} = 1 \quad \text{for} \quad i = 1 \ldots N$$
$$x_{i,j} \text{ is “fixed in repeated samples” for } i = 1 \ldots N \text{ and for } j = 2 \ldots k \qquad (9\text{-}1)$$
$$U_i \sim \text{NIID}[0, \sigma^2]$$

Equivalently,

Y_i is independently (but not identically) distributed $N[\beta_1 x_{i,1} \ldots \beta_k x_{i,k}, \sigma^2]$

$x_{i,1} + \ldots + \beta_k x_{i,k}$

in otherwords put a plus sign both before and after the "..."

abrupt, then one might do better to model β_3 as varying linearly with x_{i3}, in which case one is right back at including $(x_{i3})^2$ in the model.[7]

The assumption that the explanatory variables are "fixed in repeated samples" was discussed in Section 5.3. This assumption is unfortunately not so easily checked; nor are violations of it so easily dealt with. But dealt with they can be, using an econometric technique known as "instrumental variables estimation." A useful consideration of this very important topic must be delayed until Chapter 12, however, because it requires the additional material on probability theory covered in Chapter 11.[8]

Finally, the last piece of the *form* of the model specification is a choice as to the form with which the dependent variable enters the model. For example, will Y_i be the average income for the ith country in the sample, or would the dependent variable in the model be more appropriately specified as the logarithm of the ith country's income, or perhaps as per capita income for the ith country? This choice turns out to interact strongly with a consideration of the *statistical* assumptions on the error term – in particular on the assumptions that the model error term, U_i, is normally distributed and homoscedastic. The remainder of this chapter focuses on precisely these two assumptions: how to sensibly test whether or not they are (to a reasonable degree) satisfied and how to respond to sample indications that they are seriously violated.[9]

10.2 THE FITTING ERRORS AS LARGE-SAMPLE ESTIMATES OF THE MODEL ERRORS, $U_1 \ldots U_N$

There is a simple relationship between the fitting errors and the model errors in the Bivariate Regression Model; in that model the fitting errors are

$$
\begin{aligned}
U_i^{\text{fit}} &= Y_i - [\hat{\alpha} + \hat{\beta} x_i] \\
&= [\alpha + \beta x_i + U_i] - [\hat{\alpha} + \hat{\beta} x_i] \\
&= U_i - [\hat{\alpha} - \alpha] - [\hat{\beta} x_i + \beta x_i] \\
&= U_i - [\hat{\alpha} - \alpha] - x_i [\hat{\beta} + \beta]
\end{aligned}
\qquad (10\text{-}2)
$$

these should be Minus signs

where the model has been substituted in for Y_i and the terms rearranged.[10] Recall from Equation 6-16 that

$$\hat{\beta} = \beta + \sum_{i=1}^{N} w_i^{\text{ols}} U_i \qquad (10\text{-}3)$$

[7] With time-series data, where the data set is ordinarily sorted in increasing time-order and the usual issue is whether (and in what way) a coefficient such as β_3 varies over time, many different specifications have been proposed for the time-evolution of a regression parameter. The simplest of these are the abrupt-change specifications discussed above; another is to assume that β_3 is a linear or quadratic function of time (observation number) – this suggests an alternative specification examined in Exercise 10-1. Ashley (1984, *Economic Inquiry XXII*, 253–67) examined these (and a number of substantially more sophisticated alternatives) and found that one is usually just as well off using the straightforward dummy-variable approach described above.

[8] See also the discussion of this topic at the end of Active Learning Exercise 10b (available at www.wiley.com/college/ashley).

[9] The material in Chapters 5 and 9 showed that inclusion of an intercept in the regression model automatically implies that $E[U_i]$ is zero – and takes one a considerable way toward unbiased parameter estimates. Thus, in general no sensible person will omit the intercept; consequently, this portion of the assumptions on the model error ordinarily need not be checked at all. The non-autocorrelation assumption – that corr(U_i, U_j) equals zero for all $i \neq j$ – is frequently quite problematic for models using time-series data, but the analysis of these models requires the additional probability theory material to be covered in Chapter 11. Consequently, coverage of diagnostic checking of this portion of the assumptions on the model errors is delayed until Chapters 13 and 14.

[10] This relationship was previously derived in the "Estimating σ^2" section of Chapter 7. The relationship is almost as simple in the Multiple Regression Model – see Exercise 10-3.

he other hand, these results yield no guidance as to how one might respecify one's model so as btain a better model or more efficient (closer to BLU) parameter estimates: the results ribed below only "fix" the standard error estimates, which heteroscedasticity otherwise orts. It is also well to emphasize at the outset that these "robust" standard error estimates – ed "robust" because they are robust to the presence of heteroscedasticity – are only valid for e samples.[35]

practice, robust standard error estimates are very easy to obtain: all econometric software ages worthy of the name will compute these standard error estimates (at the user's option) with uss at all. In Stata, for example, the command:

```
gress y x1 x2 x3
```

putes the parameter estimates, a standard error estimate for each parameter estimate, and so h for a regression model with dependent variable "y" and explanatory variables "x1", "x2", and '. The command for estimating this regression model and instead computing the robust standard r estimates is simply:

```
gress y x1 x2 x3 , ~~robust~~ vce (robust)
```

he formulas which the econometric software uses for computing the robust standard errors are ightforward, also. For example, the robust standard error estimator for $\hat{\beta}$ (the OLS slope mator in the Bivariate Regression Model), is computed as the square root of the variance mator:

$$\left[\operatorname{var}(\hat{\beta})\right]^{\text{robust}} = \sum_{i=1}^{N} \left(w_i^{\text{ols}}\right)^2 \left(u_i^{\text{fit}}\right)^2 \tag{10-20}$$

ch leads to the robust standard error estimate

$$\sqrt{\left[\operatorname{var}(\hat{\beta})\right]^{\text{robust}}} = \sqrt{\sum_{i=1}^{N} (w_i^{\text{ols}})^2 (u_i^{\text{fit}})^2} \tag{10-21}$$

re w_i^{ols} is the usual OLS weight for the ith term in the expression for $\hat{\beta}$ and u_i^{fit} is the observed ng error for the ith observation. These standard error estimates are usually nowadays ed "White-Eicker standard errors" after White (1980) and Eicker (1963), who first proposed n.

Vhere do Equation 10-20 – and analogous results for computing robust standard errors for fficient estimates from the Multiple Regression Model – come from? The remainder of this ion provides a simple demonstration that the expression given in Equation 10-20 is an unbiased mator for the actual sampling variance of $\hat{\beta}$, at least to the extent that the sample is sufficiently e that the fitting errors can be substituted for the model errors. In fact, this demonstration ires only a minor amendment to the end of the derivation of the sampling variance of $\hat{\beta}$ (in the ence of heteroscedastic errors) given in equations 10-11 through 10-14.

[35] ow large is large enough? Absent a simulation study for the particular data set under consideration, one can never be in. (And the bootstrap simulation method described in Section 11.8 is, as noted there, specifically not appropriate for a el with heteroscedastic errors.) Generally speaking, experienced analysts are usually dubious about using these results less than around 40 observations and casually confident about using them when the sample substantially exceeds 100 rvations.

This amended derivation begins in precisely the same way, so equations 10-11 and 10-12 are simply repeated here:

$$\begin{aligned}
\text{var}(\hat{\beta}) &= E\left[\left[\hat{\beta} - E[\hat{\beta}]\right]^2\right] \\
&= E\left[\left[\hat{\beta} - \beta\right]^2\right] \\
&= E\left[\left[\sum_{i=1}^{N} w_i^{\text{ols}} U_i\right]^2\right] \qquad \text{(10-11)} \\
&= E\left[\sum_{\ell=1}^{N} w_\ell^{\text{ols}} U_\ell \sum_{i=1}^{N} w_i^{\text{ols}} U_i\right] \\
&= E\left[H \sum_{i=1}^{N} w_i^{\text{ols}} U_i\right]
\end{aligned}$$

where H stands for the random variable $\sum_{\ell=1}^{N} w_\ell^{\text{ols}} U_j$. Bringing the factor of H inside the sum over the running index i,

this is a note to editor

should be ℓ (like the running index) not j

$$\begin{aligned}
\text{var}(\hat{\beta}) &= E\left[\sum_{i=1}^{N} H w_i^{\text{ols}} U_i\right] \\
&= E\left[\sum_{i=1}^{N} w_i^{\text{ols}} (U_i H)\right] \qquad \text{(10-12)} \\
&= \sum_{i=1}^{N} w_i^{\text{ols}} E[U_i H]
\end{aligned}$$

where the last step uses the Linearity Property of Chapter 2.

Here is where the derivation begins to differ. Evaluating the expectation in Equation 10-12,

$$\begin{aligned}
E[U_i H] &= E\left[U_i \sum_{\ell=1}^{N} w_\ell^{\text{ols}} U_\ell\right] \\
&= E\left[\sum_{\ell=1}^{N} w_\ell^{\text{ols}} U_i U_\ell\right] \qquad \text{(10-22)} \\
&= \sum_{\ell=1}^{N} w_\ell^{\text{ols}} E[U_i U_\ell] \\
&= w_i^{\text{ols}} E\left[U_i^2\right] \\
&\approx w_i^{\text{ols}} E\left[\left(U_i^{\text{fit}}\right)^2\right]
\end{aligned}$$

where (as in Equation 10-13) the third-to-last step again uses the Linearity Property and the second-to-last step follows because $E[U_i U_\ell]$ is zero when i is unequal to ℓ, due to the nonautocorrelation assumption. The final step in the sequence of equations above is now different, however. Here, instead of evaluating $E[U_i^2]$ as σ_i^2, we instead recognize that – for sufficiently large samples – the fitting error U_i^{fit} is a good approximation to the model error, U_i.

Substituting this expression for $E[U_iH]$ from Equation 10-22 into the previous expression – Equation 10-12 – for $\text{var}(\hat{\beta})$ yields

$$\begin{aligned}\text{var}(\hat{\beta}) &= \sum_{i=1}^{N} w_i^{\text{ols}} E[U_iH] \\ &= \sum_{i=1}^{N} w_i^{\text{ols}}\left(w_i^{\text{ols}} E\left[\left(U_i^{\text{fit}}\right)^2\right]\right) \\ &= \sum_{i=1}^{N} \left(w_i^{\text{ols}}\right)^2 E\left[\left(U_i^{\text{fit}}\right)^2\right]\end{aligned} \tag{10-23}$$

Replacing ~~[illegible]~~ u_i^{fit} by U_i^{fit} and taking

~~Taking~~ the expected value of both sides of Equation 10-20 yields

$$E\left[\left[\text{var}(\hat{\beta})\right]^{\text{robust}}\right] = E\left[\sum_{i=1}^{N}\left(w_i^{\text{ols}}\right)^2\left(U_i^{\text{fit}}\right)^2\right] = \sum_{i=1}^{N}\left(w_i^{\text{ols}}\right)^2 E\left[\left(U_i^{\text{fit}}\right)^2\right] \tag{10-24}$$

which has the same right-hand side as Equation 10~~-23. Thus~~,

$$E\left[\left[\text{var}(\hat{\beta})\right]^{\text{robust}}\right] \cong \text{var}(\hat{\beta}) \tag{10-25}$$

and the robust estimator of $\text{var}(\hat{\beta})$ given in Equation 10-20 is an unbiased estimator when the sample size is sufficiently large that the model errors can be replaced by the fitting errors.

This unbiasedness result is nice, but what's really wanted is a proof that the robust standard error estimate – the square root of Equation 10-20 – provides an at least consistent estimator of the actual standard errors of $\hat{\beta}$ – i.e. of $\sqrt{\text{var}(\hat{\beta})}$. This is not difficult to do, but requires the sharper statistical tools covered in Chapter 11.[36]

For samples sufficiently large that their use is justified, the robust standard error estimates discussed in this section completely resolve the problems with hypothesis testing and confidence interval construction posed by heteroscedastic model errors. Nevertheless, it is still a good idea to test for heteroscedasticity in the fashion described in the previous section, so as to give the data a chance to point one toward a better model specification.[37]

[36] That derivation constitutes Exercise 11-7.

[37] For large samples with heteroscedasticity of unknown form – and where it has proven infeasible to improve the model specification so as to eliminate the heteroscedasticity – the best estimation technique to use is Generalized Method of Moments (GMM) estimation. GMM is covered in Section 19.4. In this context GMM essentially implements the FGLS estimator discussed above (immediately below Equations 10-17 and 10-18), by using the squared fitting errors from the OLS regression – $(u_1^{\text{fit}})^2 \ldots (u_N^{\text{fit}})^2$ – to "estimate" the weights $\omega_1^2 \ldots \omega_N^2$ and then using these weights (as in Equation 10-18) to respecify the model so that the error term is (asymptotically) homoscedastic. The alert reader will recognize this latter substitution as the same "trick" used in the White-Eicker robust standard error estimator. The advantage of GMM, where it is conveniently implemented in the software one is using, is that it provides asymptotically efficient parameter estimates as well as consistent standard error estimates. As with the robust standard error estimates, its disadvantages are that it is only asymptotically justified and that it encourages one to view heteroscedasticity as a minor problem which is easily fixed, rather than as a symptom of model misspecification which merits further analysis.

of $\hat{\beta}$ is in fact positive, so much so that the null hypothesis H_o: $\beta = 0$ could be rejected using the estimated asymptotic distribution of $\hat{\beta}$ if $\hat{\beta}$ were credibly a consistent estimator of β. But $\hat{\beta}$ is clearly not a consistent estimator because measurement error in X_i is known to be a substantial problem. However, since Equation 11-30 implies that our OLS estimate is actually converging to a value *lower* than β, we can infer that, were the inconsistency in $\hat{\beta}$ removed, then this null hypothesis would be rejected with an even smaller p-value.

11.7 ENDOGENOUS REGRESSORS: JOINT DETERMINATION – INTRODUCTION TO SIMULTANEOUS EQUATION MACROECONOMIC AND MICROECONOMIC MODELS

Another source of endogeneity in explanatory variables is where the regressor is jointly determined with the dependent variable because the equation being estimated is part of a set of simultaneous equations. Endogeneity of this sort is so common in applied economic work as to be practically endemic.

One way to see why this is the case is to note that this kind of endogeneity is often alternatively described as being due to "reverse causation." This terminology arises because the essential source of this endogeneity is the breakdown of the causality assumption embodied in the usual regression model with fixed-regressors. In that model the sample fluctuations in the dependent variable are taken to be caused by fluctuations in the explanatory variable, but not vice-versa. (The sample variation in the explanatory variable – because it is "fixed in repeated samples" – cannot be caused by fluctuations in the dependent variable.)

Consider then, as an example, the consumption function in a simple macroeconomic model. This consumption function might explain sample variation in C_i (aggregate consumption expenditures) as being due to sample variation in aggregate income, Y_i:

$$C_i = \beta_1 + \beta_2 Y_i + U_i \tag{11-34}$$

a simultaneous?

But it is clearly untenable to assume that the causality runs only from aggregate income to aggregate consumption expenditures when (per the GDP identity) aggregate consumption expenditures form the bulk of aggregate income. In fact, it is a rarity to find an interesting explanatory variable in an economic model whose sample variation cannot, to one extent or another, be ascribed to variation in the dependent variable.

Why and how does this joint determination of dependent and explanatory variable cause endogeneity – and consequent inconsistency in the OLS parameter estimator?

Consider the quintessential microeconomic model, where the price fetched and the quantity traded of some good – athletic shoes, say – are determined by supply and demand in N similar markets and the N sample observations are the consequent market equilibrium values of price and quantity.

In this framework one would model the quantity of shoes demanded in the ith market as having been determined by utility maximization subject to a budget constraint. In the simplest possible setting, this yields a demand relationship like

$$Q_i^d = aP_i + bR_i + U_i \qquad U_i \sim \text{IID}(0, \sigma_u^2) \tag{11-35}$$

Q_i^d in this equation is the quantity of shoes demanded and R_i is some measurable, fixed, characteristic of the potential shoe customers in market i, such as average income or a taste parameter. In any practical setting there would be many such exogenous characteristics, but only one is specified here – and the intercept is dropped – just to keep the equations as simple as possible.[19] P_i is the price of a shoe;

$\tilde{V}$

APPENDIX 11.1: THE ALGEBRA OF PROBABILITY LIMITS

Let $\hat{V}$ and $\hat{V}$ be two estimators, each of which is defined in such a way that its probability limit is finite and each of which is based on observations $V_1 \ldots V_N$, which are assumed to all be identically and independently distributed.[52] Then it can be shown that the following results all hold:

1. (Linearity) Assuming that a and b are fixed constants,

$$\text{plim}(a\hat{V} + b\bar{V}) = a\,\text{plim}(\hat{V}) + b\,\text{plim}(\bar{V}) \qquad \text{(A11-1)}$$

multiplication sign, just like

2. (Multiplication)

$$\text{plim}(\hat{V} \times \bar{V}) = \text{plim}(\hat{V}) \overset{X}{\cancel{+}} \text{plim}(\bar{V}) \qquad \text{(A11-2)}$$

3. (Division) Assuming that $\text{plim}(\tilde{V}) \neq 0$,

$$\text{plim}\left(\frac{\hat{V}}{\tilde{V}}\right) = \frac{\text{plim}(\hat{V})}{\text{plim}(\tilde{V})} \qquad \text{(A11-3)}$$

4. (Reduction to Ordinary Limit) Assuming that F_N is a nonstochastic function of N,

$$\text{plim}(F_N) = \lim_{N \to \infty} [F_N] \qquad \text{(A11-4)}$$

5. (Slutsky Theorem) For any continuous function $F\{\cdot\}$,[53]

$$\text{plim}(F\{\hat{V}\}) = F\{\text{plim}(\hat{V})\} \qquad \text{(A11-5)}$$

6. (Law of Large Numbers) Assuming that $E[V_i]$ is finite,

$$\text{plim}\left(\frac{1}{N}\sum_{i=1}^{N} V_i\right) = E[V_i] \qquad \text{(A11-6)}$$

7. (Central Limit Theorem – Lindeberg-Lévy)

$$V_i \sim \text{IID}(\mu, \sigma_v^2) \Rightarrow \text{plim}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} V_i\right] = \text{plim}\left[\sqrt{N}\,\overline{V}\right] = N[\mu, \sigma_v^2] \qquad \text{(A11-7)}$$

8. (Asymptotic Equivalence)

If $\text{plim}(\tilde{V}) = \text{plim}(\hat{V})$, then their limiting distributions are the same. (A11-8)

[52] Strictly speaking, $\hat{V}$ and $\tilde{V}$ are two sequences of estimators such that the probability limit of each is either a finite number or a random variable with finite variance. Hamilton (1994, Chapter 7) provides a more detailed treatment and proofs, including extensions of these same results to sequences for which the underlying observations are a serially dependent time-series rather than independently distributed observations. These extensions are used in Chapter 13. James D. Hamilton (1994), *Time Series Analysis*, Princeton University Press: Princeton, NJ.

[53] A function $F\{z\}$ is continuous if and only if the limit of $F\{z\}$ as z approaches z_o is $F\{z_o\}$ for all values of z_o – i.e., the function has no sudden jumps.

$$= \text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\left(\frac{X_i-\overline{X}}{N\widehat{\text{var}}(X)}\right)U_i\right]$$

$$= \text{plim}\left[\frac{1}{\widehat{\text{var}}(X)}\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\overline{X})U_i\right]$$

$$= \text{plim}\left[\frac{1}{\widehat{\text{var}}(X)}\right]\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\overline{X})U_i\right]$$

$$= \frac{1}{\sigma_x^2}\,\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\overline{X})U_i\right] \qquad \text{(A11-12)}$$

Note that the validity of this last step rests on the indicated probability limit existing – that is, in this case, corresponding to the probability limit of a random variable with finite variance. That this is not a problem becomes clear later in the derivation.

In the next step μ_x, the population mean of X_i, is both added and subtracted in the expression. This leads to an additional term in the expression which, after a few more steps, is shown to have probability limit zero and then disappears from the remainder of the derivation. The reason why this apparent detour is necessary is that it effectively replaces $\overline{X}$ by μ_x in our eventual result, the zero-mean variable which we will call "V_i." This replacement is crucial in showing that the V_i are independently distributed because $\overline{X}$ depends on all of $X_1 \ldots X_N$ whereas μ_x is a fixed constant. Proceeding along these lines, then,

put subscript X
"μx"
where circled

$$\text{plim}\left[\sqrt{N}\left(\hat{\beta}^{\text{OLS}}-\beta\right)\right] = \frac{1}{\sigma_x^2}\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\overline{X})U_i\right]$$

$$= \frac{1}{\sigma_x^2}\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\mu_x+\mu_x-\overline{X})U_i\right]$$

$$= \frac{1}{\sigma_x^2}\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\mu_x)U_i+\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(\mu_X-\overline{X})U_i\right]$$

$$= \frac{1}{\sigma_x^2}\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\mu_x)U_i\right]+\frac{1}{\sigma_x^2}\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(\mu_X-\overline{X})U_i\right]$$

$$= \frac{1}{\sigma_x^2}\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\mu_x)U_i\right]+\frac{1}{\sigma_x^2}\text{plim}\left[(\mu_X-\overline{X})\sqrt{N}\,\overline{U}\right]$$

$$= \frac{1}{\sigma_x^2}\text{plim}\left[\sqrt{N}\sum_{i=1}^{N}\frac{1}{N}(X_i-\mu_x)U_i\right]+\frac{1}{\sigma_x^2}\text{plim}\left[(\mu_X-\overline{X})\right]\text{plim}\left[\sqrt{N}\,\overline{U}\right]$$

Two additional comments are worthwhile at this point. First, this proof can be easily extended to a generalization of the model in Equation 11-1 in which the homoscedasticity assumption – $\text{var}(U_i) = \sigma_u^2$ – is replaced by $\text{var}(U_i) = \sigma_{u,i}^2$. This extension requires the use of a more general form of the Central Limit Theorem than Equation ~~A11-10;~~ A11-7; this more general version instead assumes that $V_i \sim \text{IID}[0, \sigma_{v,i}^2]$ and replaces σ_v^2 in the right-hand side by the large-sample limit of the average value $\sigma_{v,i}^2$ over the sample – e.g., see Davidson and MacKinnon (2004, p. 149).[57] The extension is not difficult; it ends a few steps earlier than the present development, at the point in Equation A11-15 where

$$\text{var}(V_i)\frac{1}{\sigma_x^4}E\left([Xi - \mu_x]^2 U_i^2\right) \tag{A11-16}$$

so that the final result is

$$\sqrt{N}\left(\hat{\beta}^{\text{OLS}} - \beta\right) \xrightarrow{d} N\left[0, \frac{E\left([X_i - \mu_x]^2 U_i^2\right)}{\sigma_x^4}\right] \tag{A11-17}$$

and the expectation on the right-hand side is in practice replaced by the analogous sample average, which consistently estimates it. The resulting asymptotic standard error estimates for $\hat{\beta}^{\text{OLS}}$ are identical to the White-Eicker "robust" standard error estimates discussed in Section 10.7 and readily available in most econometric packages.[58]

Second, it is also worth noting that – were one unwilling to make the assumption that $\text{cov}\left(X_i^2, U_i^2\right)$ is zero – one could still reach the result given as Equation A11-17 and use this for inference. Implementing this result amounts to simply using the White-Eicker robust standard errors even though the homoscedasticity assumption is being maintained. Asymptotically, this makes no difference, but – in more modest samples – one would expect to obtain more accurate standard error estimates using the A11-9 result.

fact appear to require that the V_i's be identically distributed – it apparently suffices that their means and variances are the same. Nor is independence clearly necessary, as uncorrelatedness sufficed in order to derive the needed result on the sampling variance of $\overline{V}$.

A great deal of work has been done proving LLNs for various different variations on the IID assumption. This work has also led to a set of "algebraic" results on probability limits which make them extremely easy to manipulate and evaluate. Appendix 11.1 lists a useful selection of these results, all of which have been proven for the case where the underlying data ~~is~~ are identically and independently distributed.[9] Looking at these results, one can see that it is a simple matter to evaluate the probability limit of a sum, product, or quotient of two estimators. Appendix 11.1 also gives a result – the "Slutsky Theorem" – which states that the probability limit of a continuous function o an estimator equals that function of probability limit of the estimator. Thus, for example, if one knows the probability limit of an estimator of a population variance, then the probability limit of the standard deviation is just the square root of this, since the square root is a continuous function.[1 Note also that the restatement of the LLN in Appendix 11.1 is more powerful than it might at firs appear, as the V_i can be any random variable (or function of several multiple random variables whose expectation is finite. Thus, for example, this result implies that the probability limit of sample variance or covariance is equal to the corresponding population variance or covariance.

The net effect of these results is to make it fairly easy to express the probability limit of practicall any estimator in terms of the combination of population moments which might correspond to wha the estimator is attempting to estimate. This ability to manipulate probability limits motivates th formal definition of the property of consistency for an estimator of some population quantity (c model parameter) μ:

Consistency of Estimator $\hat{V}$

$$\text{plim}(\hat{V}) = \mu \iff \hat{V} \text{ in a consistent estimator of } \mu \qquad (11\text{-}1$$

Thus, it is often feasible to evaluate the probability limit of a regression model parameter estimat and say something cogent as to the circumstances under which it is (or is not) consistent. Indeed, th is exactly what we will do with respect to $\hat{\beta}^{\text{OLS}}$ in the next section.

Most analysts regard consistency as an essential quality in an estimator, to the point where estimator which is not at least consistent is pretty much regarded as junk. In principle that is r really justified: one can imagine an estimator which is slightly inconsistent, but which is quite use because it has a small mean square error for modest sample sizes.[11] And a consistent estimator is r necessarily a good estimator. For example, per Exercise 11-2, if $\hat{V}$ is a consistent estimator of then so is $\hat{V} + [10^{1000}/N^{.001}]$. At least one of this pair of estimators is incredibly terrible! Still, actual estimators which we find to be consistent in the context of regression models in pract appear to be fairly useful, at least in reasonably large samples. And – as noted above – most peo simply have no use for estimators which are not at least consistent.

[9] Hamilton (1994, Chapter 7) provides a more detailed treatment of these results, including both proofs and extension these same results to sequences of random variables for which the underlying observations V_1 ... V_N are a serially correl time-series rather than independently distributed observations. An extension of this sort is needed in order to apply t results to the autoregressive models considered in Chapter 13. James D. Hamilton (1994, Chapter 7) *Time-Series Anal* Princeton University Press: Princeton.

have this property! Thus, the square

APPENDIX 12.2: PROOF THAT THE 2SLS COMPOSITE INSTRUMENT IS ASYMPTOTICALLY UNCORRELATED WITH THE MODEL ERROR TERM

The Two-Stage Least Squares composite instrument for the endogenous regressor $X_{i,2}$ in the Multiple Regression Model (Equation 12-33) is $\hat{X}_{i,2}$, which is given in Equation 12-37 as

$$\hat{X}_{i,2} = \hat{\gamma}_{l,1}^{OLS} Z_{i,1} + \hat{\gamma}_{2,1}^{OLS} Z_{i,2} + \hat{\gamma}_{3,1}^{OLS} Z_{i,3} + \hat{\gamma}_{4,1}^{OLS} X_{i,4} + \hat{\gamma}_{5,1}^{OLS} X_{i,5} + \ldots + \hat{\gamma}_{k+3,1}^{OLS} X_{i,k} \qquad \text{(A12-10)}$$

where the estimators $\hat{\gamma}_{l,1}^{OLS} \ldots \hat{\gamma}_{k+3,1}^{OLS}$ are obtained from OLS estimation of the first-stage regression equation, Equation 12-34. The purpose of this section is to show that $\hat{X}_{i,2}$ is asymptotically uncorrelated with U_i, the model error term in Equation 12-33, so long as the non-endogenous variables in the model $(X_{i,4} \ldots X_{i,k})$ and the three instruments $(Z_{i,1},\ Z_{i,2},\ \text{and}\ Z_{i,3})$ are all asymptotically uncorrelated with U_i.[46]

$\hat{X}_{i,2}$ is asymptotically uncorrelated with U_i if $\text{plim}\{\widehat{\text{cov}}(\hat{X}_{i,2}, U_i)\}$ is zero. From the definition of a sample covariance – and letting $\overline{X}_{\text{hat},2}$ stand for the sample mean of $\hat{X}_{i,2}$ – this is equivalent to requiring that

$$\begin{aligned}
\text{plim}\{\widehat{\text{cov}}(\hat{X}_{i,2}, U_i)\} &= \text{plim}\left\{\frac{1}{N}\sum_{i=1}^{N}[(\hat{X}_{i,2} - \overline{X}_{\text{hat},2})(U_i - \overline{U})]\right\} \\
&= \text{plim}\left\{\frac{1}{N}\sum_{i=1}^{N}[\hat{X}_{i,2}U_i] - \overline{X}_{\text{hat},2}\overline{U}\right\} \\
&= \text{plim}\left\{\frac{1}{N}\sum_{i=1}^{N}[\hat{X}_{i,2}U_i]\right\} - \text{plim}\{\overline{X}_{\text{hat},2}\}\text{plim}\{\overline{U}\} \\
&= \text{plim}\left\{\frac{1}{N}\sum_{i=1}^{N}[\hat{X}_{i,2}U_i]\right\}
\end{aligned} \qquad \text{(A12-11)}$$

The second equality in Equation A12-11 follows from the result obtained in Exercise 12-10; the third equality follows from the addition and multiplication properties given in Appendix 11.1. The probability limit of $\overline{U}$ is equal to the population mean of $\overline{U}$ from the Law of Large Numbers given in Appendix 11.1; this population mean is zero because (due to the inclusion of an intercept in the model) each of the U_i has mean zero.

Substituting into Equation A12-11 the definition of $\hat{X}_{i,2}$, the Two-Stage Least Squares composite instrument for the endogenous regressor $X_{i,2}$, from Equation 12-37, or A12-10, yields

Ui not U1

$$\begin{aligned}
&\text{plim}\{\widehat{\text{cov}}(\hat{X}_{i,2}, U_1)\} \\
&= \text{plim}\left\{\frac{1}{N}\sum_{i=1}^{N}\left[\left(\hat{\gamma}_{l,1}^{OLS} Z_{i,1} + \hat{\gamma}_{2,1}^{OLS} Z_{i,2} + \hat{\gamma}_{3,1}^{OLS} Z_{i,3} + \hat{\gamma}_{4,1}^{OLS} X_{i,4} + \hat{\gamma}_{5,1}^{OLS} X_{i,5} + \ldots + \hat{\gamma}_{k+3,1}^{OLS} X_{i,k}\right)U_i\right]\right\} \\
&= \text{plim}\left\{\frac{1}{N}\sum_{i=1}^{N}(\hat{\gamma}_{l,1}^{OLS} Z_{i,1} U_i + \hat{\gamma}_{2,1}^{OLS} Z_{i,2} U_i + \hat{\gamma}_{3,1}^{OLS} Z_{i,3} U_i + \hat{\gamma}_{4,1}^{OLS} X_{i,4} U_i + \hat{\gamma}_{5,1}^{OLS} X_{i,5} U_i\right.
\end{aligned}$$

Ui not U1

$$
\begin{aligned}
&= \text{plim}\Big\{\boxed{\underset{1}{\hat{\gamma}^{\text{OLS}}_{i,1}}}\widehat{\text{cov}}(Z_{i,1}, U_i) + \hat{\gamma}^{\text{OLS}}_{2,1}\widehat{\text{cov}}(Z_{i,2}, U_i) + \hat{\gamma}^{\text{OLS}}_{3,1}\widehat{\text{cov}}(Z_{i,3}, U_i) + \hat{\gamma}^{\text{OLS}}_{4,1}\widehat{\text{cov}}(X_{i,4}, U_i) \\
&\qquad + \hat{\gamma}^{\text{OLS}}_{5,1}\widehat{\text{cov}}(X_{i,5}, U_i) + \ldots + \hat{\gamma}^{\text{OLS}}_{k+3,1}\widehat{\text{cov}}(X_{i,k}, U_i)\Big\} \\
&= \text{plim}\Big\{\hat{\gamma}^{\text{OLS}}_{1,1}\Big\}\text{plim}\{\widehat{\text{cov}}(Z_{i,1}, U_i)\} + \text{plim}\Big\{\hat{\gamma}^{\text{OLS}}_{2,1}\Big\}\text{plim}\{\widehat{\text{cov}}(Z_{i,2}, U_i)\} \\
&\qquad + \text{plim}\Big\{\hat{\gamma}^{\text{OLS}}_{3,1}\Big\}\text{plim}\{\widehat{\text{cov}}(Z_{i,3}, U_i)\} \\
&\quad + \text{plim}\Big\{\hat{\gamma}^{\text{OLS}}_{4,1}\Big\}\text{plim}\{\widehat{\text{cov}}(X_{i,4}, U_i)\} + \text{plim}\Big\{\hat{\gamma}^{\text{OLS}}_{5,1}\Big\}\text{plim}\{\widehat{\text{cov}}(X_{i,5}, U_i)\} \\
&\qquad + \ldots + \text{plim}\Big\{\hat{\gamma}^{\text{OLS}}_{k+3,1}\Big\}\text{plim}\{\widehat{\text{cov}}(X_{i,k}, U_i)\}
\end{aligned}
\tag{A12-12}
$$

'here the result of Exercise 12-10 is used to convert terms like $\frac{1}{N}\sum_{i=1}^{N} Z_{i,1}\, U_i$ into $\widehat{\text{cov}}(Z_{i,1}\, U_i)$.[47]

inally, note that each term in the right-hand side of the last part of Equation A12-16 contains a ıctor which is the probability limit of the sample covariance of either one of the three instruments $Z_{i,1}$, $Z_{i,2}$, and $Z_{i,3}$) or of one of the non-endogenous variables in the model $(X_{i,4} \ldots X_{i,k})$ with the ıodel error (U_i). Consequently, if each of these variables is asymptotically uncorrelated with the ıodel error, then the composite instrument $\hat{X}_{i,2}$ is asymptotically uncorrelated with U_i, the model rror term.

everywhere* circled on pp 340, 341 $\hat{\gamma}^{OLS}_{i,1}$ should be $\hat{\gamma}^{OLS}_{1,1}$

Equation 11-6 from Section 11.4 can now be written as

$$\begin{aligned}
\hat{\varphi}_1^{\text{OLS}} &= \varphi_1 + \sum_{t=2}^{T} W_t^{\text{ols}} U_t - \overline{U}\sum_{t=2}^{T} W_t^{\text{ols}} \\
&= \varphi_1 + \sum_{t=2}^{T} W_t^{\text{ols}}\left(U_t - \overline{U}\right) \qquad \text{(13-} \\
&= \varphi_1 + \sum_{t=2}^{T}\left[\frac{Y_{t-1} - \overline{Y}}{\sum\limits_{\ell=2}^{T}\left(Y_{\ell-1} - \overline{Y}\right)^2}\right]\left(U_t - \overline{U}\right)
\end{aligned}$$

Letting G now stand for the sum $\sum\limits_{\ell=2}^{T}\left(Y_{\ell-1} - \overline{Y}\right)$, this becomes

$$\begin{aligned}
\hat{\varphi}_1^{\text{OLS}} &= \varphi_1 + \sum_{t=2}^{T}\left[\frac{Y_{t-1} - \overline{Y}}{G}\right]\left(U_t - \overline{U}\right) \\
&= \varphi_1 + \frac{1}{G}\sum_{t=2}^{T}\left(Y_{t-1} - \overline{Y}\right)\left(U_t - \overline{U}\right) \\
&= \varphi_1 + \frac{\sum\limits_{t=2}^{T}\left(Y_{t-1} - \overline{Y}\right)\left(U_t - \overline{U}\right)}{G} \\
&= \varphi_1 + \frac{\sum\limits_{t=2}^{T}\left(Y_{t-1} - \overline{Y}\right)\left(U_t - \overline{U}\right)}{\sum\limits_{\ell=2}^{\boxed{N}}\left(Y_{\ell-1} - \overline{Y}\right)^2} \qquad \text{(1} \\
&= \varphi_1 + \frac{\frac{1}{T-1}\sum\limits_{t=2}^{T}\left(Y_{t-1} - \overline{Y}\right)\left(U_t - \overline{U}\right)}{\frac{1}{T-1}\sum\limits_{\ell=2}^{\boxed{N}}\left(Y_{\ell-1} - \overline{Y}\right)^2} \\
&= \varphi_1 + \frac{\widehat{\text{cov}}\left(Y_{t-1},\, U_t\right)}{\widehat{\text{var}}\left(Y_{t-1}\right)}
\end{aligned}$$

T

Taking the probability limit of both sides of Equation 13-25 and applying the results Appendix 11.1,[33]

[33] The results on probability limits given in Appendix 11.1 assumed that the relevant random variables are indepe distributed, whereas that is plainly not the case for the serially dependent time-series $Y_2 \ldots Y_T$. Also, the tim $Y_1U_2 \ldots Y_{T-1}U_T$ – although serially uncorrelated – is not serially independent, either. It can be shown, however – Hamilton (1994, Chapter 7) – that all of the Appendix 11.1 results follow for both of these serially dependent series both of these time-series are covariance stationary and because for each the sum of the magnitudes of its autocovar bounded. (See Exercises 13-4d and 13-5.) The derivation of Equation 13-27, the asymptotic sampling distribution f however, must apply a Central Limit Theorem with regard to sequences of sample means of $Y_{t-1}U_t$, which requires sc closer to serial independence (a "martingale difference") on the part of $Y_{t-1}U_t$; this is discussed in Appendix 13

T not N

$$
\begin{aligned}
\text{plim}\left(\hat{\varphi}_1^{\text{OLS}}\right) &= \text{plim}\left(\varphi_1 + \frac{\widehat{\text{cov}}(Y_{t-1},\, U_t)}{\widehat{\text{var}}(Y_{t-1})}\right) \\
&= \text{plim}(\varphi_1) + \text{plim}\left(\frac{\widehat{\text{cov}}(Y_{t-1},\, U_t)}{\widehat{\text{var}}(Y_{t-1})}\right) \\
&= \lim_{N\to\infty}(\varphi_1) + \text{plim}\left(\frac{\widehat{\text{cov}}(Y_{t-1},\, U_t)}{\widehat{\text{var}}(Y_{t-1})}\right) \\
&= \varphi_1 + \frac{\text{plim}(\widehat{\text{cov}}(Y_{t-1},\, U_t))}{\text{plim}(\widehat{\text{var}}(Y_{t-1}))} \\
&= \varphi_1 + \frac{\text{cov}(Y_{t-1},\, U_t)}{\text{var}(Y_{t-1})} \\
&= \varphi_1
\end{aligned}
\qquad (13\text{-}26)
$$

where the second step in Equation 13-26 uses the "Linearity" property (with $a = b = 1$), the third step uses the "Reduction to Ordinary Limit" property (since φ_1 is not random), and the fourth step uses the "Division" property.[34] Finally, the last two steps in this equation use the Law of Large Numbers property to evaluate the probability limits of the sample covariance of Y_{t-1} with U_t and the sample variance of Y_{t-1}. The covariance of Y_{t-1} with U_t is zero because the MA(∞) form of the model in Equation 13-17 implies that Y_{t-1} does not depend on the current value of the model error – see Exercise 13-4.

Thus, $\hat{\varphi}_1^{\text{OLS}}$ is consistent for φ_1, albeit biased. The asymptotic sampling distribution for $\hat{\varphi}_1^{\text{OLS}}$ is derived in Appendix 13.1; the result is

$$
\sqrt{T-1}\left(\hat{\varphi}_1^{\text{OLS}} - \varphi_1\right) \xrightarrow{d} N\left[0, \frac{\sigma_u^2}{\text{var}(Y_{t-1})}\right] = N\left[0, 1 - \varphi_1^2\right] \qquad (13\text{-}27)
$$

where the second part of Equation 13-27 recognizes that variance of Y_{t-1} is identical to the variance of Y_t itself; the result from Equation 13-19 is substituted in on the right-hand side of this equation.[35]

Consequently – just as in Section 11.4 – the asymptotic sampling distribution of $\hat{\varphi}_1^{\text{OLS}}$ yields essentially the same computing formulas (for standard errors, confidence intervals, hypothesis test rejection p-values, etc.) as does the Bivariate Regression Model with a fixed-regressor. Therefore – so long as Y_t really is generated by the AR(1) model, and so long as the sample length T is sufficiently large that the asymptotic distribution $\hat{\varphi}_1^{\text{OLS}}$ is a reasonable approximation to its actual sampling distribution – the results from the usual OLS computer output can be interpreted as if the explanatory variable were fixed in repeated samples.

Broadly speaking, this aspect of the asymptotic results obtained here in the specific setting of the AR(1) Model carries over to the Multiple Regression Model, including explanatory variables in addition to Y_{t-1}: the results from the usual OLS computer output can be interpreted as if the explanatory variable were fixed in repeated samples *if* the sample is large enough for the asymptotic sampling distribution of the parameter estimators to be relevant and *if* the model is well specified in terms of error term homoscedasticity, omitted variables, and endogeneity issues.

[34] Note that the division property requires that the probability limit in what will become the denominator is not zero; this is not an issue here, because it is obvious that the variance of Y_{t-1} is strictly positive.

[35] Of course, per Exercise 13-3, if the true generating mechanism for Y_t is actually some other model – such as a higher-order AR model, or one of the nonlinear models discussed in Section 18.6 – then Equation 13-11 is omitting explanatory variables, one or more of which might be correlated with Y_{t-1}. In that case $\hat{\varphi}_1^{\text{OLS}}$ is no longer a consistent estimator of φ_1 and Equation 13-27 is not the asymptotic sampling distribution for $\hat{\varphi}_1^{\text{OLS}}$.

dure like FGLS to be reasonable, but insufficient data for it to be reasonable to
cified model.[49]

pag
377

other modeling situations the respecification of the model to include dynamics can destroy the point e enterprise. Active Learning Exercise 13b (available at www.wiley.com/college/ashley) provides a example of this situation. The point of this exercise is to empirically examine the stability over time particular kind of macroeconomic monetary policy called the "Taylor Rule." The sample data ly indicate that the implied model error term in this model for how the U.S. central bank sets the short interest rate is serially correlated to a significant degree and in a manner which changes over the ous sub-periods considered. The serial correlation can be modeled by including different dynamics ach model, but the models for the various sub-periods would then no longer be comparable with other and none would correspond to the "Taylor Rule" itself as it is commonly understood.

he most graceful approach in this circumstance is to leave the model specification itself alone and ompute what are called "Newey-West" standard errors for the model coefficient estimates and use e in the inference work. Newey-West standard error estimates are the "serial correlation" logue to the White-Eicker ("robust") standard error estimates discussed in Section 10.7 in nection with heteroscedastic model errors. Like the White-Eicker standard error estimates, they only asymptotically justified – and hence are only appropriate to use in large samples – and induce improvement (or even any change) in the parameter estimates. But they do provide consistent mates of the parameter estimator standard errors, so that one can still construct asymptotically d confidence intervals and hypothesis tests, even though the model errors are serially correlated. like the White-Eicker estimates – which required no specification at all with regard to the form of heteroscedasticity – the Newey-West standard error estimates require that one at least specify a ximum lag, beyond which the model errors are assumed to be serially uncorrelated.

The idea underlying the Newey-West standard error estimates is otherwise so similar to that lerlying the White-Eicker estimates that the simplest way to describe it is to just repeat the mula given in Chapter 10 for the White-Eicker robust standard error estimator for $\hat{\beta}$, the OLS pe estimator in the Bivariate Regression Model and then restate it as the corresponding Newey-st estimator. From Equation 10-20, the White-Eicker (robust) standard error estimate is

$$\sqrt{\left[\operatorname{var}(\hat{\beta})\right]^{\text{robust}}} = \sqrt{\sum_{i=1}^{N} \left(w_i^{\text{ols}}\right)^2 \left(u_i^{\text{fit}}\right)^2} \qquad (10\text{-}21)$$

ere w_i^{ols} is the usual OLS weight for the ith term in the expression for $\hat{\beta}$ and u_i^{fit} is the observed ing error for the ith observation. The corresponding Newey-West estimator, allowing for serial rrelation out to (for example) one lag is

$$\sqrt{\left[\operatorname{var}(\hat{\beta})\right]^{\text{Newey-West}}} = \sqrt{\sum_{i=1}^{\overset{T}{N}} \left(w_i^{\text{ols}}\right)^2 \left(u_i^{\text{fit}}\right)^2 + \sum_{i=2}^{N} \left(w_i^{\text{ols}} w_{i-1}^{\text{ols}}\right)\left(u_i^{\text{fit}} u_{i-1}^{\text{fit}}\right)} \qquad (13\text{-}37)$$

all become t

ote that the Newey-West estimator subsumes the White-Eicker terms, so it is also robust against teroscedasticity in the model errors.[50]

[49] In that circumstance – with model parameters $\beta_0 \ldots \beta_k$ and AR model parameters $\varphi_0 \ldots \varphi_p$ – one might estimate the rameters in the equation analogous to Equation 13-35 directly by minimizing the sum of squared fitting errors with respect these $k+p+2$ parameters. Since the model is now a nonlinear function of the parameters, this procedure is called onlinear least squares" rather than "ordinary least squares" and is numerically more challenging because it involves a merical minimization of a function of $k+p+2$ variables rather than just solving linear equations. This minimization is sentially identical to the analogous Hildreth-Lu FGLS procedure described in the body of the text. Many econometrics ograms provide a facility for estimating regression models using nonlinear least squares – e.g., the "nl" command in Stata.

[50] Equation 13-37 is correct but a bit misleading, as the additional terms which would be added so as to account for serial rrelation at larger lags will enter with diminished weights. This down-weighting is necessary so as to ensure that the Newey-Vest variance estimate is always positive – e.g., see Davidson and MacKinnon (1993, p. 611).

```
-------------+------------------------------           F(  4,   519) =   23.04
       Model |  .005673387     4  .001418347           Prob > F      =  0.0000
    Residual |  .031950486   519  .000061562           R-squared     =  0.1508
-------------+------------------------------           Adj R-squared =  0.1442
       Total |  .037623873   523  .000071939           Root MSE      =  .00785

------------------------------------------------------------------------------
   chg_coint |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       coint |
         L1. |   -.096222   .0249325    -3.86   0.000    -.1452031    -.047241
   chg_coint |
         L1. |  -.2029439   .0446796    -4.54   0.000     -.290719   -.1151688
         L2. |  -.1837483    .043891    -4.19   0.000     -.269974   -.0975225
         L3. |  -.2233371   .0426831    -5.23   0.000      -.30719   -.1394842
       _cons |  -.0000405   .0003428    -0.12   0.906    -.0007139    .0006329
------------------------------------------------------------------------------
```

page
429

The ADF test statistic here is −3.86, which is less than the 5% critical point (−3.768) in the second line of Table 14-2, so one can reject the null hypothesis that these cointegration equation fitting errors are I(1) at the 5% (although not the 1%) level of significance.[50] Thus, if one rejects the null hypothesis that the cointegrating regression fitting errors are I(1) – and hence accepts the hypothesis that $log(C_t)$, $log(Y_t)$, and R_t are cointegrated – then there is only a 5% chance that this conclusion is simply an artifact of sampling error in the Engle-Granger ADF test regression.

The estimated cointegrating regression model thus indicates that "$coint_t$," defined by the equation:

$$coint_t \equiv log(C_t) - 1.02573\,log(Y_t) + .00506R_t + .29725 \qquad (14\text{-}15)$$

is an I(0) time-series with zero mean. Equation 14-15 can therefore be interpreted as an estimated long-run relationship between these three time-series. Importantly, the existence of this long-run relationship explains why the time plot in Figure 14-11 shows that $log(C_t)$ does not wander off at random from its trend value: were $log(C_t)$ to do so, then it would depart overmuch from the expected value for it implied by Equation 14-15.

The estimated long-run relationship implied by the cointegration of these three time-series strongly suggests that $coint_{t-1}$, the lagged fitting errors from the cointegration regression model, "belongs" as an explanatory variable in the model for the changes in $log(C_t)$; explanatory variables such as $coint_{t-1}$ are called "error-correction terms." In fact, this interpretation of these fitting errors predicts that $coint_{t-1}$ will enter the model for $\Delta log(C_t)$ with a *negative* coefficient. In this way a value of $log(C_t)$ which is larger than is consistent with the long-run relationship pushes the expected value of $\Delta log(C_t)$ down, tending to bring $log(C_t)$ back in consonance with the long-run relationship. Similarly, a negative coefficient on $coint_{t-1}$ implies that a value of $log(C_t)$ which is smaller than is consistent with the long-run relationship pushes the expected value of $\Delta log(C_t)$ up, again tending to bring $log(C_t)$ back toward the long-run relationship.

Adding $coint_{t-1}$, obtained by lagging Equation 14-15 one month, to the model for $\Delta log(C_t)$ from Section 14.4 yields what is called the "Vector Error-Correction Model" or "VECM" for $\Delta log(C_t)$:

[50] The critical points from the second line of Table 14-2 are appropriate because there are now two variables in the cointegratiag regression model.

cointegra ting

and p_re denote the variable names for the corresponding "quasi-differenced" variables, then the Stata command:

```
ivregress 2sls cig_re (cig_lag1_re p_re = q1 q2 q3) tax_re , vce(cluster statenum)
```

also yields consistent two-stage least squares estimates of β_{cig}, β_{tax}, and β_{price}, presuming that the variables and instruments are all uncorrelated with ν_i. As in Section 16.1, if the parameter estimates of greatest interest are not significantly different across these two models, then one would prefer the more efficient Random Effects Model estimates.

Unfortunately, it is not immediately obvious what actual variables one could credibly use as the instruments $Z1_{i,t}$, $Z2_{i,t}$, and $Z3_{i,t}$ (or $Q1_{i,t}$, $Q2_{i,t}$, and $Q3_{i,t}$) in these models. In some regression models the underlying economic theory implies that certain variables will be exogenous; utilizing those implications in justifying the estimation procedure, however, makes the consistency of the model parameter estimators conditional on the validity of the theory.[10] This is fine for testing the theory – but only if one can also consistently estimate the model's parameters in a separate fashion that is *not* conditional on the theory's validity. Which leaves the analyst, again, needing credibly valid instruments.

Fortunately, there is an alternative approach, the "First-Differences Model," which provides a solution – albeit an imperfect one – to the problem of consistently estimating coefficients such as β_{cig}, β_{tax}, and β_{price} in a model like Equation 16-41.

Aside from the pooled regression model, which basically ignores the heterogeneity induced by the country/state/individual-specific component (ν_i) altogether, the First-Differences Model is actually the simplest framework so far considered. In the vocabulary of Chapter 13, it consists of simply considering the model "in changes." Thus, for the cigarette consumption model considered here (Equation 16-41), the First-Differences Model is just

$$\begin{aligned}\Delta CIG_{i,t} &= \beta_{cig}\Delta CIG_{i,t-1} + \beta_{tax}\Delta TAX_{i,t} + \beta_{price}\Delta P_{i,t} + \Delta\nu_i + \Delta\varepsilon_{i,t} \\ &= \beta_{cig}\Delta CIG_{i,t-1} + \beta_{tax}\Delta TAX_{i,t} + \beta_{price}\Delta P_{i,t} + \Delta\varepsilon_{i,t}\end{aligned} \qquad (16\text{-}47)$$

where $\Delta CIG_{i,t}$ is $CIG_{i,t} - CIG_{i,t-1}$, $\Delta CIG_{i,t-1}$ is $CIG_{i,t-1} - CIG_{i,t-2}$, $\Delta TAX_{i,t}$ is $TAX_{i,t} - TAX_{i,t-1}$, and so forth for the other variables.

Note that ν_i, the state-specific component of the error term in Equation 16-41, is eliminated by this transformation, along with the intercept (had there been one in Equation 16-41) and any explanatory variables which do not vary over time. Thus, the first-difference transformation at one stroke eliminates the non homogeneity induced by the panel nature of the state-level cigarette consumption data, but – like the "within" transformation of the Fixed Effects Model – it also eliminates the possibility of quantifying how $CIG_{i,t}$ varies with any time-invariant explanatory variables.[11]

Also, one can (and ordinarily should) still include an intercept (α) and a set of year-dummy variables in a model specification like Equation 16-47. The inclusion of these terms guarantees that

[10] For example, as noted in Footnotes 16-1 and 16-7, Equation 16-41 is a subset of the "myopic" model considered in the Becker, Grossman, and Murphy (1994) analysis of cigarette addiction. Their more sophisticated ("rational") theory of addiction – which was the point of their paper – implies the additional inclusion of $CIG_{i,t+1}$ as an explanatory variable. Moreover, their theory also implies that current cigarette consumption, $CIG_{i,t}$, depends on past and future values of $P_{i,t}$ only through their impact on $CIG_{i,t-1}$ and $CIG_{i,t+1}$. Consequently – conditional on CIG_{t-1} and $CIG_{i,t+1}$ – both past and future values of $P_{i,t}$ provide valid instruments for use in estimating the regression parameters in their "rational" model setting. 15.5

[11] The Hausman-Taylor estimator discussed at the end of Section 14.3 allows for some time-invariant explanatory variables, but assumes that all of the explanatory variables are strictly exogenous with respect to $\varepsilon_{i,t}$ and also does not allow for heteroscedasticity and/or serial correlation in $\varepsilon_{i,t}$; work on generalizing this approach is currently in progress.

p. 545

w can one tell whether a particular, ...
iance stationary? Just as with diagnostically checking the assumptions of the Multiple
ssion Model – e.g., of homoscedastic and serially uncorrelated model errors – one can never
or sure. Still, sensible and useful diagnostic checking of these regression assumptions is
ble in practice, so long as one has a reasonable amount of data.[21]
ecking a time-series for covariance stationarity is a similar enterprise. The regression model
ostic checking used data plots and auxiliary regressions; covariance stationarity checking
entrates on two tools: a time-plot of the sample data and a "sample correlogram."
e sample correlogram is simply a plot (and/or tabulation) of the sample autocorrelations – i.e.,
2, $r_3 \ldots r_k$, where $k \leq 20$ is usually plenty for quarterly data and $k \leq 60$ – so as to include lags up
e years – is usually plenty for monthly data. A sample correlogram for a time-series is very easy
oduce in most econometric software packages; for a time-series named *ydat*, for example, the
mands in Stata are:[22]

```
et tee
dat, lags(60)
rgram ydat, lags(60)
```

efore looking as some illustrative sample output from these commands, it is helpful to first
mine some useful results on the asymptotic sampling distribution of the estimator R_k, obtained
Bartlett (1946) and Marriot and Pope (1954).[23] Understanding these results – and their
itations – is essential to effectively interpreting the estimates, $r_1 \ldots r_k$. In the way of
mitations," please note at the outset that these results were all derived under the assumption
t N is large and that the time-series $Y_1 \ldots Y_N$ is covariance stationary and (jointly) normally
tributed. As we will see, a bit can be said about how large N needs to be, but the best advice with
gard to data which are highly nonnormal is to pretty much ignore the Bartlett results.[24]
Assuming, then, that the time-series $Y_1 \ldots Y_N$ is covariance stationary and (jointly) normally
stributed, Bartlett (1946) finds that the asymptotic sampling distribution of R_k is

$$\sqrt{N}(R_k - \rho_k) \xrightarrow{d} N\left[0, \sum_{\cancel{i=1}}^{\infty} \rho_i^2\right] \longrightarrow i = -\infty \tag{17-11}$$

hus, R_k is consistent (as an estimator of ρ_k), asymptotically normal, and has an asymptotic variance
f $(1/N) \sum_{i=\cancel{1}\,-\infty}^{\infty} \rho_i^2$. It is commonplace, therefore, for econometric software to blithely replace ρ_i by r_i in

For example, using the plots and auxiliary regressions summarized in Equation 15-3 of Section 15.6. Note that the
omoscedasticity assumption corresponds precisely to a portion of what is required for covariance stationarity. In fact – to
ake the connection explicit – the assumptions made about the model error term (U_t) in the Multiple Regression Model
orrespond precisely to assuming that U_t is covariance stationary, serially uncorrelated, and normally distributed.

[2]2 As noted in earlier chapters, Stata needs to be informed (once) that the data are in time order, using the "tsset" command to
pecify a variable which increases monotonically with time; here, it is assumed that a user-created variable named *tee* has this
roperty. These commands compute $r_1 \ldots r_{60}$; 40 lags is the default value; the "ac" command makes the plot and the
'corrgram" command makes the tabulation; most people find the plot far more useful than the tabulation.

[23] See also discussion in Kendall and Stuart (1963, Volume 1). Bartlett, M. S. (1946), "On the Theoretical Specification
of Sampling Properties of Autocorrelated Time-Series," *Journal of the Royal Statistical Society B 8*, pp. 27–41; Marriott,
F. H. C., and J. A. Pope (1954), "Bias in the Estimation of Autocorrelations," *Biometrika 41(3)*, pp. 390–402; Kendall, M. G.,
and A. Stuart (1963), *The Advanced Theory of Statistics 1*, Griffin: London.

[24] Non-normality becomes a substantial issue when working with dispersion data – such as volatility measures for financial
returns data. It is also an issue for nonlinear time-series modeling, because nonlinear generating mechanisms generally induce
non-normality in the data. In this regard, note that transforming each observation in a non-normally distributed time-series so
that a histogram of the observations looks normally distributed does not eliminate any nonlinearity in the way the time-series
relates to its own recent past; it also does not induce *joint* normality in the time-series.

Estimated Bartlett 95% Confidence Interval for ρ_k
(Assuming $\rho_k, \rho_{k+1}, \rho_{k+2}, \ldots$ are all zero.)

$$\left[-1.96\sqrt{\frac{1}{N}\sum_{i=0}^{k-1} r_i^2},\; 1.96\sqrt{\frac{1}{N}\sum_{i=0}^{k-1} r_i^2}\right] \tag{17-12}$$

under the tacit assumption that ρ_k, ρ_{k+1}, ρ_{k+2}, etc. are all zero. That assumption may or may not be remotely reasonable, but it is probably the only reasonable assumption to build into the software.

Of course, using Equation 17-11, one can easily compute one's own Bartlett-based asymptotic 95% confidence interval for ρ_k as

Bartlett 95% Confidence Interval for ρ_k
(Assuming $\rho_m, \rho_{m+1}, \rho_{m+2}, \ldots$ are all zero.)

$$\left[r_k - 1.96\sqrt{\frac{1}{N}\sum_{i=1}^{m} r_i^2},\; r_k + 1.96\sqrt{\frac{1}{N}\sum_{i=1}^{m} r_i^2}\right] \tag{17-13}$$

which is asymptotically valid so long as $Y_1 \ldots Y_N$ are covariance stationary and (jointly) normally distributed, with ρ_k equal to zero for values of k greater than a value of m one would have to choose. In practice, the squared sample autocorrelations become small at large lags – i.e., for large values of k – for data on time-series which are clearly covariance stationary, so the value of m one chooses to use in estimating such a confidence interval is actually not critical, so long as one makes it fairly large.[25]

The sample autocorrelation (R_k) is a consistent estimator of ρ_k for any covariance stationary time-series – and follows the asymptotic sampling distribution of Equation 17-11 for jointly normal time-series – but R_k is not generally an unbiased estimator. A bit is known about this finite-sample bias in R_k, however. In particular, Marriot and Pope (1954) looked at the *asymptotic* bias in R_k for time-series generated by several simple processes. For jointly normal Y_t generated by an AR(1) model with parameter value φ_1, they find that R_1, for example, is biased downward in amount equal to (1+ $3\rho_1$)/N in large samples.[26] Based on these results (and considerable experience with simulated data) it seems likely that R_k is always biased downward – i.e., toward zero – at low lags when the time-series is positively autocorrelated at low lags, and probably in greater degree for small N than the Marriott-Pope result suggests.

Finally, Bartlett (1946) also provides an expression for the correlation between the sampling errors made by R_k and those made by $R_{k+\ell}$ – i.e., at a different lag – assuming, as before, that the sample is large and that the time-series Y_t is both covariance stationary and jointly normally distributed. Their general expression for this correlation is a bit opaque, but it is worthwhile looking at a special case of it – which is by no means misleading – in which the difference in the two lags is one and the population autocorrelations are assumed to all be zero for lags larger than three. In that case, the Bartlett result is

$$\operatorname{corr}(R_k, R_{k+1}) = 2\frac{(\rho_1 + \rho_1\rho_2 + \rho_2\rho_3)}{(1 + 2\rho_1^2 + 2\rho_2^2 + 2\rho_3^2)}. \tag{17-14}$$

add twos here

[25] The reason for this is that the value of $(\rho_k)^2$ must dwindle to zero for large k in order for the variance of the time-series to be bounded; this will be demonstrated in Section 17.5, when the MA(∞) representation of a time-series is examined.

[26] Again, see Equation 13-11 in Section 13.4 for the definition of the AR(1) model (and φ_1); autoregressive models are analyzed more completely in Section 17.6.

p546

Finally, this simple bilinear model also illustrates a fundamental source of another feature of many economic time-series: autoregressive conditional heteroscedasticity, or "ARCH." The homoscedasticity assumption in a regression model is the assumption that the variance of the model error term is a constant, whereas heteroscedasticity is the failure of this assumption. Here the issue is homoscedasticity or heteroscedasticity of the time-series itself, which is the dependent variable in the model. The assumption that this dependent variable is strictly stationary (and hence also covariance stationary) implies that the *unconditional* variance of the time-series is a constant over time; hence, the time-series is unconditionally homoscedastic. This result, however, does not rule out the possibility that the *conditional* variance of a time-series – and, in particular, its variance conditional on having observed its own recent past – might vary over time.

Indeed, it has become well known in the last 15 years that the variance of the daily and monthly returns to many financial asset time-series vary over time and, moreover, vary in such a way as to be positively correlated with the magnitude of recent returns fluctuations. Predicting the variance of a financial return time-series is crucial to pricing related financial assets – "options" and the like – so a great deal of effort has gone into producing such variance prediction models. One of the most popular of these is the GARCH(p, q) model, where the current variance in the model error term is taken to be a function of the squares of model errors in the recent past:

$$\begin{aligned} Y_t &= \mu + \sqrt{h_t}\, U_t \\ h_t &= \alpha_o + \sum_{i=1}^{q} \alpha_i\, U_{t-i}^2 + \sum_{j=1}^{q} \beta_j\, h_{t-j} \end{aligned} \tag{18-18}$$

Here U_t is taken to be serially uncorrelated – with mean zero and unit variance – and the variance of Y_t, conditional on its own past, is then given by the value of h_t. Much attention has gone into estimating and utilizing GARCH(p, q) models, and a number of variations on this theme have been elaborated and applied; Enders (2010, Chapter 3) provides an excellent introduction to this literature.[41]

The GARCH(p, q) model (and its variants) are usually considered to be nonlinear models in their own right, but usually do not allow for nonlinear serial dependence in the conditional mean of Y_t – only in its conditional variance – and hence are generally not helpful in forecasting Y_t itself. This is because the GARCH(p, q) formulation is fundamentally a "tack on" to the model for Y_t, focusing solely on the variance of the model error term. In contrast, it is easy to show that the bilinear model of Equation 18-13 *endogenously* induces positive conditional heteroscedasticity in Y_t^{BILINEAR}:[42]

$$\begin{aligned} \text{var}\left(Y_{N+1}^{\text{BILINEAR}} \middle| y_N, y_{N-1}, \ldots\right) &= 1 \\ \text{var}\left(Y_{N+2}^{\text{BILINEAR}} \middle| y_N, y_{N-1}, \ldots\right) &= \beta^2 y_N^2 + 1 \end{aligned} \tag{18-19}$$

Thus, if Y_t is generated by Equation 18-13, then the conditional variance of Y_t one period hence is still a constant, but the conditional variance of Y_t two periods hence is automatically a direct

[41] Note that the parameters $\alpha_1 \ldots \alpha_q$ are generally constrained to be positive, the parameter p is often set to one, and the estimated value of the parameter β_1 is usually close to one: this yields values of h_t which are positive and smoothly varying over time. See also: Heracleous and Spanos (2006) for a fundamentally better approach. Enders, W. (2010), *Applied Econometric Time-Series*, Wiley: Hoboken. Heracleous, M., and A. Spanos (2006), "The Student's t Dynamic Linear Regression: Re-examining Volatility Modeling," in *Econometric Analysis of Financial and Economic Time-Series (Part A); Advances in Econometrics 20*. Elsevier: Amsterdam.

[42] The derivation of Equation 18-19 is not very difficult, but a bit too challenging for a chapter exercise here. Consequently, this derivation – and the generalization of this result ~~about the endogenous~~ origin of conditional heteroscedasticity in nonlinear models – are both left to Ashley (2010). ~~Ashley, R. (2010).~~ "On the Origins of Conditional Heteroscedasticity in Time-Series," ~~available at Web sites econpapers.repec.org/paper/vpiwpaper/e07-23.htm and ashleymac.econ.vt.edu/working_papers/origins_of_conditional_heteroscedasticity.pdf~~. Ashley (2012), Korean Economic Review 28, 1-5.

for footnote #42
page 630 in chapter 18

Ashley, R., M. J. Hinich, and D. M. Patterson. 1990. Nonlinear Serial Dependence in Industrial Stock Returns. In Advances in Mathematical Programming and Financial Planning 2. (K. D. Lawrence, Guerard, J.B., and Reeves, G.R., eds.) Vol. 2. JAI Press, London. p. 163-182.

PUBLICATIONS (JOURNAL ARTICLES)

Ashley, R. 2012. "On the Origins of Conditional Heteroscedasticity in Time Series" *Korean Economic Review 28*: 1-5.
url: http://ashleymac.econ.vt.edu/working_papers/origins_of_conditional_heteroscedasticity.pdf

Ashley, R. and H. Ye. 2012. "On the Granger Causality Between Median Inflation and Price Dispersion" *Applied Economics 44*, 4221-4238.
url: http://ashleymac.econ.vt.edu/working_papers/price_dispersion_causality.pdf

Ashley, R., S. Ball and C. Eckel. 2010. "Motives for Giving: A Re-Analysis of Two Classic Public Goods Experiments" *Southern Economic Journal 77*: 15-26.
url: http://ashleymac.econ.vt.edu/working_papers/E2004_1.pdf.

Ashley, R., and D.M. Patterson. 2010. "Apparent Long Memory in Time Series as an Artifact of a Time-Varying Mean: Considering Alternatives to the Fractionally Integrated Model." *Macroeconomic Dynamics 14:* 59-87. url: http://ashleymac.econ.vt.edu/working_papers/long_memory.pdf.

Ashley, R., and D.M. Patterson. 2010. "A Test of the GARCH(1,1) Specification for Daily Stock Returns." *Macroeconomic Dynamics 14:* 137-144.
url: http://ashleymac.econ.vt.edu/working_papers/NLInsight_GARCH.pdf

Ashley, R., and R. Verbrugge. 2009. "To Difference or Not to Difference: A Monte Carlo Investigation of Inference in Vector Autoregression Models." *International Journal of Data Analysis Techniques and Strategies1(3)*: 242-274. url: http://ashleymac.econ.vt.edu/working_papers/varsim.pdf.

Ashley, R. 2009. "Assessing the Credibility of Instrumental Variables Inference with Imperfect Instruments via Sensitivity Analysis." *Journal of Applied Econometrics 24*: 325-337.
url: http://ashleymac.econ.vt.edu/working_papers/E2003_8.pdf

Ashley, R., and R. Verbrugge. 2009. "Frequency Dependence in Regression Model Coefficients: An Alternative Approach for Modeling Nonlinear Dynamic Relationships." *Econometric Reviews 28*: 4-20.
url: http://ashleymac.econ.vt.edu/working_papers/freq_depend.pdf.

Rusticelli, E., R. Ashley, E. B. Dagum, and D.M. Patterson. 2009. "A New Bispectral Test for Nonlinear Serial Dependence." *Econometric Reviews 28*: 279-293.
url: http://ashleymac.econ.vt.edu/working_papers/maximal_bispectral_9_05.pdf.

Ashley, R. 2008. "Growth May Be Good for the Poor, But Decline Is Disastrous: on the Non-Robustness of the Dollar-Kraay Result." *International Review of Economics and Finance* 17: 333-338.
url: http://ashleymac.econ.vt.edu/working_papers/growthgood.pdf.

Ashley, R., and R. Verbrugge. 2006. "Comments on 'A Critical Investigation on Detrending Procedures for Non-Linear Processes." *Journal of Macroeconomics* 28:192-194.

Ashley, R., and D.M. Patterson. 2006. Evaluating the Effectiveness of State-Switching Models for U.S. Real Output" *Journal of Business and Economic Statistics 24(3):*266-77.

erval estimates and *p*-values for hypothesis tests. But OLS does not provide an
^{2}; the estimator S^2 is tacked on separately.
he MLE framework generates an estimator for σ^2 in a graceful and natural way. In the
le – but now no longer artificially assuming that the value of σ^2 is known – σ^2 in the
function is simply replaced by $\hat{\sigma}^2_{\text{guess}}$ and $L(y_1 \dots y_N; \hat{\beta}^{\text{guess}}, \hat{\sigma}^2_{\text{guess}}, x^1 \dots x_N)$ is now
ith respect to both $\hat{\beta}^{\text{guess}}$ and $\hat{\sigma}^2_{\text{guess}}$. The MLE estimator of σ^2 is thus just the value of

ch

$$\left.\frac{\hat{\beta}^{\text{guess}}, \hat{\sigma}^2_{\text{guess}}, x_1 \dots x_N)\Bigr)}{\partial\hat{\sigma}^2_{\text{guess}}}\right)_{\hat{\beta}^{\text{guess}} = \hat{\beta}^{\text{MLE}}} \tag{19-8}$$

$$= \frac{\partial}{\partial\hat{\sigma}^2_{\text{guess}}}\left[-\frac{N}{2}\ln\left(2\pi\hat{\sigma}^2_{\text{guess}}\right) - \frac{1}{2\hat{\sigma}^2_{\text{guess}}}\sum_{i=1}^{N}\left(y_i - \hat{\beta}^{\text{guess}}x_i\right)^2\right]$$

19-2b

After just a bit of algebra (Exercise 19-1), this yields the MLE estimator of σ^2:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}^{\text{MLE}}x_i)^2 = \frac{1}{N}\text{SSE} = \left(\frac{N-1}{N}\right)S^2 \tag{19-9}$$

Maximum Likelihood Estimation framework leads to an estimator of σ^2 in a coherent,
onsistent way. Unlike S^2 – which was constructed so as to be an unbiased estimator of
is clearly biased; but $\hat{\sigma}^2_{\text{MLE}}$ turns out to have very nice large-sample (asymptotic)

this is the second big advantage of MLE over OLS estimation: the MLE estimator of *any*
ameter – call it θ – can be shown to have every good asymptotic property one could think
virtue of being a maximum likelihood estimator:[4]

is a consistent and (asymptotically) unbiased estimator of θ.

is an asymptotically efficient estimator of θ – that is, there is no consistent estimator of θ
a smaller asymptotic variance than that of $\hat{\theta}_{\text{MLE}}$.

asymptotic sampling distribution of $\hat{\theta}_{\text{MLE}}$ is easy to obtain, because $\hat{\theta}_{\text{MLE}}$ is (asymptoti-
) normally distributed with (asymptotic) mean θ and (asymptotic) sampling variance equal
$1/E[\partial^2 L(\theta)/\partial\theta^2]$, which is usually easy to calculate – see Exercises 19-1 and 19-2 for
trative examples.[5]

$\cdot$) is a continuous function, then $g(\hat{\theta}_{\text{MLE}})$ is the maximum likelihood estimator of $g(\theta)$, and
ce partakes of all the nice asymptotic properties listed above.

971, Sections 8.4 and 8.5) gives a particularly accessible derivation of all of the results given below for the scalar
-parameter) case. He also discusses the very mild regularity conditions (basically on the smoothness of the
g density function) which are needed to support these results; these are all satisfied by regression models with
distributed errors. Except for the more sophisticated notation, these conditions (and results) are very similar where
fficients are estimated, except that the value of *k* must be finite; that is, *k* cannot rise with the sample length, *N*. Theil,
), *Principles of Econometrics*, Wiley: New York.

at the log-likelihood function in the foregoing expression is in terms of the population parameter, θ – e.g., see
19-1b. The indicated second partial derivative is generally easy to obtain whenever (as here) one has an analytic
on for $L(\theta)$; its dependence on the model errors is also usually simple, so evaluating the expectation is generally easy.
d asymptotic variances for MLE estimators are readily obtained where $L(\hat{\theta})$ must be maximized numerically, using
esults.

$$
\begin{aligned}
\mathrm{Prob}(Y_i = 1|x_i,\alpha,\beta) &= \mathrm{Prob}(\tilde{Y}_i > 0|x_i,\alpha,\beta) \\
&= \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(\tilde{Y}_i-\alpha-\beta x_i)^2}{2}} d\tilde{Y} \\
&= \int_{-(\alpha+\beta x_i)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\
&= 1 - F(\alpha \mp \beta x_i)
\end{aligned}
\tag{19-11}
$$

where $F(\cdot)$ is the cumulative density function for the unit normal distribution.

In contrast, if y_i is zero, then the latent variable must be non-positive, which (according to Equation 19-10) occurs with probability equal to the probability that $\alpha + \beta x_i + U_i$ is non-positive. This probability is

$$
\begin{aligned}
\mathrm{Prob}(Y_i = 0|x_i,\alpha,\beta) &= \mathrm{Prob}(\tilde{Y}_i \le 0|x_i,\alpha,\beta) \\
&= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{\frac{-(\tilde{Y}_i-\alpha-\beta x_i)^2}{2}} d\tilde{Y} \\
&= \int_{-\infty}^{-\alpha \mp \beta x_i} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\
&= F(\alpha \mp \beta x_i)
\end{aligned}
\tag{19-12}
$$

Thus, the likelihood of observing the value y_i can be written as

$$
\{F(\overset{-\alpha}{\alpha} \mp \beta x_i)\}^{1-y_i}\{1 - F(\overset{-\alpha}{\alpha} \mp \beta x_i)\}^{y_i}
\tag{19-13}
$$

Hence, replacing α by $\hat{\alpha}^{\text{guess}}$ and β by $\hat{\beta}^{\text{guess}}$, the log-likelihood of observing the entire sample, $y_i \dots y_N$, is

same

$F(-\hat{\alpha}^{guess} - \hat{\beta}^{guest} x_i)$

$$
\begin{aligned}
&L\left(y_1 \dots y_N; \hat{\alpha}^{\text{guess}}, \hat{\beta}^{\text{guess}}, x_1 \dots x_N\right) \\
&= \sum_{i=1}^{N} \left[(1-y_i)\ln\left\{F(\hat{\alpha}^{\text{guess}} \mp \hat{\beta}^{\text{guess}} x_i)\right\} + y_i \ln\left\{1 - \cancel{F(\hat{\alpha}^{\text{guess}} \mp \hat{\beta}^{\text{guess}} x_i)}\right\}\right]
\end{aligned}
\tag{19-14}
$$

Excellent computing approximations for the cumulative distribution function of the normal distribution (and its derivatives) are readily available and $L\left(y_1 \dots y_N; \hat{\alpha}^{\text{guess}}, \hat{\beta}^{\text{guess}}, x_1 \dots x_N\right)$ is very well-behaved as an optimand, so it is not difficult to numerically maximize it over $\hat{\alpha}^{\text{guess}}$ and $\hat{\beta}^{\text{guess}}$ to obtain $\hat{\alpha}^{\text{MLE}}$ and $\hat{\beta}^{\text{MLE}}$. This procedure is called "Probit Regression."

The estimation of such probit regression models is extremely easy in practice, because all of the numerical manipulations for maximizing log-likelihood functions like Equation 19-14 are already programmed into modern econometric software. For example, the Stata command for estimating the model of Equation 19-10 is just "probit *y x*". The Stata syntax for including additional explanatory variables (e.g., $z1_i$ and $z2_i$) in the latent variable model is straightforward: "probit *y x z*1 *z*2".