4.5

We can compare estimators based on their asymptotic variances. The complication in comparing two consistent estimators is that both converge to the true parameter as the sample size increases. Moreover, it usually happens (as in our example 4.9), that they converge at the same rate – that is, in both cases, the asymptotic variance of the two estimators are of the same order, such as $O(1/n)$. In such a situation, we can sometimes compare the asymptotic variances for the same n to resolve the ranking. The least absolute deviations estimator as an alternative to least squares provides an example.

AU: Is capital Roman "Oh" OK?

Example 4.5 Least Squares vs. Least Absolute Deviations – A Monte Carlo Study

We noted earlier (Section 4.2) that while it enjoys several virtues, least squares is not the only available estimator for the parameters of the linear regresson model. Least absolute deviations (LAD) is an alternative. (The LAD estimator is considered in more detail in Section 7.3.1.) The LAD estimator is obtained as

$$\mathbf{b}_{LAD} = \text{the minimizer of } \sum_{i=1}^{n} | y_i - \mathbf{x}_i'\mathbf{b}_0 |,$$

in contrast to thelinear least squares estimator,

$$\mathbf{b}_{LS} = \text{the minimizer of } \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b}_0)^2.$$

Suppose the regression model is defined by

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

conditional

where the distribution of ε_i has conditional mean zero, constant variance σ^2 and *median* zero as well – the distribution is symmetric – and $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. That is, all the usual regression assumptions, but with the normality assumption replaced by symmetry of the distribution. Then, under our assumptions, $\mathbf{b}_{LS}$ is a consistent and asymptotically normally distributed estimator with asymptotic covariance matrix given in Theorem 4.4, which we will call $\sigma^2\mathbf{A}$. As Koenker and Bassett (1978, 1982), Huber (1987), Rogers (1993) and Koenker (2005) have discussed, under these assumptions, $\mathbf{b}_{LAD}$ is also consistent. A good estimator of the asymptotic variance of $\mathbf{b}_{LAD}$ would be $(1/2)^2[1/f(0)]^2\mathbf{A}$ where $f(0)$ is the density of ε at its median, zero. This means that we can compare these two estimators based on their asymptotic variances. The ratio of the asymptotic variance of the k'th element of $\mathbf{b}_{LAD}$ to the corresponding element of $\mathbf{b}_{LS}$ would be

$$q_k = \text{Var}(b_{k,LAD})/\text{Var}(b_{k,LS}) = (1/2)^2(1/\sigma^2)[1/f(0)]^2.$$

If ε actually did have a normal distribution with mean (and median) zero, then

minus

$$f(\varepsilon) = (2\pi\sigma^2)^{-1/2}\exp(-\varepsilon^2/(2\sigma^2))$$

so $f(0) = (2\pi\sigma^2)^{-1/2}$ and for this special case $q_k = \pi/2$. Thus, if the disturbances are normally distributed, then LAD will be asymptotically less efficient by a factor of $\pi/2 = 1.573$.

The usefulness of the LAD estimator arises precisely in cases in which we cannot assume normally distributed disturbances. Then, it becomes unclear which is the better estimator. It has been found in a long body of research that the advantage of the LAD estimator is most likely to appear in small samples when the distribution of ε has thicker tails than the normal – that is, when

outlying values of y_i are more likely. As the sample size grows larger, one can expect the LS estimator to regain its superiority. We will explore this aspect of the estimator in a small **Monte Carlo study**.

Examples 2.6 and 3.4 note an intriguing feature of the fine art market. At least in some settings, large paintings sell for more at auction than small ones. Appendix Table 4.1 contains the sale prices, widths and heights of 430 Monet paintings. These paintings sold at auction for prices ranging from $10,000 up to as much as $33 million. A linear regression of the log of the price on a constant term, the log of the surface area and the aspect ratio produces the results in the top line of Table 4.4. This is the focal point of our analysis. In order to study the different behaviors of the LS and LAD estimators, we will do the following Monte Carlo study[7]: We will draw without replacement 100 samples of R observations from the 430. For each of the 100 samples, we will compute $\mathbf{b}_{LS,r}$ and $\mathbf{b}_{LAD,r}$. We then compute the average of the 100 vectors and the sample variance of the 100 observations.[8] The sampling variability of the 100 sets of results corresponds to the notion of "variation in repeated samples." For this experiment, we will do this for R = 10, 50 and 100. The overall sample size is fairly large, so it is reasonable to take the full sample results as at least approximately the "true parameters." The standard errors reported for the full sample LAD estimator are computed using **bootstrapping**. Briefly, the procedure is carried out by drawing B - we used B = 100 - samples of n (430) observations *with replacement*, from the full sample of n observations. The estimated variance of the LAD estimator is then obtained by computing the mean squared deviation of these B estimates around the full sample LAD estimates (not the mean of the B estimates). This procedure is discussed in detail in Section 15.6.

If the assumptions underlying our regression model are correct, we should observe the following:

(1) Since both estimators are consistent, the averages should resemble the main results above, the more so as R increases.
(2) As R increases, the sampling variance of the estimators should decline.
(2) We should observe generally that the standard deviations of the LAD estimates are larger than the corresponding values for the LS estimator.
(3) When R is small, the LAD estimator should compare more favorably to the LS estimator, but as R gets larger, any advantage of the LS estimator should become apparent.

[7] Being a Monte Carlo study that uses a random number generator, there is a question of replicability. The study was done with NLOGIT and is replicable. The program can be found on the website for the text. The qualitative results, if not the precise numerical values, can be reproduced with other programs that allow random sampling from a data set.

[8] Note that the sample size R is not a negligible fraction of the population size, 430 for each replication. However, this does not call for a finite population correction of the variances in Table 4.8. We are not computing the variance of a sample of R observations drawn from a population of 430 paintings. We are computing the variance of a sample of R statistics each computed from a different subsample of the full population. There are a bit less than 10^{20} different samples of 10 observations we can draw. The number of different samples of 50 or 100 is essentially infinite.

A kernel density estimate for the distribution of the least squares residuals appears in Figure 4.4. There is a bit of skewness in the distribution, so a main assumption underlying our experiment may be violated to some degree. Results of the experiments are shown in Table 4.4. The force of the asymptotic results can be seen most clearly in the column for the coefficient on logArea. The decline of the standard deviation as R increases is evidence of the consistency of both estimators. In each pair of results (LS,LAD), we can also see that the estimated standard deviation of the LAD estimator is greater by a factor of about 1.2 to 1.4, which is also to be expected. Based on the normal distribution, we would have expected this ratio to be $\sqrt{1.573} = 1.254$.

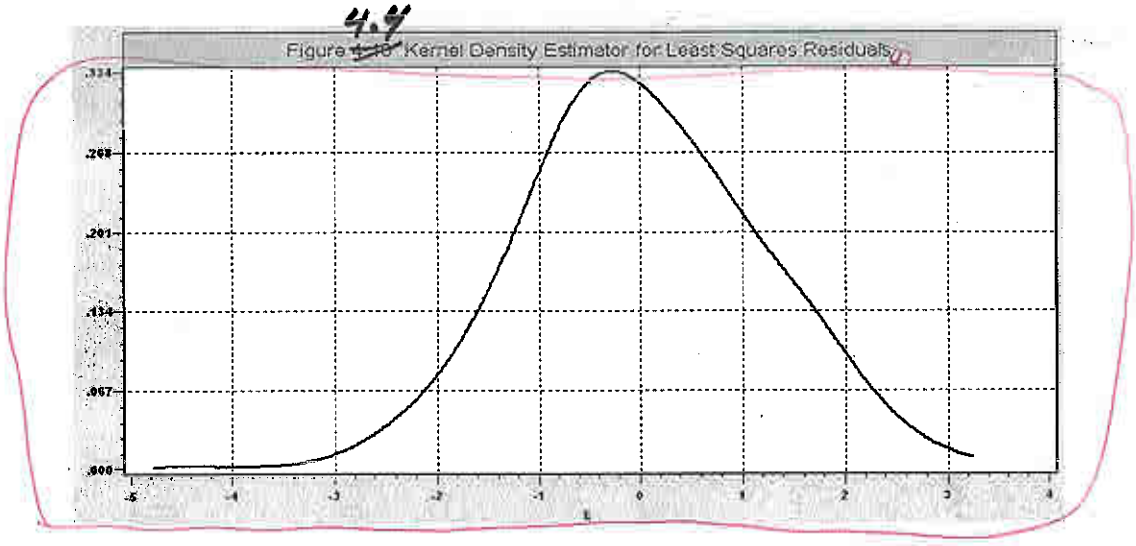


TABLE 4.4 Estimated Equations for Art Prices

	Constant		Log Size		Aspect Ratio	
Full Sample	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
LS	-8.42653	0.61184	1.33372	0.09072	-0.16537	0.12753
LAD	-7.62436	0.89055	1.20404	0.13626	-0.21260	0.13628
R = 10						
LS	-9.39384	6.82900	1.40481	1.00545	0.39446	2.14847
LAD	-8.97714	10.24781	1.34197	1.48038	0.35842	3.04773
R=50						
LS	-8.73099	2.12135	1.36735	0.30025	-0.06594	0.52222
LAD	-8.91671	2.51491	1.38489	0.36299	-0.06129	0.63205
R=100						
LS	-8.36163	1.32083	1.32758	0.17836	-0.17357	0.28977
LAD	-8.05195	1.54190	1.27340	0.21808	-0.20700	0.29465

4.4.6 MAXIMUM LIKELIHOOD ESTIMATION

We have motivated the least squares estimator in two ways: First, we obtained Theorem 4.1 which states that the least squares estimator mimics the coefficients in the minimum mean squared error predictor of y in the joint distribution of y and $\mathbf{x}$. Second, Theorem 4.2, the Gauss-Markov Theorem states that the least squares estimator is the *minimum variance linear unbiased* estimator of $\boldsymbol{\beta}$ under the assumptions of the model. Neither of these results relies on Assumption A6, normality of the distribution of ε. A natural question at this point would be what is the role of this assumption? There are two. First, the assumption of normality will produce the basis for determining the appropriate endpoints for confidence intervals in Sections 4.5 and 4.6. But, we found in Section 4.4.2 that based on the central limit theorem, we could base inference on the asymptotic normal distribution of $\mathbf{b}$, even if the disturbances were not normally distributed. That would seem to make the normality assumption no longer necessary, which is largely true, but for a second result.

If the disturbances are normally distributed, then the least squares estimator is also the **maximum likelihood estimator (MLE)**. We will examine maximum likelihood estimation in detail in Chapter ~~13~~ 14, so we will describe it only briefly at this point. The end result is that by virtue of being an MLE, least squares is *asymptotically efficient among consistent and asymptotically normally distributed estimators.* This is a large sample counterpart to the Gauss-Markov theorem (known formally as the Cramér-Rao bound). What the two theorems have in common is that they identify the least squares estimator as the most efficient estimator in the assumed class of estimators. They differ in the class of estimators assumed:

Gauss-Markov:	Linear and unbiased estimators;
ML:	Based on normally distributed disturbances, consistent and asymptotically normally distributed estimators.

These are not "nested." Notice, for example, that the MLE result does not require unbiasedness or linearity. Gauss-Markov does not require normality or consistency. The Gauss-Markov Theorem is a finite sample result while the Cramér-Rao bound is an asymptotic (large sample) property. The important aspect of the development concerns the efficiency property. Efficiency, in turn, relates to the question of how best to use the sample data for statistical inference. In general, it is difficult to establish that an estimator is efficient without being specific about the candidates. The Gauss-Markov theorem is a powerful result for the linear regression model. However, it has no counterpart in any other modeling context, so once we leave the linear model, we will require different tools for comparing estimators. The principle of maximum likelihood allows the analyst to assert asymptotic efficiency for the estimator, but only for the specific distribution assumed. Example 4.6 establishes that $\mathbf{b}$ is the MLE in the regression model with normally distributed disturbances. Example 4.7 then considers a case in which the regression disturbances are not normally distributed and, consequently, $\mathbf{b}$ is less efficient than the MLE.

Example 4.6 MLE with Normally Distributed Disturbances

With normally distributed disturbances, $y_i|\mathbf{x}_i$ is normally distributed with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and variance σ^2, so the density of $y_i|\mathbf{x}_i$ is

$$f(y_i|\mathbf{x}_i) = \frac{\exp\left[-\frac{1}{2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right]}{\sqrt{2\pi\sigma^2}}$$

The log likelihood for a sample of n independent observations is equal to the log of the joint density of the observed random variables. For a random sample, the joint density would be the product, so the log likelihood, given the data, which is written $\ln L(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})$ would be the sum of the logs of the densities. This would be (after a bit of manipulation),

$$\ln L(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = -(n/2)[\ln\sigma^2 + \ln 2\pi + (1/\sigma^2)\frac{1}{n}\Sigma_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2].$$

The values of $\boldsymbol{\beta}$ and σ^2 that maximize this function are the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2. As we will explore further in Chapter 14, the functions of the data that maximize this function with respect to $\boldsymbol{\beta}$ and σ^2 are the least squares coefficient vector, $\mathbf{b}$, and the mean squared residual, $\mathbf{e}'\mathbf{e}/n$. Once again, we leave for Chapter 14 a derivation of the following result,

$$\text{Asy.Var}\left[\hat{\boldsymbol{\beta}}_{ML}\right] = -E[\partial^2 \ln L/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}']^{-1} = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}],$$

which is exactly what appears in Section 4.3.6. This shows that the least squares estimator is the maximum likelihood estimator. It is consistent, normally distributed, and, under the assumption of normality, by virtue of Theorem 14.4, asymptotically efficient.

It is important to note that the properties of an MLE depend on the specific distribution assumed for the observed random variable. If some nonnormal distribution is specified for ε and it emerges that $\mathbf{b}$ is not the MLE, then least squares may not be efficient. The following case illustrates.

In Chapter 16, we will show that if the disturbances are normally distributed, then the least squares estimator is also the **maximum likelihood estimator**. Maximum likelihood estimators are asymptotically efficient among consistent and asymptotically normally distributed estimators. This gives us a partial result, albeit a somewhat narrow one since to claim it, we must assume normally distributed disturbances. If some other distribution is specified for $\boldsymbol{\varepsilon}$ and it emerges that $\mathbf{b}$ is not the maximum likelihood estimator, then least squares may not be efficient.

Example 4.7 The Gamma Regression Model

Greene (1980a) considers estimation in a regression model with an asymmetrically distributed disturbance,

$$y = (\alpha + \sigma\sqrt{P}) + \mathbf{x}'\boldsymbol{\beta} + (\varepsilon - \sigma\sqrt{P}) = \alpha^* + \mathbf{x}'\boldsymbol{\beta} + \varepsilon^*,$$

where ε has the gamma distribution in Section B.4.5 [see (B-39)] and $\sigma = \sqrt{P}/\lambda$ is the standard deviation of the disturbance. In this model, the covariance matrix of the least squares estimator of the slope coefficients (not including the constant term) is,

$$\text{Asy. Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1},$$

whereas for the maximum likelihood estimator (which is not the least squares estimator),[14]

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_{ML}] \approx [1 - (2/P)]\sigma^2(\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1}.$$

But for the asymmetry parameter, this result would be the same as for the least squares estimator. We conclude that the estimator that accounts for the asymmetric disturbance distribution is more efficient asymptotically.

4.9.6 MORE GENERAL DATA GENERATING PROCESSES

The asymptotic properties of the estimators in the classical regression model were established under the following assumptions:

- **A1.** ***Linearity:*** $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$.
- **A2.** ***Full rank:*** The $n \times K$ sample data matrix, $\boldsymbol{X}$ has full column rank.
- **A3.** ***Exogeneity of the independent variables:*** $E[\varepsilon_i \mid x_{j1}, x_{j2}, \ldots, x_{jK}] = 0$, $i, j = 1, \ldots, n$.
- **A4.** ***Homoscedasticity and nonautocorrelation.***
- **A5.** ***Data generating mechanism-independent observations.***

The following are the crucial results needed: For consistency of $\mathbf{b}$, we need (4-21) and (4-24),

$$\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \text{plim}\,\overline{\mathbf{Q}}_n = \mathbf{Q}, \quad \text{a positive definite matrix,}$$

$$\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \text{plim}\,\overline{\mathbf{w}}_n = E[\overline{\mathbf{w}}_n] = \mathbf{0}.$$

[14] The matrix $\mathbf{M}^0$ produces data in the form of deviations from sample means. (See Section A.2.8.) In Greene's model, P must be greater than 2.

Another example that is somewhat similar to the model in Example 4.7 is the stochastic frontier model developed in Chapter 18. In these two cases in particular, the distribution of the disturbance is asymmetric. The maximum likelihood estimators are computed in a way that specifically accounts for this while the least squares estimator treats observations above and below the regression line symmetrically. That difference is the source of the asymptotic advantage of the MLE for these two models.

4.5 Interval Estimation

The objective of interval estimation is to present the best estimate of a parameter with an explicit expression of the uncertainty attached to that estimate. A general approach, for estimation of a parameter θ, would be

$$\hat{\theta} \pm \text{sampling variability}. \qquad \textbf{(4-37)}$$

(We are assuming that the interval of interest would be symmetic around $\hat{\theta}$.) Following the logic that the range of the sampling variability should convey the degree of (un)certainty, we consider the logical extremes. We can be absolutely (100%) certain that the true value of the parameter we are estimating lies in the range $\hat{\theta} \pm \infty$. Of course, this is not particularly informative. At the other extreme, we should place no certainty (0%) on the range $\hat{\theta} \pm 0$. The probability that our estimate precisely hits the true parameter value should be considered zero. The point is to choose a value of α – 0.05 or 0.01 is conventional – such that we can attach the desired confidence (probability), $100(1-\alpha)\%$, to the interval in (4-37). We consider how to find that range, then apply the procedure to three familiar problems, interval estimation for one of the regression parameters, estimating a function of the parameters and predicting the value of the dependent variable in the regression using a specific setting of the independent variables. For this purpose, we depart from assumption A6 that the disturbances are normally distributed. We will then relax that assumption and rely instead on the asymptotic normality of the estimator.

4.5.1 Forming a Confidence Interval for a Coefficient

From(4-18), we have that $\mathbf{b}|\mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$. It follows that for any particular element of $\mathbf{b}$, say b_k,

$$b_k \sim N[\beta_k, \sigma^2 S^{kk}]$$

where S^{kk} denotes the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. By standardizing the variable, we find

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \qquad \textbf{(4-38)}$$

has a standard normal distribution. Note that z_k which is a function of b_k, β_k, σ^2 and S^{kk}, nonetheless has a distribution that involves none of the model parameters or the data; z_k is a **pivotal statistic**. Using our conventional 95% confidence level, we know that Prob[$-1.96 \leq z_k \leq 1.96$]. By a simple manipulation, we find that

$$\text{Prob}\left[b_k - 1.96\sqrt{\sigma^2 S^{kk}} \leq \beta_k \leq b_k + 1.96\sqrt{\sigma^2 S^{kk}}\right] = 0.95. \qquad \textbf{(4-39)}$$

Note that this is a statement about the probability that the random interval $b_k \pm$ the sampling variability contains β_k, not the probability that β_k lies in the specified interval. If we wish to use some other level of confidence, not 95%, then the 1.96 in (4-39) is replaced by the appropriate $z_{(1-\alpha/2)}$. (We are using the notation $z_{(1-\alpha/2)}$ to denote the value of z such that for the standard normal variable z, Prob[$z \leq z_{(1-\alpha/2)}$] = 1 - $\alpha/2$. Thus, $z_{0.975} = 1.96$, which corresponds to $\alpha = 0.05$.)

We would have our desired confidence interval in (4-39), save for the complication that σ^2 is not known, so the interval is not operational. It would seem natural to use s^2 from the regression. This is, indeed, an appropriate approach. The quantity

$$\frac{(n-K)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \qquad \textbf{(4-40)}$$

is an idempotent quadratic form in a standard normal vector, $(\boldsymbol{\varepsilon}/\sigma)$. Therefore, it has a chi-squared distribution with degrees of freedom equal to the rank($\mathbf{M}$) = trace($\mathbf{M}$) = n-K. (See Section B11.4 for the proof of this result.) The chi-squared variable in (4-40) is independent of the standard normal variable in (4-38). To prove this, it suffices to show that

$$\left(\frac{\mathbf{b}-\boldsymbol{\beta}}{\sigma}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$$

is independent of $(n-K)s^2/\sigma^2$. In Section B.11.7 (Theorem B.12), we found that a sufficient condition for the independence of a linear form $\mathbf{Lx}$ and an idempotent quadratic form $\mathbf{x}'\mathbf{Ax}$ in a standard normal vector $\mathbf{x}$ is that $\mathbf{LA} = \mathbf{0}$. Letting $\boldsymbol{\varepsilon}/\sigma$ be the $\mathbf{x}$, we find that the requirement here would be that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M} = \mathbf{0}$. It does, as seen in (3-15). The general result is central in the derivation of many test statistics in regression analysis.

THEOREM 4.4 Independence of b and s^2

If $\boldsymbol{\varepsilon}$ is normally distributed, then the least squares coefficient estimator $\mathbf{b}$ is statistically independent of the residual vector $\mathbf{e}$ and therefore, all functions of $\mathbf{e}$, including s^2.

Therefore, the ratio

$$t_k = \frac{(b_k - \beta_k)/\sqrt{\sigma^2 S^{kk}}}{\sqrt{[(n-K)s^2/\sigma^2]/(n-K)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \qquad (4\text{-}13)$$

has a t distribution with $(n-K)$ degrees of freedom.[10] We can use t_k to test hypotheses or form confidence intervals about the individual elements of $\boldsymbol{\beta}$.

The result in (4-41) differs from (4-38) in the use of s^2 instead of σ^2, and in the pivotal distribution, t with $(n-K)$ degrees of freedom, rather than standard normal. It follows that a confidence interval for β_k can be formed using

$$\text{Prob}\left[b_k - t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}} \le \beta_k \le b_k + t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}}\right] = 1 - \alpha, \qquad (4\text{-}42)$$

where $t_{(1-\alpha/2),[n-K]}$ is the appropriate critical value from the t distribution. Here, the distribution of the pivotal statistic depends on the sample size through $(n-K)$, but, once again, not on the parameters or the data. The practical advantage of (4-42) is that it does not involve any unknown parameters. A confidence interval for β_k can be based on (4-42)

[10] See (B-36) in Section B.4.2. It is the rationi of a standard normal variable to the square root of a chi-squared variable divided by its degrees of freedom.

TABLE 4.3 Regression Results for a Demand Equation

Sum of squared residuals:	0.120871
Standard error of the regression:	0.050712
R^2 based on 52 observations	0.958443

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	−21.21109	0.75322	−28.160
ln P_G	−0.021206	0.04377	−0.0485
ln *Income*/*Pop*	1.095874	0.07771	14.102
ln P_{nc}	−0.373612	0.15707	−2.379
ln P_{uc}	0.02003	0.10330	0.194

Example 4.4 Confidence Interval for the Income Elasticity of Demand for Gasoline

Using the gasoline market data discussed in Example 2.3, we estimated the following demand equation using the 52 observations:

$$\ln(G/Pop) = \beta_1 + \beta_2 \ln P_G + \beta_3 \ln(Income/Pop) + \beta_4 \ln P_{nc} + \beta_5 \ln P_{uc} + \varepsilon.$$

Least squares estimates of the model parameters with standard errors and t ratios are given in Table 4.3.

To form a confidence interval for the income elasticity, we need the critical value from the t distribution with $n - K = 52 - 5 = 47$ degrees of freedom. The 95 percent critical value is 2.012. Therefore a 95 percent confidence interval for β_3 is 1.095874 ± 2.012 (0.07771) = [0.9395, 1.2522].

We are interested in whether the demand for gasoline is income inelastic. The hypothesis to be tested is that β_3 is less than 1. For a one-sided test, we adjust the critical region and use the t_α critical point from the distribution. Values of the sample estimate that are greatly inconsistent with the hypothesis cast doubt on it. Consider testing the hypothesis

$$H_0 : \beta_3 < 1 \quad \text{versus} \quad H_1 : \beta_3 \geq 1.$$

The appropriate test statistic is

$$t = \frac{1.095874 - 1}{0.07771} = 1.234.$$

The critical value for a one-tailed test using the t distribution with 47 degrees of freedom is 1.678, which is greater than 1.234. We conclude that the data are consistent with the hypothesis that the income elasticity is less than one, so we do not reject the null hypothesis.

4.7.3 CONFIDENCE INTERVAL FOR A LINEAR COMBINATION OF COEFFICIENTS: THE OAXACA DECOMPOSITION

With normally distributed disturbances, the least squares coefficient estimator, $\mathbf{b}$, is normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. In Example 4.4, we showed how to use this result to form a confidence interval for one of the elements of $\boldsymbol{\beta}$. By extending those results, we can show how to form a confidence interval for a linear function of the parameters. **Oaxaca** (1973) and **Blinder's** (1973) **decomposition** provides a frequently used application.[4]

Let $\mathbf{w}$ denote a $K \times 1$ vector of known constants. Then, the linear combination $c = \mathbf{w}'\mathbf{b}$ is normally distributed with mean $\gamma = \mathbf{w}'\boldsymbol{\beta}$ and variance $\sigma_c^2 = \mathbf{w}'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{w}$,

[4] See Bourguignon et al. (2002) for an extensive application.

4.5.2 Confidence Intervals Based on Large Samples

If the disturbances are not normally distributed, then the development in the previous section, which departs from this assumption, is not useable. But, the large sample results in Section 4.4 provide an alternative approach. Based on the development that we used to obtain Theorem 4.4 and (4-35), we have that the limiting distribution of the statistic

$$z_n = \frac{\sqrt{n}(b_k - \beta_k)}{\sqrt{\frac{\sigma^2}{n} Q^{kk}}}$$

is standard normal, where $\mathbf{Q} = [\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$ and Q^{kk} is the kth diagonal element of $\mathbf{Q}$. Based on the Slutsky theorem (D.16), we may replace σ^2 with a consistent estimator, s^2, and obtain a statistic with the same limiting distribution. And, of course, we estimate $\mathbf{Q}$ with $(\mathbf{X}'\mathbf{X}/n)^{-1}$. This gives us precisely (4-41), which states that under the assumptions in Section 4.4, the "t" statistic in (4-41) converges to standard normal even if the disturbances are not normally distributed. The implication would be that to employ the asymptotic distribution of $\mathbf{b}$, we should use (4-42) to compute the confidence interval, but use the critical values from the standard normal table (e.g., 1.96) rather than from the t distribution. In practical terms, if the degrees of freedom in (4-42) are moderately large, say greater than 100, then the t distribution will be indistinguishable from the standard normal, and this large sample result would apply in any event. For smaller sample sizes, however, in the interest of conservatism, one might be advised to use the critical values from the t table rather the standard normal, even in the absence of the normality assumption. In the application in Example 4.8, based on a sample of 52 observations, we formed a confidence interval for the income elasticity of demand using the critical value of 2.012 from the t table with 47 degrees of freedom. If we chose to base the interval on the asymptotic normal distribution, rather than the standard normal, we would use the 95% critical value of 1.96. One might think this is a bit optimistic, however, and retain the value 2.012, again, in the interest of conservatism.

Example 4.9 Confidence Interval Based on the Asymptotic Distribution

In Example 4.4, we analyzed a dynamic form of the demand equation for gasoline,

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(Income/Pop) + \ldots + \gamma \ln(G/POP)_{t-1} + \varepsilon_t.$$

In this model, the long run price and income elasticities are $\theta_P = \beta_2/(1-\gamma)$ and $\theta_I = \beta_3/(1-\gamma)$. We computed estimates of these two nonlinear functions using the least squares and the delta method, Theorem 4.5. The point estimates were -0.411358 and 0.970522, respectively. The estimated asymptotic standard errors were 0.152296 and 0.162386. In order to form confidence intervals for θ_P and θ_I, we would generally use the asymptotic distribution, not the finite sample distribution. Thus, the two confidence intervals are

$$\hat{\theta}_P = -0.411358 \pm 1.96\ (0.152296) = [-0.709858, -0.112858]$$

and

$$\hat{\theta}_I = 0.970523 \pm 1.96\ (0.162386) = [0.652246, 1.288800].$$

In a sample of 51 observations, one might argue that using the critical value for the limiting normal distribution might be a bit optimistic. If so, using the critical value for the *t* distribution with 51-6 = 45 degrees of freedom would give a slightly wider interval. For example, for the the income elasticity the interval would be 0.970523 ± 2.014 (0.162386) = [0.643460, 1.297585]. We do note, this is a practical adjustment. The statistic based on the asymptotic standard error does not actually have a *t* distribution with 45 degrees of freedom.

TABLE 4.3 Regression Results for a Demand Equation

Sum of squared residuals:	0.120871		
Standard error of the regression:	0.050712		
R^2 based on 52 observations	0.958443		
Variable	***Coefficient***	***Standard Error***	***t Ratio***
Constant	−21.21109	0.75322	−28.160
ln P_G	−0.021206	0.04377	−0.0485
ln *Income*/*Pop*	1.095874	0.07771	14.102
ln P_{nc}	−0.373612	0.15707	−2.379
ln P_{uc}	0.02003	0.10330	0.194

Example 4.4 Confidence Interval for the Income Elasticity of Demand for Gasoline

Using the gasoline market data discussed in Example 2.3, we estimated the following demand equation using the 52 observations:

$$\ln(G/Pop) = \beta_1 + \beta_2 \ln P_G + \beta_3 \ln(Income/Pop) + \beta_4 \ln P_{nc} + \beta_5 \ln P_{uc} + \varepsilon.$$

Least squares estimates of the model parameters with standard errors and t ratios are given in Table 4.3.

To form a confidence interval for the income elasticity, we need the critical value from the t distribution with $n - K = 52 - 5 = 47$ degrees of freedom. The 95 percent critical value is 2.012. Therefore a 95 percent confidence interval for β_3 is 1.095874 ± 2.012 (0.07771) = [0.9395,1.2522].

We are interested in whether the demand for gasoline is income inelastic. The hypothesis to be tested is that β_3 is less than 1. For a one-sided test, we adjust the critical region and use the t_α critical point from the distribution. Values of the sample estimate that are greatly inconsistent with the hypothesis cast doubt on it. Consider testing the hypothesis

$$H_0 : \beta_3 < 1 \quad \text{versus} \quad H_1 : \beta_3 \geq 1.$$

The appropriate test statistic is

$$t = \frac{1.095874 - 1}{0.07771} = 1.234.$$

The critical value for a one-tailed test using the t distribution with 47 degrees of freedom is 1.678, which is greater than 1.234. We conclude that the data are consistent with the hypothesis that the income elasticity is less than one, so we do not reject the null hypothesis.

4.5 CONFIDENCE INTERVAL FOR A LINEAR COMBINATION OF COEFFICIENTS: THE OAXACA DECOMPOSITION

With normally distributed disturbances, the least squares coefficient estimator, $\mathbf{b}$, is normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. In Example 4.4, we showed how to use this result to form a confidence interval for one of the elements of $\boldsymbol{\beta}$. By extending those results, we can show how to form a confidence interval for a linear function of the parameters. **Oaxaca** (1973) and **Blinder's** (1973) **decomposition** provides a frequently used application.[11]

Let $\mathbf{w}$ denote a $K \times 1$ vector of known constants. Then, the linear combination $c = \mathbf{w}'\mathbf{b}$ is normally distributed with mean $\gamma = \mathbf{w}'\boldsymbol{\beta}$ and variance $\sigma_c^2 = \mathbf{w}'[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{w}$,

[11] See Bourgignon et al. (2002) for an extensive application.

which we estimate with $s_c^2 = \mathbf{w}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{w}$. With these in hand, we can use the earlier results to form a confidence interval for γ:

$$\text{Prob}[c + t_{\alpha/2}s_c \leq \gamma \leq c + t_{\alpha/2}s_c] = 1 - \alpha. \qquad (4-44)$$

This general result can be used, for example, for the sum of the coefficients or for a difference.

Consider, then, Oaxaca's application. In a study of labor supply, separate wage regressions are fit for samples of n_m men and n_f women. The underlying regression models are

$$\ln \text{wage}_{m,i} = \mathbf{x}'_{m,i}\boldsymbol{\beta}_m + \varepsilon_{m,i}, \quad i = 1, \ldots, n_m$$

and

$$\ln \text{wage}_{f,j} = \mathbf{x}'_{f,j}\boldsymbol{\beta}_f + \varepsilon_{f,j}, \quad j = 1, \ldots, n_f.$$

The regressor vectors include sociodemographic variables, such as age, and human capital variables, such as education and experience. We are interested in comparing these two regressions, particularly to see if they suggest wage discrimination. Oaxaca suggested a comparison of the regression functions. For any two vectors of characteristics,

$$\begin{aligned} E[\ln \text{wage}_{m,i}] - E[\ln \text{wage}_{f,j}] &= \mathbf{x}'_{m,i}\boldsymbol{\beta}_m - \mathbf{x}'_{f,j}\boldsymbol{\beta}_f \\ &= \mathbf{x}'_{m,i}\boldsymbol{\beta}_m - \mathbf{x}'_{m,i}\boldsymbol{\beta}_f + \mathbf{x}'_{m,i}\boldsymbol{\beta}_f - \mathbf{x}'_{f,j}\boldsymbol{\beta}_f \\ &= \mathbf{x}'_{m,i}(\boldsymbol{\beta}_m - \boldsymbol{\beta}_f) + (\mathbf{x}_{m,i} - \mathbf{x}_{f,j})'\boldsymbol{\beta}_f. \end{aligned}$$

The second term in this decomposition is identified with differences in human capital that would explain wage differences naturally, assuming that labor markets respond to these differences in ways that we would expect. The first term shows the differential in log wages that is attributable to differences unexplainable by human capital; holding these factors constant at $\mathbf{x}_m$ makes the first term attributable to other factors. Oaxaca suggested that this decomposition be computed at the means of the two regressor vectors, $\bar{\mathbf{x}}_m$ and $\bar{\mathbf{x}}_f$, and the least squares coefficient vectors, $\mathbf{b}_m$ and $\mathbf{b}_f$. If the regressions contain constant terms, then this process will be equivalent to analyzing $\overline{\ln y_m} - \overline{\ln y_f}$.

We are interested in forming a confidence interval for the first term, which will require two applications of our result. We will treat the two vectors of sample means as known vectors. Assuming that we have two independent sets of observations, our two estimators, $\mathbf{b}_m$ and $\mathbf{b}_f$, are independent with means $\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_f$ and covariance matrices $\sigma_m^2(\mathbf{X}_m'\mathbf{X}_m)^{-1}$ and $\sigma_f^2(\mathbf{X}_f'\mathbf{X}_f)^{-1}$. The covariance matrix of the difference is the sum of these two matrices. We are forming a confidence interval for $\bar{\mathbf{x}}_m'\mathbf{d}$ where $\mathbf{d} = \mathbf{b}_m - \mathbf{b}_f$. The estimated covariance matrix is

$$\text{Est. Var}[\mathbf{d}] = s_m^2(\mathbf{X}_m'\mathbf{X}_m)^{-1} + s_f^2(\mathbf{X}_f'\mathbf{X}_f)^{-1}. \qquad (4-45)$$

Now, we can apply the result above. We can also form a confidence interval for the second term; just define $\mathbf{w} = \bar{\mathbf{x}}_m - \bar{\mathbf{x}}_f$ and apply the earlier result to $\mathbf{w}'\mathbf{b}_f$.

4.7.4 TESTING THE SIGNIFICANCE OF THE REGRESSION

A question that is usually of interest is whether the regression equation as a whole is significant. This test is a joint test of the hypotheses that *all* the coefficients except the

AU: "Prediction," "forecasting," and "ex post prediction" are not in chap list

4.6 PREDICTION AND FORECASTING

After the estimation of the model parameters, a common use of regression modeling is for prediction of the dependent variable. We make a distinction between "prediction" and "forecasting," most easily based on the difference between cross section and time series modeling. **Prediction** (which would apply to either case) involves using the regression model to compute fitted (predicted) values of the dependent variable, either within the sample or for observations outside the sample. The same set of results will apply to cross sections, panels and time series. We consider these methods first. **Forecasting,** while largely the same exercise, explicitly gives a role to "time" and often involves lagged dependent variables and disturbances that are correlated with their past values. This exercise usually involves predicting future outcomes. An important difference between predicting and forecasting (as defined here) is that for predicting, we are usually examining a "scenario" of our own design. Thus, in the example below in which we are predicting the prices of Monet paintings, we might be interested in predicting the price of a hypothetical painting of a certain size and aspect ratio, or one that actually exists in the sample. In the time series context, we will often try to forecast an event such as real investment next year, not based on a hypothetical economy, but based on our best estimate of what economic conditions will be next year. We will use the term **ex post prediction** (or **ex post forecast**) for the cases in which the data used in the regression equation to make the prediction are either observed or constructed experimentally by the analyst. This would be the first case considered here. An **ex ante forecast** (in the time series context) will be one that requires the analyst to forecast the independent variables first before it is possible to forecast the dependent variable. In an exercise for this chapter, real investment is forecasted using a regression model that contains real GDP and the consumer price index. In order to forecast real investment, we must first forecast real GDP and the price index. Ex ante forecasting is considered briefly here and again in ~~Section 20.X~~ Chapter 20.

4.6.1 Prediction Intervals

Suppose that we wish to predict the value of y^0 associated with a regressor vector $\mathbf{x}^0$. The actual value would be

$$y^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0.$$

It follows from the Gauss-Markov theorem that (4-45)

$$\hat{y}^0 = \mathbf{x}^{0\prime}\mathbf{b} \quad (\text{4-23})$$

is the minimum variance linear unbiased estimator of $E[y^0|\mathbf{x}^0] = \mathbf{x}^{0\prime}\boldsymbol{\beta}$. The **prediction error** is

$$e^0 = \hat{y}^0 - y^0 = (\mathbf{b} - \boldsymbol{\beta})'\mathbf{x}^0 + \varepsilon^0.$$

The **prediction variance** of this estimator is

$$\text{Var}[e^0|\mathbf{X},\mathbf{x}^0] = \sigma^2 + \text{Var}[(\mathbf{b}-\boldsymbol{\beta})'\mathbf{x}^0 \,|\, \mathbf{X},\mathbf{x}^0] = \sigma^2 + \mathbf{x}^{0\prime}\left[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right]\mathbf{x}^0. \qquad (4\text{-}46)$$

If the regression contains a constant term, then an equivalent expression is

$$\text{Var}[e^0|\mathbf{X},\mathbf{x}^0] = \sigma^2\left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}\left(x_j^0 - \bar{x}_j\right)\left(x_k^0 - \bar{x}_k\right)\left(\mathbf{Z}'\mathbf{M}^0\mathbf{Z}\right)^{jk}\right], \qquad (4\text{-}47)$$

where $\mathbf{Z}$ is the K-1 columns of $\mathbf{X}$ not including the constant, $\mathbf{Z}'\mathbf{M}^0\mathbf{Z}$ is the matrix of sums of squares and products for the columns of $\mathbf{X}$ in deviations from their means [see (3-21)] and the "jk" superscript indicates the jk element of the inverse of the matrix. This result suggests that the width of a confidence interval (i.e., a **prediction interval**) depends on the distance of the elements of $\mathbf{x}^0$ from the center of the data. Intuitively, this idea makes sense; the farther the forecasted point is from the center of our experience, the greater is the degree of uncertainty. Figure 4.5 shows the effect for the bivariate case. Note that the prediction variance is composed of three parts. The second and third become progressively smaller as we accumulate more data (i.e., as n increases). But, the first term, σ^2 is constant, which implies that no matter how much data we have, we can never predict perfectly.

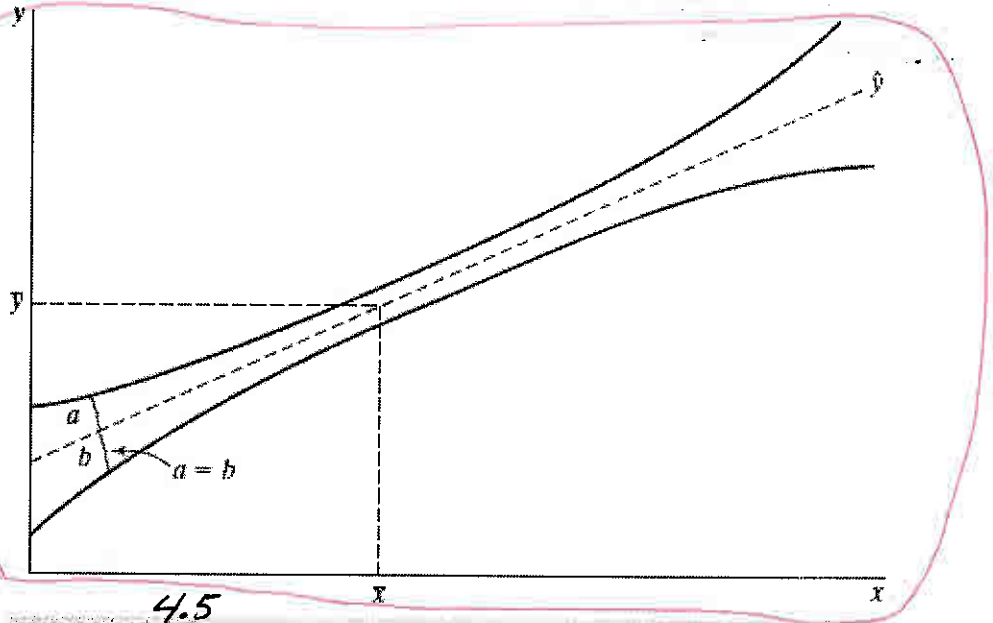


FIGURE 4.5 Prediction Intervals

The prediction variance can be estimated by using s^2 in place of σ^2. A confidence (prediction) interval for y^0 would then be formed using

$$\text{prediction interval} = \hat{y}^0 \pm t_{(1-\alpha/2),[n-K]}\,\text{se}\left(e^0\right) \qquad \textbf{(4-48)}$$

where $t_{(1-\alpha/2),[n-K]}$ is the appropriate critical value for $100(1-\alpha)\%$ significance from the t table for n-K degrees of freedom and $\text{se}(e^0)$ is the square root of the prediction variance.

4.6.2 Predicting *y* When the Regression Model Describes Log *y*

It is common to use the regression model to describe a function of the dependent variable, rather than the variable, itself. In Example 4.5 we model the sale prices of Monet paintings using

$$\ln Price = \beta_1 + \beta_2 \ln Area + \beta_3 AspectRatio + \varepsilon$$

(area is width times height of the painting and aspect ratio is the height divided by the width). The log form is convenient in that the coefficient provides the elasticity of the dependent variable with respect to the independent variable. I.e., in this model, $\beta_2 = \partial E[\ln Price | \ln Area, AspectRatio]/\partial \ln Area$. However, the equation in this form is less interesting for prediction purposes than one that predicts the price, itself. The natural approach for a predictor of the form

$$\ln y^0 = \mathbf{x}^{0\prime}\mathbf{b}$$

would be to use

$$\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b}).$$

The problem is that $E[y|\mathbf{x}^0]$ is not equal to $\exp(E[\ln y|\mathbf{x}^0])$. The appropriate conditional mean function would be

$$\begin{aligned} E[y|\mathbf{x}^0] &= E[\exp(\mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0)|\mathbf{x}^0] \\ &= \exp(\mathbf{x}^{0\prime}\boldsymbol{\beta})\, E[\exp(\varepsilon^0)|\mathbf{x}^0]. \end{aligned}$$

The second term is not $\exp(E[\varepsilon^0|x^0]) = 1$ in general. The precise result if $\varepsilon^0|\mathbf{x}^0$ is normally distributed with mean zero and variance σ^2 is $E[\exp(\varepsilon^0)|\mathbf{x}^0] = \exp(\sigma^2/2)$. (See Section B.4.4.) The implication for normally distributed disturbances would be that an appropriate predictor for the conditional mean would be

$$\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b} + s^2/2) > \exp(\mathbf{x}^{0\prime}\mathbf{b}), \qquad \textbf{(4-49)}$$

which would seem to imply that the naïve predictor would systematically underpredict *y*. However, this is not necessarily the appropriate interpretation of this result. The inequality implies that the naïve predictor will systematically underestimate the conditional mean function, not necessarily the realizations of the variable, itself. The pertinent question is whether the conditional mean function is the desired predictor for the exponent of the dependent variable in the log regression. The conditional median might be more interesting, particularly for a financial variable such as income, expenditure or the price of a painting. If the distribution of the variable in the log regression is symmetrically distributed (as they are when the disturbances are normally distributed), then the exponent will be asymmetrically distributed with a long tail in the positive direction, and the mean will exceed the median, possibly vastly so. In such cases, the median is often a preferred estimator of the center of a distribution. For estimating the median, rather then the mean, we would revert to the original naïve predictor, $\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b})$.

Given the preceding, we consider estimating $E[\exp(y)|\mathbf{x}^0]$. If we wish to avoid the normality assumption, then it remains to determine what one should use for $E[\exp(\varepsilon^0)|\mathbf{x}^0]$. Duan (1983) suggested the consistent estimator (assuming that the expectation is a constant, i.e., that the regression is homoscedastic),

$$\hat{E}[\exp(\varepsilon^0)\,|\,\mathbf{x}^0] = h^0 = \frac{1}{n}\sum_{i=1}^{n}\exp(e_i), \qquad \textbf{(4-50)}$$

where e_i is a least squares residual in the original log form regression. Then, Duan's **smearing estimator** for prediction of y^0 is

$$\hat{y}^0 = h^0 \exp(\mathbf{x}^{0\prime}\mathbf{b}).$$

4.6.3 Prediction Interval for *y* When the Regression Model Describes log *y*

We obtained a prediction interval in (4-48) for $\ln y|\mathbf{x}^0$ in the loglinear model $\ln y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$,

$$\left[\ln \hat{y}^0_{LOWER}, \ln \hat{y}^0_{UPPER}\right] = \left[\mathbf{x}^{0\prime}\mathbf{b} - t_{(1-\alpha/2),[n-K]}\,\mathrm{se}\left(e^0\right), \mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha/2),[n-K]}\,\mathrm{se}\left(e^0\right)\right].$$

For a given choice of α, say, 0.05, these values give the .025 and .975 quantiles of the distribution of $\ln y|\mathbf{x}^0$. If we wish specifically to estimate these quantiles of the distribution of $y|\mathbf{x}^0$, not $\ln y|\mathbf{x}^0$, then we would use;

$$\left[\hat{y}^0_{LOWER}, \hat{y}^0_{UPPER}\right] = \left\{\exp\left[\mathbf{x}^{0\prime}\mathbf{b} - t_{(1-\alpha/2),[n-K]}\,\mathrm{se}\left(e^0\right)\right], \exp\left[\mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha/2),[n-K]}\,\mathrm{se}\left(e^0\right)\right]\right\}. \qquad \textbf{(4-51)}$$

This follows from the result that if $\mathrm{Prob}[\ln y \le \ln L] = 1 - \alpha/2$, then $\mathrm{Prob}[y \le L] = 1 - \alpha/2$. The result is that the natural estimator is the right one for estimating the specific quantiles of the distribution of the original variable. However, if the objective is to find an interval estimator for $y|\mathbf{x}^0$ that is as narrow as possible, then this approach is not optimal. If the distribution of y is asymmetric, as it would be for a loglinear model with normally distributed disturbances, then the naïve interval estimator is longer than necessary. Figure 4.6 shows why. We suppose that (L,U) in the figure is the prediction interval formed by (4-51). Then, the probabilities to the left of L and to the right of U each equal $\alpha/2$. Consider alternatives $L_0 = 0$ and U_0 instead. As we have constructed the figure, area (probability) between L_0 and L equals the area between U_0 and U. But, because the density is so much higher at L, the distance $(0,U_0)$, the dashed interval, is visibly shorter than that between (L,U). The sum of the two tail probabilities is still equal to α, so this provides a shorter prediction interval. We could improve on (4-51) by using, instead, $(0,U_0)$ where U_0 is simply $\exp[\mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha),[n-K]}\,\mathrm{se}(e^0)]$ (i.e., we put the entire tail area to the right of the upper value). However, while this is an improvement, it goes too far, as we now demonstrate.

Consider finding directly the shortest prediction interval. We treat this as an optimization problem:

$$\text{Minimize } (L,U)\text{: } I = U - L \text{ subject to } F(L) + [1 - F(U)] = \alpha,$$

where F is the CDF of the random variable y (not $\ln y$). That is, we seek the shortest interval for which the two tail probabilities sum to our desired α (usually 0.05). Formulate this as a Lagrangean problem,

Minimize (L,U,λ): $I^* = U - L + \lambda\,[F(L) + (1-F(U)) - \alpha]$.

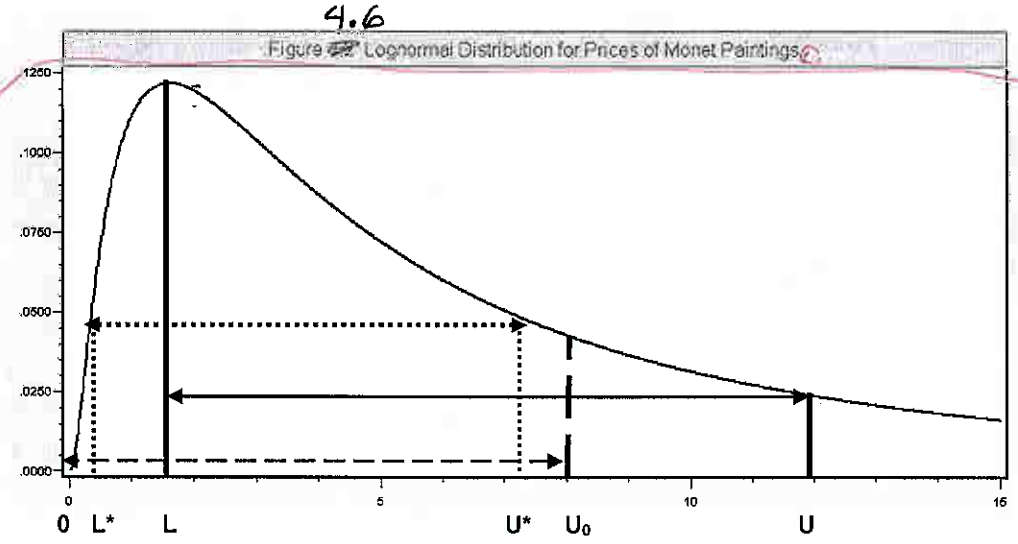
The solutions are found by equating the three partial derivatives to zero:

$$\partial I^*/\partial L = -1 + \lambda f(L) = 0,$$
$$\partial I^*/\partial U = 1 - \lambda f(U) = 0,$$
$$\partial I^*/\partial \lambda = F(L) + [1 - F(U)] - \alpha = 0,$$

where $f(L) = F'(L)$ and $f(U) = F'(U)$ are the derivtives of the cdf, which are the densities of the random variable at L and U, respectively. The third equation enforces the restriction that the two tail areas sum to α, but does not force them to be equal. By adding the first two equations, we find that $\lambda[f(L)-f(U)] = 0$, which, if λ is not zero, means that the solution is obtained by locating (L^*,U^*) such that the tail areas sum to α and the densities are equal. Looking again at Figure 4.6, we can see that the solution we would seek is (L^*,U^*) where $0 < L^* < L$ and $U^* < U_0$. This is the shortest interval, and it is shorter than both $[0,U_0]$ and $[L,U]$

This derivation would apply for any distribution, symmetric or otherwise. For a symmetric distribution, however, we would obviously return to the symmetric interval in (4-51). It provides the correct solution for when the distribution is asymmetric. In Bayesian analysis, the counterpart when we examine the distribution of a parameter conditioned on the data, is the **highest posterior density interval**. (See Section 16.4.2.) For practical application, this computation requires a specific assumption for the distribution of $y|\mathbf{x}^0$, such as lognormal. Typically, we would use the smearing estimator specifically to avoid the distributional assumption. There also is no simple formula to use to locate this interval, even for the lognormal distribution. A crude grid search would probably be best, though the computations are each very simple. What this derivation does establish is that one can do substantially better than the naïve interval estimator, for example using $[0,U_0]$.

4.6



Example 4.10 Pricing Art

In Example 4.5, we suggested an intriguing feature of the market for Monet paintings, that larger paintings sold at auction for more than than smaller ones. In this example, we will examine that proposition empirically. Table F4.1 contains data on 430 auction prices for Monet paintings, with data on the dimensions of the paintings and several other variables that we will examine in later examples. Figure 4.7 shows a histogram for the sample of sale prices (in $Million). Figure 4.8 shows a histogram for the logs of the prices.

FIG 4.7

FIG 4.8

4.7

$million

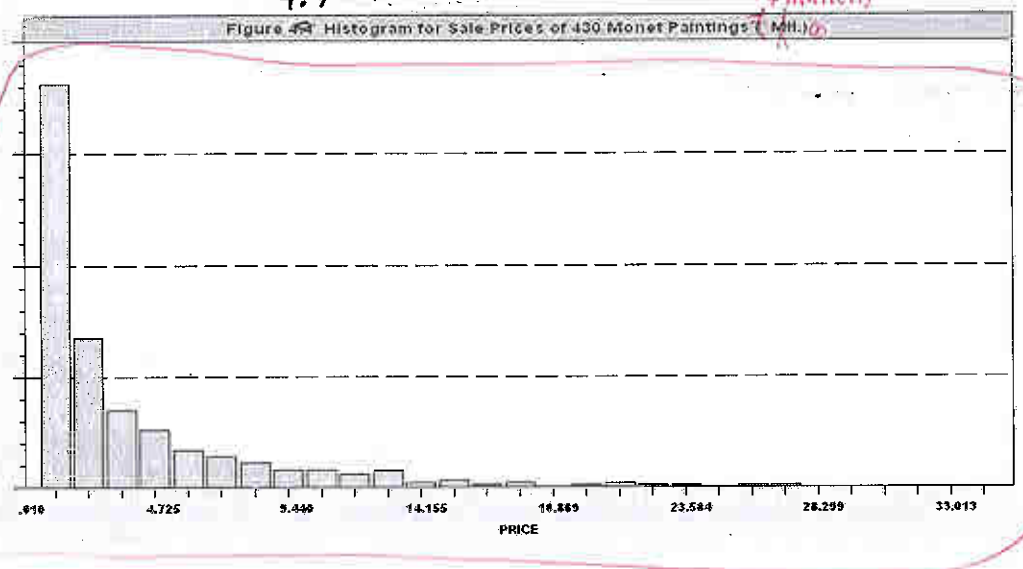
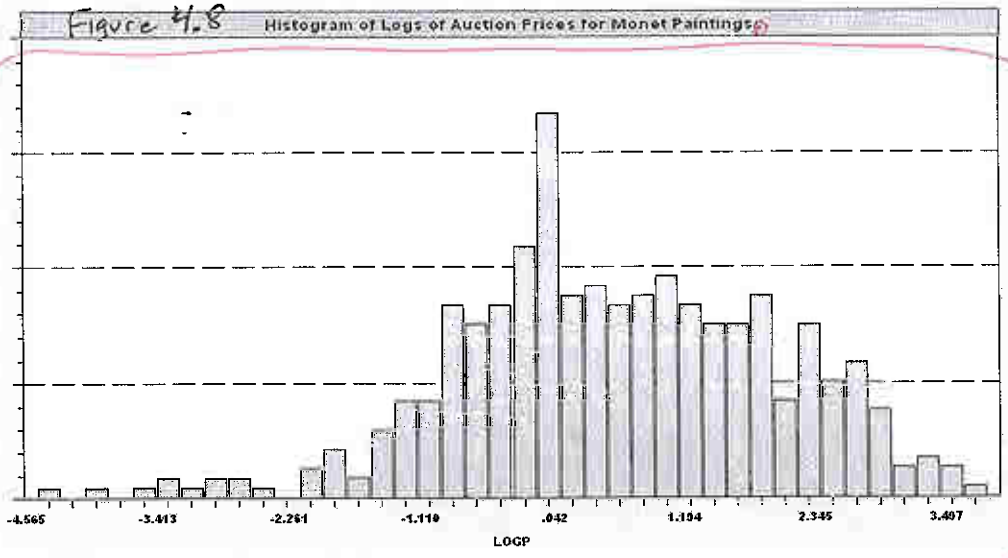


Figure 4.8



Results of the linear regression of lnPrice on lnArea (height times width) and Aspect Ratio (height divided by width) are given in Table 4.6.

TABLE 4.6 Estimated Equation for Log Price

Mean of log Price	.33274
Sum of squared residuals	519.17235
Standard error of regression	1.10266
R-squared	.33620
Adjusted R-squared	.33309
Number of observations	430

Variable	Coefficient	Standard Error	t	Mean of X
Constant	-8.42653	.61183	-13.77	1.00000
LOGAREA	1.33372	.09072	14.70	6.68007
ASPECT	-.16537	.12753	-1.30	0.90759

Estimated Asymptotic Covariance Matrix

	Constant	LogArea	AspectRatio
Constant	.37434	-.05429	-.00974
LogArea	-.05429	.00823	-.00075
AspectRatio	-.00974	-.00075	.01626

We consider using the regression model to predict the price of one of the paintings, a 1903 painting of Charing Cross Bridge that sold for $3,522,500. The painting is 25.6" high and 31.9" wide. (This is observation 60 in the sample.) The log area equals ln(25.6×31.9) = 6.705198 and the aspect ratio equals 25.6/31.9 = 0.802508. The prediction for the log of the price would be

$$\ln P \mid \mathbf{x}^0 = -8.42653 + 1.33372(6.705198) - 0.16537(0.802508) = 0.383636.$$

Note that the mean log price is 0.33274, so this painting is expected to be sell for roughly 5% more than the average painting, based on its dimensions. The estimate of the prediction variance is computed using (4-47); $s_p = 1.104027$. The sample is large enough to use the critical value from the standard normal table, 1.96, for a 95% confidence interval. A prediction interval for the log of the price is therefore

$$0.383636 \pm 1.96(1.104027) = [-1.780258, 2.547529].$$

For predicting the price, the naïve predictor would be exp(0.383636) = $1.476411M, which is far under the actual sale price of $3.5225M. To compute the smearing estimator, we require the mean of the exponents of the residuals, which is 1.813045. The revised point estimate for the price would thus be 1.813045×1.47641 = $2.660844M – this is better, but still fairly far off. This particular painting seems to have sold for relatively more than history (the data) would have predicted.

To compute an interval estimate for the price, we begin with the naïve prediction by simply exponentiating the lower and upper values for the log price, which gives a prediction interval for 95% confidence of [$0.168595M, $12.77503M]. Using the method suggested in Section 4.6.3, however, we are able to narrow this interval to [0.021261,9.027543], a range of $9M compared to the range based on the simple calculation of $12.2M. The interval divides the .05 tail probability into 0.00063 on the left and .04937 on the right. The search algorithm is outlined below.

Grid Search Algorithm for Optimal Prediction Interval, [LO,UO]:

$\mathbf{x}^0 = (1, \log(25.6 \times 31.9), 25.6/31.9)'$;

$\hat{\mu}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b})$, $\hat{\sigma}_p^0 = \sqrt{s^2 + \mathbf{x}^{0\prime}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x}^0}$;

Confidence interval for $\log P|\mathbf{x}^0$: [Lower,Upper] = $[\hat{\mu}^0 - 1.96\hat{\sigma}_p^0,\ \hat{\mu}^0 + 1.96\hat{\sigma}_p^0]$;

Naïve confidence interval for $Price|\mathbf{x}^0$: L1=exp(Lower) ; U1=exp(Upper);

Initial value of L was .168595, LO = this value;

Grid search for optimal interval, decrement by Δ = .005 (chosen ad hoc);

Decrement LO and compute companion UO until densities match;

(*) LO = LO - Δ = new value of LO;

$$f(\text{LO}) = \left[LO\hat{\sigma}_p^0\sqrt{2\pi}\right]^{-1}\exp\left[-\tfrac{1}{2}\left((\ln LO - \hat{\mu}^0)/\hat{\sigma}_p^0\right)^2\right];$$

$F(\text{LO}) = \Phi((\ln(\text{LO}) - \hat{\mu}^0)/\hat{\sigma}_p^0)$ = left tail probability;

$\text{UO} = \exp(\hat{\sigma}_p^0\,\Phi^{-1}[F(\text{LO}) + .95] + \hat{\mu}^0)$ = next value of UO;

$$f(\text{UO}) = \left[UO\hat{\sigma}_p^0\sqrt{2\pi}\right]^{-1}\exp\left[-\tfrac{1}{2}\left((\ln UO - \hat{\mu}^0)/\hat{\sigma}_p^0\right)^2\right];$$

$1-F(\text{UO}) = 1 - \Phi((\ln(\text{UO}) - \hat{\mu}^0)/\hat{\sigma}_p^0)$ = right tail probability;

Compare f(LO) to f(UO). If not equal, return to (*). If equal, exit.

4.6.4 Forecasting

The preceding assumes that $\mathbf{x}^0$ is known with certainty, ex post, or has been forecasted perfectly, ex ante. If $\mathbf{x}^0$ must, itself, be forecasted (an ex ante forecast), then the formula for the forecast variance in (4-XX) would have to be modified to incorporate the uncertainty in forecasting $\mathbf{x}^0$. This would be analogous to the term σ^2 in the prediction variance that accounts for the implicit prediction of ε^0. This will vastly complicate the computation. Most authors view it as simply intractable. Beginning with Feldstein (1971), derivation of firm analytical results for the correct forecast variance for this case remain to be derived except for simple special cases. The one qualitative result that seems certain is that (4-XX) will understate the true variance. McCullough (1996) presents an alternative approach to computing appropriate forecast standard errors based on the method of bootstrapping. (See Chapter XX.)

Various measures have been proposed for assessing the predictive accuracy of forecasting models.[10] Most of these measures are designed to evaluate **ex post forecasts,** that is, forecasts for which the independent variables do not themselves have to be forecasted. Two measures that are based on the residuals from the forecasts are the **root mean squared error,**

$$\text{RMSE} = \sqrt{\frac{1}{n^0}\sum_i (y_i - \hat{y}_i)^2}$$

and the **mean absolute error,**

$$\text{MAE} = \frac{1}{n^0}\sum_i |y_i - \hat{y}_i|,$$

where n^0 is the number of periods being forecasted. (Note that both of these, as well as the measures below, are backward looking in that they are computed using the observed data on the independent variable.) These statistics have an obvious scaling problem — multiplying values of the dependent variable by any scalar multiplies the measure by that scalar as well. Several measures that are scale free are based on the **Theil *U* statistic:**[13]

$$\text{U} = \sqrt{\frac{(1/n^0)\sum_i (y_i - \hat{y}_i)^2}{(1/n^0)\sum_i y_i^2}}.$$

This measure is related to R^2 but is not bounded by zero and one. Large values indicate a poor forecasting performance. An alternative is to compute the measure in terms of the changes in y:

$$\text{U}_\Delta = \sqrt{\frac{(1/n^0)\sum_i (\Delta y_i - \Delta\hat{y}_i)^2}{(1/n^0)\sum_i (\Delta y_i)^2}}$$

where $\Delta y_i = y_i - y_{i-1}$ and $\Delta\hat{y}_i = \hat{y}_i - y_{i-1}$, or, in percentage changes, $\Delta y_i = (y_i - y_{i-1})/y_{i-1}$ and $\Delta\hat{y}_i = (\hat{y}_i - y_{i-1})/y_{i-1}$. These measures will reflect the model's ability to track turning points in the data.

[10] See Theil (1961) and Fair (1984).

[13] Theil (1961).

4.7 DATA PROBLEMS

The analysis to this point has assumed that the data in hand, $\mathbf{X}$ and $\mathbf{y}$, are well measured and correspond to the assumptions of the model in Table 2.1 and to the variables described by the underlying theory. At this point, we consider several ways that "real world," observed nonexperimental data fail to meet the assumptions. Failure of the assumptions generally has implications for the performance of the estimators of the model parameters—unfortunately, none of them good. The cases we will examne are:

- Multicollinearity: Although the full rank assumption, A2, is met, it almost fails. ("Almost" is a matter of degree, and sometimes a matter of interpretation.) Multicollinearity leads to imprecision in the estimator, though not to any systematic biases in estimation.
- Missing values: Gaps in $\mathbf{X}$ and/or $\mathbf{y}$ can be harmless. In many cases, the analyst can (and should) simply ignore them, and just use the complete data in the sample. In other cases, when the data are missing for reasons that are related to the outcome being studied, ignoring the problem can lead to inconsistency of the estimators.
- Measurement error: Data often correspond only imperfectly to the theoretical construct that appears in the model—individual data on income and education are familiar examples. Measurement error is never benign. The least harmful case is measurement error in the dependent variable. In this case, at least under probably reasonable assumptions, the implication is to degrade the fit of the model to the data compared to the (unfortunately hypothetical) case in which the data are accurately measured. Measurement error in the regressors is malignant—it produces systematic biases in estimation that are difficult to remedy.

4.7.1 MULTICOLLINEARITY

The Gauss–Markov theorem states that among all linear unbiased estimators, the least squares estimator has the smallest variance. Although this result is useful, it does not assure us that the least squares estimator has a small variance in any absolute sense. Consider, for example, a model that contains two explanatory variables and a constant. For either slope coefficient,

$$\text{Var}[b_k \mid \mathbf{X}] = \frac{\sigma^2}{(1 - r_{12}^2)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2} = \frac{\sigma^2}{(1 - r_{12}^2)S_{kk}}, \quad k = 1, 2. \tag{4-18}$$

If the two variables are perfectly correlated, then the variance is infinite. The case of an exact linear relationship among the regressors is a serious failure of the assumptions of the model, not of the data. The more common case is one in which the variables are highly, but not perfectly, correlated. In this instance, the regression model retains all its assumed properties, although potentially severe statistical problems arise. The problem faced by applied researchers when regressors are highly, although not perfectly, correlated include the following symptoms:

- Small changes in the data produce wide swings in the parameter estimates.
- Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high.
- Coefficients may have the "wrong" sign or implausible magnitudes.

For convenience, define the data matrix, $\mathbf{X}$, to contain a constant and $K - 1$ other variables measured in deviations from their means. Let $\mathbf{x}_k$ denote the kth variable, and let $\mathbf{X}_{(k)}$ denote all the other variables (including the constant term). Then, in the inverse matrix, $(\mathbf{X}'\mathbf{X})^{-1}$, the kth diagonal element is

$$\begin{aligned}(\mathbf{x}_k'\mathbf{M}_{(k)}\mathbf{x}_k)^{-1} &= [\mathbf{x}_k'\mathbf{x}_k - \mathbf{x}_k'\mathbf{X}_{(k)}(\mathbf{X}_{(k)}'\mathbf{X}_{(k)})^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k]^{-1} \\ &= \left[\mathbf{x}_k'\mathbf{x}_k\left(1 - \frac{\mathbf{x}_k'\mathbf{X}_{(k)}(\mathbf{X}_{(k)}'\mathbf{X}_{(k)})^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k}{\mathbf{x}_k'\mathbf{x}_k}\right)\right]^{-1} \\ &= \frac{1}{(1 - R_k^2)S_{kk}},\end{aligned} \tag{4-19}$$

where R_k^2 is the R^2 in the regression of x_k on all the other variables. In the multiple regression model, the variance of the kth least squares coefficient estimator is σ^2 times this ratio. It then follows that the more highly correlated a variable is with the other variables in the model (collectively), the greater its variance will be. In the most extreme case, in which $\mathbf{x}_k$ can be written as a linear combination of the other variables so that $R_k^2 = 1$, the variance becomes infinite. The result

$$\text{Var}[b_k \mid \mathbf{X}] = \frac{\sigma^2}{(1 - R_k^2)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}, \tag{4-20}$$

shows the three ingredients of the precision of the kth least squares coefficient estimator:

- Other things being equal, the greater the correlation of x_k with the other variables, the higher the variance will be, due to multicollinearity.

- Other things being equal, the greater the variation in x_k, the lower the variance will be. This result is shown in Figure 4.3.
- Other things being equal, the better the overall fit of the regression, the lower the variance will be. This result would follow from a lower value of σ^2. We have yet to develop this implication, but it can be suggested by Figure 4.3 by imagining the identical figure in the right panel but with all the points moved closer to the regression line.

Since nonexperimental data will never be orthogonal ($R_k^2 = 0$), to some extent multicollinearity will always be present. When is multicollinearity a problem? That is, when are the variances of our estimates so adversely affected by this intercorrelation that we should be "concerned"? Some computer packages report a **variance inflation factor** (VIF), $1/(1 - R_k^2)$, for each coefficient in a regression as a diagnostic statistic. As can be seen, the VIF for a variable shows the increase in $\text{Var}[b_k]$ that can be attributable to the fact that this variable is not orthogonal to the other variables in the model. Another measure that is specifically directed at $\mathbf{X}$ is the **condition number** of $\mathbf{X}'\mathbf{X}$, which is the square root of the ratio of the largest characteristic root of $\mathbf{X}'\mathbf{X}$ (after scaling each column so that it has unit length) to the smallest. Values in excess of 20 are suggested as indicative of a problem [Belsley, Kuh, and Welsch (1980)]. (The condition number for the Longley data of Example 4.6 is over 15,000!)

AU: "condition number" not in chap. list.

AU: Confirm Exm 4.11 is OK

4.11

Example 4.6 Multicollinearity in the Longley Data

The data in Appendix Table F4.2 were assembled by J. Longley (1967) for the purpose of assessing the accuracy of least squares computations by computer programs. (These data are still widely used for that purpose.) The Longley data are notorious for severe multicollinearity. Note, for example, the last year of the data set. The last observation does not appear to be unusual. But, the results in Table 4.5 show the dramatic effect of dropping this single observation from a regression of employment on a constant and the other variables. The last coefficient rises by 600 percent, and the third rises by 800 percent.

Several strategies have been proposed for finding and coping with multicollinearity.[3] Under the view that a multicollinearity "problem" arises because of a shortage of information, one suggestion is to obtain more data. One might argue that if analysts had such additional information available at the outset, they ought to have used it before reaching this juncture. More information need not mean more observations, however. The obvious practical remedy (and surely the most frequently used) is to drop variables suspected of causing the problem from the regression—that is, to impose on the regression an assumption, possibly erroneous, that the "problem" variable

TABLE 4.5 Longley Results: Dependent Variable is Employment

	1947–1961	*Variance Inflation*	*1947–1962*
Constant	1,459,415		1,169,087
Year	−721.756	143.4638	−576.464
GNP deflator	−181.123	75.6716	−19.7681
GNP	0.0910678	132.467	0.0643940
Armed Forces	−0.0749370	1.55319	−0.0101453

[3] See Hill and Adkins (2001) for a description of the standard set of tools for diagnosing collinearity.