does not appear in the model. In doing so, one encounters the problems of specification that we will discuss in Section 7.2. If the variable that is dropped actually belongs in the model (in the sense that its coefficient, β_k, is not zero), then estimates of the remaining coefficients will be biased, possibly severely so. On the other hand, overfitting—that is, trying to estimate a model that is too large—is a common error, and dropping variables from an excessively specified model might have some virtue. Several other practical approaches have also been suggested. An approach sometimes used [see, e.g., Gurmu, Rilstone, and Stern (1999)] is to use a small number, say L, of **principal components** constructed from the K original variables. [See Johnson and Wichern (2005).] The problem here is that if the original model in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ were correct, then it is unclear what one is estimating when one regresses $\mathbf{y}$ on some small set of linear combinations of the columns of $\mathbf{X}$. Algebraically, it is simple; at least for the principal components case, in which we regress $\mathbf{y}$ on $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$ to obtain $\mathbf{d}$, it follows that $E[\mathbf{d}] = \boldsymbol{\delta} = \mathbf{C}_L'\boldsymbol{\beta}$. In an economic context, if $\boldsymbol{\beta}$ has an interpretation, then it is unlikely that $\boldsymbol{\delta}$ will. (How do we interpret the price elasticity minus twice the income elasticity?)

Using diagnostic tools to detect multicollinearity could be viewed as an attempt to distinguish a bad model from bad data. But, in fact, the problem only stems from a prior opinion with which the data seem to be in conflict. A finding that suggests multicollinearity is adversely affecting the estimates seems to suggest that but for this effect, all the coefficients would be statistically significant and of the right sign. Of course, this situation need not be the case. If the data suggest that a variable is unimportant in a model, then, the theory notwithstanding, the researcher ultimately has to decide how strong the commitment is to that theory. Suggested "remedies" for multicollinearity might well amount to attempts to force the theory on the data.

4.8.2 MISSING OBSERVATIONS

It is common for data sets to have gaps, for a variety of reasons. Perhaps the most frequent occurrence of this problem is in survey data, in which respondents may simply fail to respond to the questions. In a time series, the data may be missing because they do not exist at the frequency we wish to observe them; for example, the model may specify monthly relationships, but some variables are observed only quarterly. In panel data sets, the gaps in the data may arise because of **attrition** from the study. This is particularly common in health and medical research, when individuals choose to leave the study—possibly because of the success or failure of the treatment that is being studied.

There are several possible cases to consider, depending on why the data are missing. The data may be simply unavailable, for reasons unknown to the analyst and unrelated to the completeness or the values of the other observations in the sample. This is the most benign situation. If this is the case, then the complete observations in the sample constitute a usable data set, and the only issue is what possibly helpful information could be salvaged from the incomplete observations. Griliches (1986) calls this the **ignorable case** in that, for purposes of estimation, if we are not concerned with efficiency, then we may simply delete the incomplete observations and ignore the problem. Rubin (1976, 1987) and Little and Rubin (1987) label this case **missing completely at random,** or MCAR.

4.7.2 PRETEST ESTIMATION

As a response to what appears to be a "multicollinearity problem," it is often difficult to resist the temptation to drop what appears to be an offending variable from the regression, if it seems to be the one causing the problem. This "strategy" creates a subtle dilemma for the analyst. Consider the partitioned multiple regression

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

If we regress $\mathbf{y}$ only on $\mathbf{X}_1$, the estimator is biased;

$$E[\mathbf{b}_1|\mathbf{X}] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2}\boldsymbol{\beta}_2.$$

The covariance matrix of this estimator is

$$\text{Var}[\mathbf{b}_1|\mathbf{X}] = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

(Keep in mind, this variance is around the $E[\mathbf{b}_1|\mathbf{X}]$, not around $\boldsymbol{\beta}_1$.) If $\boldsymbol{\beta}_2$ is not actually zero, then in the multiple regression of $\mathbf{y}$ on $(\mathbf{X}_1,\mathbf{X}_2)$, the variance of $\mathbf{b}_{1.2}$ around its mean, $\boldsymbol{\beta}_1$ would be

$$\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}] = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$$

where

$$\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2',$$

or

$$\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}] = \sigma^2[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1]^{-1}.$$

We compare the two covariance matrices. It is simpler to compare the inverses. [See result (A-120).] Thus,

$$\{\text{Var}[\mathbf{b}_1|\mathbf{X}]\}^{-1} - \{\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}]\}^{-1} = (1/\sigma^2)\,\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1,$$

which is a nonnegative definite matrix. The implication is that the variance of $\mathbf{b}_1$ is not larger than the variance of $\mathbf{b}_{1.2}$ (since its inverse is at least as large). It follows that although $\mathbf{b}_1$ is biased, its variance is never larger than the variance of the unbiased estimator. In any realistic case (i.e., if $\mathbf{X}_1'\mathbf{X}_2$ is not zero), in fact it will be smaller. We get a useful comparison from a simple regression with two variables measured as deviations from their means. Then, $\text{Var}[b_1|\mathbf{X}] = \sigma^2/S_{11}$ where $S_{11} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2$ and $\text{Var}[b_{1.2}|\mathbf{X}] = \sigma^2/[S_{11}(1-r_{12}^2)]$ where r_{12}^2 is the squared correlation between x_1 and x_2.

The result in the preceding paragraph poses a bit of a dilemma for applied researchers. The situation arises frequently in the search for a model specification. Faced with a variable that a researcher suspects should be in their model, but that is causing a problem of multicollinearity, the analyst faces a choice of omitting the relevant variable or including it and estimating its (and all the other variables') coefficient imprecisely. This presents a choice between two estimators, b_1 and $b_{1.2}$. In fact, what researchers usually do actually creates a third estimator. It is common to include the problem variable provisionally. If its t ratio is sufficiently large, it is retained; otherwise it is discarded. This third estimator is called a **pretest estimator.** What is known about pretest estimators is not encouraging. Certainly they are biased. How badly depends on the unknown parameters. Analytical results suggest that the pretest estimator is the least precise of the three when the researcher is most likely to use it. [See Judge et al. (1985).] The conclusion to be drawn is that as a general rule, the methodology leans away from estimation strategies that include ad hoc remedies for multicollinearity.

AU: "pretest estimator" is not in chap. list.

4.7.3 PRINCIPAL COMPONENTS

A device that has been suggested for "reducing" multicollinearity [see, e.g., Gurmu, Rilstone, and Stern (1999)] is to use a small number, say L, of **principal components** constructed as linear combinations of the K original variables. [See Johnson and Wichern (2005, Chapter 8).] (The mechanics are illustrated in Example 4.6.) The argument against using this approach is that if the original specification in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ were correct, then it is unclear what one is estimating when one regresses $\mathbf{y}$ on some small set of linear combinations of the columns of $\mathbf{X}$. For a set of $L < K$ principal components, if we regress $\mathbf{y}$ on $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$ to obtain $\mathbf{d}$, it follows that $E[\mathbf{d}] = \boldsymbol{\delta} = \mathbf{C}_L'\boldsymbol{\beta}$. (The proof is considered in the exercises.) In an economic context, if $\boldsymbol{\beta}$ has an interpretation, then it is unlikely that $\boldsymbol{\delta}$ will. (E.g., how do we interpret the price elasticity minus twice the income elasticity?)

This orthodox interpretation cautions the analyst about mechanical devices for coping with multicollinearity that produce uninterpretable mixtures of the coefficients. But, there are also situations in which the model is built on a platform that might well involve a mixture of some measured variables. For example, one might be interested in a regression model that contains "ability," ambiguously defined. As a measured counterpart, the analyst might have in hand standardized scores on a set of tests, none of which individually has any particular meaning in the context of the model. In this case, a mixture of the measured test scores might serve as one's preferred proxy for the underlying variable. The study in Example 4.8 describes another natural example.

EXAMPLE 4.8 Predicting Movie Success

Predicting the box office success of movies is a favorite exercise for econometricians. [See, e.g., Litman (1983), Ravid (1999), De Vany (2003), De Vany and Walls (1999, 2002,2003) and Simonoff and Sparrow (2000).] The traditional predicting equation takes the form

$$\textit{Box Office Receipts} = f(\textit{Budget, Genre, MPAA Rating, Star Power, Sequel, etc.}) + \varepsilon.$$

Coefficients of determination on the order of .4 are fairly common. Notwithstanding the relative power of such models, the common wisdom in Hollywood is "nobody knows." There is tremendous randomness in movie success, and few really believe they can forecast it with any reliability.[15] Versaci (2009) added a new element to the model, "internet buzz." Internet buzz is vaguely defined to be internet traffic and interest on familiar websites such as RottenTomatoes.com, ImDB.com, Fandango.com and traileraddict.com. None of these by themselves defines internet buzz. But, collectively, activity on these websites, say 3 weeks before a movie's opening might be a useful predictor of upcoming success. Versaci's data set (Table F4.3) contains data for 62 movies released in 2009, including four internet buzz variables, all measured three weeks prior to the release of the movie:

$buzz_1$ = number of internet views of movie trailer at traileraddict.com,
$buzz_2$ = number of message board comments about the movie at ComingSoon.net,
$buzz_3$ = total number of "can't wait" (for release) plus "don't care" votes at Fandango.com,
$buzz_4$ = percentage of Fandango votes that are "can't wait."

[15] The assertion that "nobody knows" will be tested on a newly (April, 2010) formed futures exchange where investors can place early bets on movie success (and producers can hedge their own bets). See http://www.cantorexchange.com/ for discussion. The real money exchange was created by Cantor Fitzgerald after they purchased the popular culture website Hollywood Stock Exchange.

We have aggregated these into a single principal component as follows: We first computed the logs of $buzz_1$ – $buzz_3$ to remove the scale effects. We then standardized the four variables, so z_k contains the original variable minus its mean, $\bar{z}_k$, then divided by its standard deviation, s_k. Let **Z** denote the resulting 62×4 matrix (z_1,z_2,z_3,z_4). Then **V** = (1/61)**Z'Z** is the sample correlation matrix. Let $\mathbf{c}_1$ be the characteristic vector of **V** associated with the largest characteristic root. The first principal component (the one that explains most of the variation of the four variables) is $\mathbf{Zc}_1$. (The roots are 2.4142, 0.7742, 0.4522, 0.3585 so the first principal component explains 2.4142/4 or 60.3% of the variation. Table 4.8 shows the regression results for the sample of 62 2009 movies. It appears that internet buzz adds substantially to the predictive power of the regression. The R^2 of the regression nearly doubles, from .34 to .58 when Internet buzz is added to the model. As we will discuss in Chapter 5, buzz is also a highly "significant" predictor of success.

TABLE 4.8 Regression Results for Movie Success

	Internet Buzz Model			Traditional Model		
e'e	22.30215			35.66514		
R^2	.58883			.34247		
Variable	Coefficient	Std.Error	t	Coefficient	Std.Error	t
Constant	15.4002	.64273	23.96	13.5768	.68825	19.73
ACTION	-.86932	.29333	-2.96	-.30682	.34401	-.89
COMEDY	-.01622	.25608	-.06	-.03845	.32061	-.12
ANIMATED	-.83324	.43022	-1.94	-.82032	.53869	-1.52
HORROR	.37460	.37109	1.01	1.02644	.44008	2.33
G	.38440	.55315	.69	.25242	.69196	.36
PG	.53359	.29976	1.78	.32970	.37243	.89
PG13	.21505	.21885	.98	.07176	.27206	.26
LOGBUDGT	.26088	.18529	1.41	.70914	.20812	3.41
SEQUEL	.27505	.27313	1.01	.64368	.33143	1.94
STARPOWR	.00433	.01285	.34	.00648	.01608	.40
BUZZ	.42906	.07839	5.47			

4.7.4

~~4.8~~ Missing Values, ~~Measurement Error~~ and Data Imputation

It is common for data sets to have gaps, for a variety of reasons. Perhaps the most frequent occurrence of this problem is in survey data, in which respondents may simply fail to respond to the questions. In a time series, the data may be missing because they do not exist at the frequency we wish to observe them; for example, the model may specify monthly relationships, but some variables are observed only quarterly. In panel data sets, the gaps in the data may arise because of **attrition** from the study. This is particularly common in health and medical research, when individuals choose to leave the study—possibly because of the success or failure of the treatment that is being studied.

There are several possible cases to consider, depending on why the data are missing. The data may be simply unavailable, for reasons unknown to the analyst and unrelated to the completeness or the values of the other observations in the sample. This is the most benign situation. If this is the case, then the complete observations in the sample constitute a usable data set, and the only issue is what possibly helpful information could be salvaged from the incomplete observations. Griliches (1986) calls this the **ignorable case** in that, for purposes of estimation, if we are not concerned with efficiency, then we may simply delete the incomplete observations and ignore the problem. Rubin (1976, 1987) and Little and Rubin (1987, 2002) label this case **missing completely at random,** or *MCAR*. A second case, which has attracted a great deal of attention in the econometrics literature, is that in which the gaps in the data set are not benign but are systematically related to the phenomenon being modeled. This case happens most often in surveys when the data are "self-selected" or "self-reported."[16] For example, if a survey were designed to study expenditure patterns and if high-income individuals tended to withhold information about their income, then the gaps in the data set would represent more than just missing information. The clinical trial case is another instance. In this (worst) case, the complete observations would be qualitatively different from a sample taken at random from the full population. The missing data in this situation are termed **not missing at random,** or *NMAR*. We treat this second case in Chapter 18 with the subject of **sample selection,** so we shall defer our discussion until later.

The intermediate case is that in which there is information about the missing data contained in the complete observations that can be used to improve inference about the model. The incomplete observations in this **missing at random** (MAR) case are also ignorable, in the sense that unlike the *NMAR* case, simply using the complete data does not induce any biases in the analysis, so long as the underlying process that produces the missingness in the data does not share parameters with the model that is being estimated, which seems likely. [See Allison (2002).] This case is unlikely, of course, if "missingness" is based on the values of the dependent variable in a regression. Ignoring the incomplete observations when they are *MAR* but not *MCAR* does ignore information that is in the sample and therefore sacrifices some efficiency. Researchers have used a variety of **data imputation** methods to fill gaps in data sets. The (by far) simplest case occurs when the gaps occur in the data on the regressors. For the case of missing data on the regressors, it helps to consider the simple regression and multiple regression cases separately. In the first case, $\mathbf{X}$ has two columns: the column of 1s for the constant and a column with some blanks where the missing data would be if we had them. The **zero-order method** of replacing each missing x with $\bar{x}$ based on the observed data results in no changes and is equivalent to dropping the incomplete data. (See Exercise 7 in Chapter 3.) However, the R^2 will be lower. An alternative, **modified zero-order regression** fills the second column of $\mathbf{X}$ with zeros and adds a variable

[16] The vast surveys of Americans' opinions about sex by Ann Landers (1984, passim) and Shere Hite (1987) constitute two celebrated studies that were surely tainted by a heavy dose of self-selection bias. The latter was pilloried in numerous publications for purporting to represent the population at large instead of the opinions of those strongly enough inclined to respond to the survey. The former was presented with much greater modesty.

that takes the value 1 for missing observations and zero for complete ones.[15] We leave it as an exercise to show that this is algebraically identical to simply filling the gaps with $\bar{x}$. There also is the possibility of computing fitted values for the missing x's by a regression of x on y in the complete data. The sampling properties of the resulting estimator are largely unknown, but what evidence there is suggests that this is not a beneficial way to proceed.[16]

These same methods can be used when there are multiple regressors. Once again, it is tempting to replace missing values of $\mathbf{x}_k$ with simple means of complete observations or with the predictions from linear regressions based on other variables in the model for which data are available when $\mathbf{x}_k$ is missing. In most cases in this setting, a general characterization can be based on the principle that for any missing observation, the "true" unobserved x_{ik} is being replaced by an erroneous proxy that we might view as $\hat{x}_{ik} = x_{ik} + u_{ik}$, that is, in the framework of **measurement error.** Generally, the least squares estimator is biased (and inconsistent) in the presence of measurement error such as this. (We will explore the issue in Chapter 17.) A question does remain: Is the bias likely to be reasonably small? As intuition should suggest, it depends on two features of the data: (a) how good the prediction of x_{ik} is in the sense of how large the variance of the measurement error, u_{ik}, is compared to that of the actual data, x_{ik}, and (b) how large a proportion of the sample the analyst is filling.

The regression method replaces each missing value on an $\mathbf{x}_k$ with a single prediction from a linear regression of $\mathbf{x}_k$ on other exogenous variables — in essence, replacing the missing x_{ik} with an estimate of it based on the regression model. In a Bayesian setting, some applications that involve unobservable variables (such as our example for a binary choice model in Chapter 17) use a technique called **data augmentation** to treat the unobserved data as unknown "parameters" to be estimated with the structural parameters, such as $\boldsymbol{\beta}$ in our regression model. Building on this logic researchers, e.g., Rubin (1987) and Allison (2002) have suggested taking a similar approach in classical estimation settings. The technique involves a data imputation step that is similar to what was suggested earlier, but with an extension that recognizes the variability in the estimation of the regression model used to compute the predictions. To illustrate, we consider the case in which the independent variable, $\mathbf{x}_k$ is drawn in principle from a normal population, so it is a continuously distributed variable with a mean, a variance, and a joint distribution with other variables in the model. Formally, an imputation step would involve the following calculations:

(1) Using as much information (complete data) as the sample will provide, linearly regress $\mathbf{x}_k$ on other variables in the model (and/or outside it, if other information is available), $\mathbf{Z}_k$, and obtain the coefficient vector $\mathbf{d}_k$ with associated asymptotic covariance matrix $\mathbf{A}_k$ and estimated disturbance variance s_k^2.

(2) For purposes of the imputation, we draw an observation from the estimated asymptotic normal distribution of $\mathbf{d}_k$, that is $\mathbf{d}_{k,m} = \mathbf{d}_k + \mathbf{v}_k$ where $\mathbf{v}_k$ is a vector of random draws from the normal distribution with mean zero and covariance matrix $\mathbf{A}_k$.

(3) For each missing observation in $\mathbf{x}_k$ that we wish to impute, we compute, $x_{i,k,m} = \mathbf{d}_{k,m}'\mathbf{z}_{i,k} + s_{k,m}u_{i,k}$ where $s_{k,m}$ is s_k divided by a random draw from the chi squared distribution with degrees of freedom equal the number of degrees of freedom in the imputation regression.

[15] See Maddala (1977a, p. 202).

[16] Afifi and Elashoff (1966, 1967) and Haitovsky (1968). Griliches (1986) considers a number of other possibilities.

AU: Do you mean "inputted" - to enter data - rather than "imputed" - to blame? Check KT also

At this point, the iteration is the same as considered earlier, where the missing values are imputed using a regression, albeit, a much more elaborate procedure. The regression is then computed using the complete data and the imputed data for the missing observations, to produce coefficient vector $\mathbf{b}_m$ and estimated covariance matrix, $\mathbf{V}_m$. This constitutes a single round. The technique of **multiple imputation** involves repeating this set of steps M times. The estimators of the parameter vector and the appropriate asymptotic covariance matrix are

KT

AU: Term "multiple imputation" not in chap list.

$$\hat{\boldsymbol{\beta}} = \bar{\mathbf{b}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{b}_m,$$

$$\hat{\mathbf{V}} = \bar{\mathbf{V}} + \mathbf{B} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{V}_m + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{m=1}^{M}(\mathbf{b}_m - \bar{\mathbf{b}})(\mathbf{b}_m - \bar{\mathbf{b}})'.$$

Researchers differ on the effectiveness or appropriateness of multiple imputation. When all is said and done, the measurement error in the imputed values remains. It takes very strong assumptions to establish that the multiplicity of iterations will suffice to average away the effect of this error. Very elaborate techniques have been developed for the special case of joint normally distributed cross sections of regressors such as suggested above. However, the typical application to survey data involves gaps due to nonresponse to qualitative questions with binary answers. The efficacy of the theory is much less well developed for imputation of binary, ordered, count or other qualitative variables.

The more manageable case is missing values of the dependent variable, y_i. Once again, it must be the case that y_i is at least *MAR* and that the mechanism that is determining presence in the sample does not share parameters with the model itself. Assuming the data on $\mathbf{x}_i$ are complete for all observations, one might consider filling the gaps in the data on y_i by a two-step procedure: (1) estimate $\boldsymbol{\beta}$ with $\mathbf{b}_c$ using the complete observations, $\mathbf{X}_c$ and $\mathbf{y}_c$, then (2) fill the missing values, $\mathbf{y}_m$, with predictions, $\hat{\mathbf{y}}_m = \mathbf{X}_m\mathbf{b}_c$, and recompute the coefficients. We leave as an exercise (Exercise 17) to show that the second step estimator is exactly equal to the first. However, the variance estimator at the second step, s^2 must underestimate σ^2, intuitively because we are adding to the sample a set of observations that are fit perfectly. [See Cameron and Trivedi (2005, Chapter 27).] So, this is not a beneficial way to proceed. The flaw in the method comes back to the device used to impute the missing values for y_i. Recent suggestions that appear to provide some improvement involve using a randomized version, $\hat{\mathbf{y}}_m = \mathbf{X}_m\mathbf{b}_c + \hat{\boldsymbol{\varepsilon}}_m$, where $\hat{\boldsymbol{\varepsilon}}_m$ are random draws from the (normal) population with zero mean and estimated variance $s^2[\mathbf{I} + \mathbf{X}_m(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_m']$. (The estimated variance matrix corresponds to $\mathbf{X}_m\mathbf{b}_c + \boldsymbol{\varepsilon}_m$.) This defines an iteration. After reestimating $\boldsymbol{\beta}$ with the augmented data, one can return to re-impute the augmented data with the new $\hat{\boldsymbol{\beta}}$, then recompute $\mathbf{b}$, and so on. The process would continue until the estimated parameter vector stops changing. (A subtle point to be noted here: The same random draws should be used in each iteration. If not, there is no assurance that the iterations would ever converge.)

In general, not much is known about the properties of estimators based on using predicted values to fill missing values of y. Those results we do have are largely from simulation studies based on a particular data set or pattern of missing data. The results of these Monte Carlo studies are usually difficult to generalize. The overall conclusion seems to be that in a single-equation regression context, filling in missing values of y leads to biases in the estimator which are difficult to quantify. The only reasonably clear result is that imputations are more likely to be beneficial if the proportion of observations that are being filled is small—the smaller the better.

4.7.5 Measurement Error and Proxy Variables

There are any number of cases in which observed data are imperfect measures of their theoretical counterparts in the regression model. Examples include income, education, ability, health, "the interest rate," output, capital, and so on. Mismeasurement of the variables in a model will generally produce adverse consequences for least squares estimation. Remedies are complicated and sometimes require heroic assumptions. In this section, we will provide a brief sketch of the issues. We defer to Section 8.5 a more detailed discussion of the problem of measurement error, the most common solution (instrumental variables estimation) and some applications.

It is convenient to distinguish between measurement error in the dependent variable and measurement error in the regressor(s). For the second case, it is also useful to consider the simple regression case then extend it to the multiple regression model. Consider a model to describe expected income in a population,

$$I^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon \qquad (4\text{-}55)$$

where I^* is the intended total income variable. Suppose the observed counterpart is I, earnings. How I relates to I^* is unclear; it is common to assume that the measurement error is additive, so $I = I^* + w$. Inserting the expression for I into (4-55) gives

$$\begin{aligned} I &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon + w \\ &= \mathbf{x}'\boldsymbol{\beta} + v, \end{aligned} \qquad (4\text{-}56)$$

which appears to be a slightly more complicated regression, but otherwise similar to what we started with. As long as w and $\mathbf{x}$ are uncorrelated, that is the case. If w is a homoscedastic, zero mean error that is uncorrelated with $\mathbf{x}$, then the only difference between (4-55) and (4-56) is that the disturbance variance in (4-56) is $\sigma_w^2 + \sigma_\varepsilon^2 > \sigma_\varepsilon^2$. Otherwise both are regressions and, evidently $\boldsymbol{\beta}$ can be estimated consistently by least squares in either case. The cost of the measurement error is in the precision of the estimator, since the asymptotic variance of the estimator in (4-56) is $(\sigma_v^2/n)[\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$ while it is $(\sigma_\varepsilon^2/n)[\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$ if $\boldsymbol{\beta}$ is estimated using (4-55). The cost of the measurement error also costs some fit. To see this, note that the R^2 in the sample regression in (4-55) is

$$R_*^2 = 1 - (\mathbf{e}'\mathbf{e}/n)/(\mathbf{I}^{*\prime}\mathbf{M}^0\mathbf{I}^*/n).$$

The numerator converges to σ_ε^2 while the denominator converges to the total variance of I^*, which would approach $\sigma_\varepsilon^2 + \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}$ where $\mathbf{Q} = \text{plim}(\mathbf{X}'\mathbf{X}/n)$. Therefore,

$$\text{plim } R_*^2 = \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}/[\sigma_\varepsilon^2 + \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}].$$

The counterpart for (4-56), R^2, differs only in that σ_ε^2 is replaced by $\sigma_v^2 > \sigma_\varepsilon^2$ in the denominator. It follows that

$$\text{plim } R_*^2 - \text{plim } R^2 > 0.$$

This implies that the fit of the regression in (4-56) will, at least broadly in expectation, be inferior to that in (4-55). (The preceding is an asymptotic approximation that might not hold in every finite sample.)

These results demonstrate the implications of measurement error in the dependent variable. We note, in passing, that if the measurement error is not additive, if it is correlated with $\mathbf{x}$, or if it has any other features such as heteroscedasticity, then the preceding results are lost, and nothing in general can be said about the consequence of the measurement error. Whether there is a "solution" is likewise

ambiguous question. The preceding shows that the it would be better to have the underlying variable if possible. In the absence, would it be preferable to use a proxy? Unfortunately, I already is a proxy, so unless there exists an available I' which has smaller measurement error variance, we have reached an impasse. On the other hand, it does seem that the outcome is fairly benign. The sample does not contain as much information as we might hope, but it does contain sufficient information consistently to estimate β and to do appropriate statistical inference based on the information we do have.

The more difficult case occurs when the measurement error appears in the independent variable(s). For simplicity, we retain the symbols I and I^* for our observed and theoretical variables. Consider a simple regression,

$$y = \beta_1 + \beta_2 I^* + \varepsilon,$$

where y is the perfectly measured dependent variable and the same measurement equation, $I = I^* + w$ applies now to the independent variable. Inserting I into the equation and rearranging a bit, we obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 I + (\varepsilon - \beta_2 w) \\ &= \beta_1 + \beta_2 I + v. \end{aligned} \qquad (4\text{-}57)$$

It appears that we have obtained (4-56) once again. Unfortunately, this is not the case, because $\text{Cov}[I,v] = \text{Cov}[I^* + w, \varepsilon - \beta_2 w] = -\beta_2 \sigma_w^2$. Since the regressor in (4-57) is correlated with the disturbance, least squares regression in this case is inconsistent. There is a bit more that can be derived – this is pursued in Section 8.5, so we state it here without proof. In this case,

$$\text{plim } b_2 = \beta_2 [\sigma_*^2 / (\sigma_*^2 + \sigma_w^2)]$$

where σ_*^2 is the marginal variance of I^*. The scale factor is less than one, so the least squares estimator is biased twoard zero. The larger is the measurement error variance, the worse is the bias. (This is called **least squares attenuation**.) Now, suppose there are additional variables in the model;

$$y = \mathbf{x}'\boldsymbol{\beta}_1 + \beta_2 I^* + \varepsilon.$$

In this instance, almost no useful theoretical results are forthcoming. The following fairly general conclusions can be drawn – once again, proofs are deferred to Section 8.5:

(1) The least squares estimator of β_2 is still biased toward zero.
(2) All the elements of the estimator of $\boldsymbol{\beta}_1$ are biased, in unknown directions, even though the variables in $\mathbf{x}$ are not measured with error.

Solutions to the "measurement error problem" come in two forms. If there is outside information on certain model parameters, then it is possible to deduce the scale factors (using the **method of moments**) and undo the bias. For the obvious example, in (4-57), if σ_w^2 were known, then it would be possible to deduce σ_*^2 from $\text{Var}[I] = \sigma_*^2 + \sigma_w^2$ and thereby compute the necessary scale factor to undo the bias. This sort of information is generally not available. A second approach that has been used in many applications is the technique of instrumental variables. This is developed in detail for this setting in Section 8.5.

4.7.6 Outliers and Influential Observations

Figure 4.9 shows a scatter plot of the data on sale prices of Monet paintings that were used in Example 4.10. Two points have been highlighted. The one marked "I" and noted with the square overlay shows the smallest painting in the data set. The circle marked "O" highlights a painting that fetched an unusually low price, at least in comparison to what the regression would have predicted. (It was not the least costly painting in the sample; but, it was the one most poorly predicted by the regression.) Since least squares is based on squared deviations, the estimator is likely to be strongly influenced by extreme observations such as these, particularly if the sample is not very large.

An "influential observation" is one that is likely to have a substantial impact on the least squares regression coefficient(s). For a simple regression such as the one shown in Figure 4.9, Belsley, Kuh and Welsh (1980) defined an influence measure, for observation i,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{\sum_{j=1}^{n}(x_j - \bar{x}_n)^2} \qquad \textbf{(4-58)}$$

where $\bar{x}_n$ and the summation in the denominator of the fraction are computed without this observation. (The measure derives from the difference between $\mathbf{b}$ and $\mathbf{b}_{(i)}$ where the latter is computed without the particular observation. We will return to this shortly.) It is suggested that an observation should be noted as influential if $h_i > 2/n$. The decision is whether to drop the observation or not. We should note, observations with high "leverage" are arguably not "outliers" (which remains to be defined), because the analysis is conditional on x_i. To underscore the point, referring to Figure 4.9, this observation would be marked even if it fell precisely on the regression line -- the source of the influence is the numerator of the second term in h_i, which is unrelated to the distance of the point from the line. In our example, the "influential observation" happens to be the result of Monet's decision to paint a small painting. The point is that in the absence of an underlying theory that explains (and justifies) the extreme values of x_i, eliminating such observations is an algebraic exercise that has the effect of forcing the regression line to be fitted with the values of x_i closest to the means.

The change in the linear regression coefficient vector in a multiple regression when an observation is added to the sample is

$$\mathbf{b} - \mathbf{b}_{(i)} = \Delta\mathbf{b} = \frac{1}{1 + \mathbf{x}_i'\left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{x}_i}\left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{x}_i\left(y_i - \mathbf{x}_i'\mathbf{b}_{(i)}\right) \qquad \textbf{(4-59)}$$

where $\mathbf{b}$ is computed with observation i in the sample, $\mathbf{b}_{(i)}$ is computed without observation i and $\mathbf{X}_{(i)}$ does not include observation i. (See Exercise 6 in Chapter 3.) It is difficult to single out any particular feature of the observation that would drive this change. The influence measure,

$$\begin{aligned} h_{ii} &= \mathbf{x}_i'\left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{x}_i \\ &= \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}\left(x_{i,j} - \bar{x}_{n,j}\right)\left(x_{i,k} - \bar{x}_k\right)\left(\mathbf{Z}_{(i)}'\mathbf{M}^0\mathbf{Z}_{(i)}\right)^{jk} \end{aligned} \qquad \textbf{(4-60)}$$

has been used to flag influential observations. (See, once again, Belsley, Kuh and Welsh (1980) and Cook (1977).) In this instance, the selection criterion would be $h_{ii} > 2(K\text{-}1)/n$. Squared deviations of the elements of $\mathbf{x}_i$ from the means of the variables appear in h_{ii}, so it is also operating on the difference of $\mathbf{x}_i$ from the center of the data. [See the expression for the forecast variance in Section 4.6.1 for an application.]

In principle, an "outlier," is an observation that appears to be outside the reach of the model, perhaps because it arises from a different data generating process. Point "O" in Figure 4.9 appears to be a candidate. Outliers could arise for several reasons. The simplest explanation would be actual data errors. Assuming the data are not erroneous, it then remains to define what constitutes an outlier. Unusual residuals are an obvious choice. But, since the distribution of the disturbances would anticipate for a certain small percentage of extreme observations in any event, simply singling out observations with large residuals is actually a dubious exercise. On the other hand, one might suspect that the outlying observations are actually generated by a different population. "Studentized" residuals are constructed with this in mind by computing the regression coefficients and the residual variance without observation i for each observation in the sample, then standardizing the modified residuals. The ith studentized residual is

$$e(i) = \frac{e_i}{(1-h_{ii})} \Big/ \sqrt{\frac{\mathbf{e}'\mathbf{e} - e_i^2/(1-h_{ii})}{n-1-K}} \qquad (4-61)$$

where $\mathbf{e}$ is the residual vector for the full sample, based on $\mathbf{b}$, including e_i the residual for observation i. In principle, this residual has a t distribution with n-1-K degrees of freedom (or a standard normal distribution asymptotically). Observations with large studentized residuals, i.e., greater than 2.0, would be singled out as outliers.

There are several complications that arise with isolating outlying observations in this fashion. First, there is no a priori assumption which observations are from the alternative population, if this is the view. From a theoretical point of view, this would suggest a skepticism about the model specification. If the sample contains a substantial proportion of outliers, then the properties of the estimator based on the reduced sample are difficult to derive. In our application below, the procedure deletes 4.7% of the sample (20 observations). Finally, it will usually occur that observations that were not outliers in the original sample will become "outliers" when the original set of outliers is removed. It is unclear how one should proceed at this point. (Using the Monet paintings data, the first round of studentizing the residuals removes 20 observations. After 16 iterations, the sample size stabilizes at 316 of the original 430 observations, a reduction of 26.5%.) Table 4.9 shows the original results (from Table 4.6) and the modified results with 20 outliers removed. Since 430 is a relatively large sample, the modest change in the results is to be expected.

It is difficult to draw a firm general conclusions from this exercise. It remains likely that in very small samples, some caution and close scrutiny of the data are called for. If it is suspected at the outset that a process that is prone to large observations is at work, it may be useful to consider a different estimator altogether, such as least absolute deviations, or even a different model specification that accounts for this possibility. For example, the idea that the sample may contain some observations that are generated by a different process lies behind the latent class model that is discussed in Chapters 14 and 18.

TABLE 4.9 Estimated Equations for Log Price

Number of observations	430	410
Mean of log Price	0.33274	.36043
Sum of squared residuals	519.17235	383.17982
Standard error of regression	1.10266	0.97030
R-squared	0.33620	0.39170
Adjusted R-squared	0.33309	0.38871

Variable	Coefficient		Standard Error		t	
	n=430	n=410	n=430	n=410	n=430	n=410
Constant	-8.42653	-8.67356	.61183	.57529	-13.77	-15.08
LOGAREA	1.33372	1.36982	.09072	.08472	14.70	16.17
ASPECT	-.16537	-.14383	.12753	.11412	-1.30	-1.26

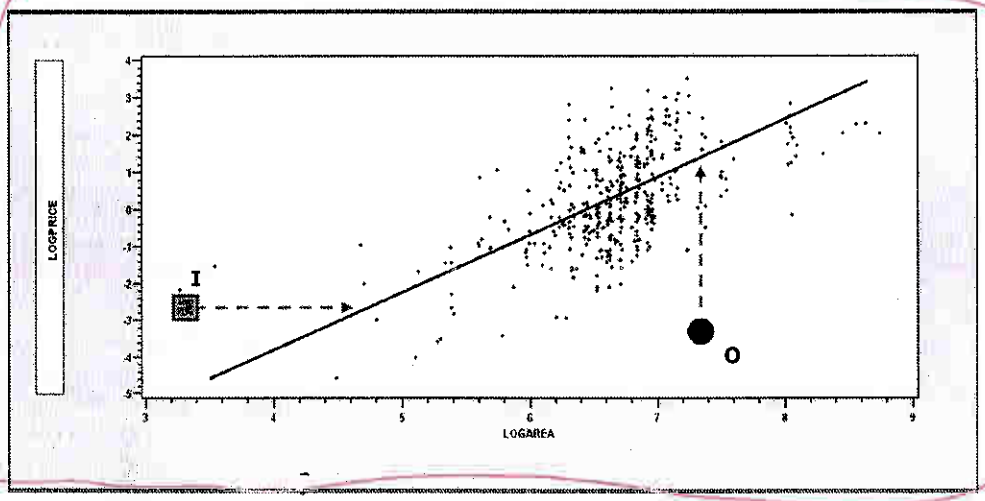

Figure 4.9 Log Price vs. Log Area for Monet Paintings

4.8 SUMMARY AND CONCLUSIONS

This chapter has examined a set of properties of the least squares estimator that will apply in all samples, including unbiasedness and efficiency among unbiased estimators. The formal assumptions of the linear model are pivotal in the results of this chapter. All of them are likely to be violated in more general settings than the one considered here. For example, in most cases examined later in the book, the estimator has a possible bias, but that bias diminishes with increasing sample sizes. For purposes of forming confidence intervals and testing hypotheses, the assumption of normality is narrow, so it was necessary to extend the model to allow nonnormal disturbances. These and other "large sample" extensions of the linear model were considered in Section 4.4. The crucial results developed here were the consistency of the estimator and a method of obtaining an appropriate covariance matrix and large sample distribution that provides the basis for forming confidence intervals and testing hypotheses. Statistical inference in the form of interval estimation for the model parameters and for values of the dependent variable were considered in Sections 4.5 and 4.6. This development will continue in Chapter 5 where we will consider hypothesis testing and model selection.

Finally, we considered some practical problems that arise when data are less than perfect for the estimation and analysis of the regression model, including multicollinearity, missing observations, measurement error and outliers.

Key Terms and Concepts

- Assumptions
- Asymptotic covariance matrix
- Asymptotic distribution
- Asymptotic efficiency
- Asymptotic normality
- Asymptotic properties
- Attrition
- Confidence interval
- Consistency
- Consistent estimator
- Data imputation
- Efficient scale

Au: The following terms were not bold KTs in text:

Assumptions
Asymptotic covariance matrix
Asymptotic efficiency
Confidence interval
~~Ergodic~~
Gauss–Markov theorem
~~Identification~~
~~Indicator~~
Lindeberg–Feller central limit theorem
Mean squared error
Minimum mean squared error
Minimum variance, linear unbiased estimator
~~Missing observations~~
~~Orthogonal random variables~~
~~Panel data~~
Sampling distribution
~~Stationary process~~
Stochastic regressors
~~t ratio~~

4-70

- Ergodic
- Estimator
- Finite sample properties
- Gauss–Markov theorem
- Grenander conditions
- Identification
- Ignorable case
- Indicator
- Lindeberg–Feller central limit theorem
- Linear estimator
- Linear unbiased estimator
- Maximum likelihood estimator
- Mean square convergence
- Mean squared error
- Measurement error
- Minimum mean squared error
- Minimum variance linear unbiased estimator
- Missing at random
- Missing completely at random
- Missing observations
- Modified zero-order regression
- Multicollinearity
- Not missing at random
- Oaxaca's and Blinder's decomposition
- Optimal linear predictor
- Orthogonal random variables
- Panel data
- Principal components
- Probability limit
- Sample selection
- Sampling distribution
- Sampling variance
- Semiparametric
- Standard error
- Standard error of the regression
- Stationary process
- Statistical properties
- Stochastic regressors
- t ratio
- Variance inflation factor
- Zero-order method

add new terms on next page msp 4-71

Exercises

1. Suppose that you have two independent unbiased estimators of the same parameter θ, say $\hat{\theta}_1$ and $\hat{\theta}_2$, with different variances v_1 and v_2. What linear combination $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ is the minimum variance unbiased estimator of θ?
2. Consider the simple regression $y_i = \beta x_i + \varepsilon_i$ where $E[\varepsilon \mid x] = 0$ and $E[\varepsilon^2 \mid x] = \sigma^2$
   a. What is the minimum mean squared error linear estimator of β? [Hint: Let the estimator be $(\hat{\beta} = \mathbf{c}'\mathbf{y})$. Choose $\mathbf{c}$ to minimize $\text{Var}(\hat{\beta}) + (E(\hat{\beta} - \beta))^2$. The answer is a function of the unknown parameters.]
   b. For the estimator in part a, show that ratio of the mean squared error of $\hat{\beta}$ to that of the ordinary least squares estimator b is

$$\frac{\text{MSE}[\hat{\beta}]}{\text{MSE}[b]} = \frac{\tau^2}{(1+\tau^2)}, \quad \text{where } \tau^2 = \frac{\beta^2}{[\sigma^2/\mathbf{x}'\mathbf{x}]}.$$

   Note that τ is the square of the population analog to the "t ratio" for testing the hypothesis that $\beta = 0$, which is given in (4-14). How do you interpret the behavior of this ratio as $\tau \to \infty$?
3. Suppose that the classical regression model applies but that the true value of the constant is zero. Compare the variance of the least squares slope estimator computed without a constant term with that of the estimator computed with an unnecessary constant term.
4. Suppose that the regression model is $y_i = \alpha + \beta x_i + \varepsilon_i$, where the disturbances ε_i have $f(\varepsilon_i) = (1/\lambda)\exp(-\varepsilon_i/\lambda)$, $\varepsilon_i \geq 0$. This model is rather peculiar in that all the disturbances are assumed to be nonnegative. Note that the disturbances have $E[\varepsilon_i \mid x_i] = \lambda$ and $\text{Var}[\varepsilon_i \mid x_i] = \lambda^2$. Show that the least squares slope is unbiased but that the intercept is biased.
5. Prove that the least squares intercept estimator in the classical regression model is the minimum variance linear unbiased estimator.

Insert on msp 4-69 and 4-70 where indicated

- Prediction error
- Pivotal statistic
- Prediction interval
- Prediction Variance
- Ex ante forecast
- Ex post forecast
- Root mean squared error
- Mean absolute error
- Theil U statistic
- Bootstrap
- Point estimation
- Interval estimation

- Highest posterior density interval

AU; KT in text on msp 4-32 is "bootstrapping." Here also?

6. As a profit-maximizing monopolist, you face the demand curve $Q = \alpha + \beta P + \varepsilon$. In the past, you have set the following prices and sold the accompanying quantities:

Q	3	3	7	6	10	15	16	13	9	15	9	15	12	18	21
P	18	16	17	12	15	15	4	13	11	6	8	10	7	7	7

Suppose that your marginal cost is 10. Based on the least squares regression, compute a 95 percent confidence interval for the expected value of the profit-maximizing output.

7. The following sample moments for $x = [1, x_1, x_2, x_3]$ were computed from 100 observations produced using a random number generator:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 100 & 123 & 96 & 109 \\ 123 & 252 & 125 & 189 \\ 96 & 125 & 167 & 146 \\ 109 & 189 & 146 & 168 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 460 \\ 810 \\ 615 \\ 712 \end{bmatrix}, \quad \mathbf{y}'\mathbf{y} = 3924.$$

The true model underlying these data is $y = x_1 + x_2 + x_3 + \varepsilon$.

a. Compute the simple correlations among the regressors.
b. Compute the ordinary least squares coefficients in the regression of y on a constant x_1, x_2, and x_3.
c. Compute the ordinary least squares coefficients in the regression of y on a constant x_1 and x_2, on a constant x_1 and x_3, and on a constant x_2 and x_3.
d. Compute the variance inflation factor associated with each variable.
e. The regressors are obviously collinear. Which is the problem variable?

8. Consider the multiple regression of $\mathbf{y}$ on K variables $\mathbf{X}$ and an additional variable $\mathbf{z}$. Prove that under the assumptions A1 through A6 of the classical regression model, the true-variance of the least squares estimator of the slopes on $\mathbf{X}$ is larger when $\mathbf{z}$ is included in the regression than when it is not. Does the same hold for the sample estimate of this covariance matrix? Why or why not? Assume that $\mathbf{X}$ and $\mathbf{z}$ are nonstochastic and that the coefficient on $\mathbf{z}$ is nonzero.
9. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, assuming that the true value of $\boldsymbol{\beta}$ is zero, what is the exact expected value of $F[K, n-K] = (R^2/K)/[(1-R^2)/(n-K)]$?
10. Prove that $E[\mathbf{b}'\mathbf{b}] = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{k=1}^{K} (1/\lambda_k)$ where $\mathbf{b}$ is the ordinary least squares estimator and λ_k is a characteristic root of $\mathbf{X}'\mathbf{X}$.
11. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, what is plim $F[K, n-K]$ = plim $\frac{R^2/K}{(1-R^2)/(n-K)}$, assuming that the true value of $\boldsymbol{\beta}$ is zero?
12. Let e_i be the ith residual in the ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ in the classical regression model, and let ε_i be the corresponding true disturbance. Prove that plim$(e_i - \varepsilon_i) = 0$.
13. For the simple regression model $y_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N[0, \sigma^2]$, prove that the sample mean is consistent and asymptotically normally distributed. Now consider the alternative estimator $\hat{\mu} = \sum_i w_i y_i$, $w_i = \frac{i}{(n(n+1)/2)} = \frac{i}{\sum_i i}$. Note that $\sum_i w_i = 1$. Prove that this is a consistent estimator of μ and obtain its asymptotic variance. [Hint: $\sum_i i^2 = n(n+1)(2n+1)/6$.]

4-73

14. In the discussion of the instrumental variables estimator, we showed that the least squares estimator $\mathbf{b}$ is biased and inconsistent. Nonetheless, $\mathbf{b}$ does estimate something: plim $\mathbf{b} = \boldsymbol{\theta} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\boldsymbol{\gamma}$. Derive the asymptotic covariance matrix of $\mathbf{b}$, and show that $\mathbf{b}$ is asymptotically normally distributed.
15. Suppose we change the assumptions of the model to **AS5**: $(\mathbf{x}_i, \varepsilon)$ are an independent and identically distributed sequence of random vectors such that $\mathbf{x}_i$ has a finite mean vector, $\boldsymbol{\mu}_{\mathbf{x}}$, finite positive definite covariance matrix $\boldsymbol{\Sigma}_{\mathbf{xx}}$ and finite fourth moments $E[x_j x_k x_l x_m] = \phi_{jklm}$ for all variables. How does the proof of consistency and asymptotic normality of $\mathbf{b}$ change? Are these assumptions weaker or stronger than the ones made in Section 4.1?
16. Now, assume only finite second moments of $\mathbf{x}$; $E[x_i^2]$ is finite. Is this sufficient to establish consistency of $\mathbf{b}$? (Hint: the Cauchy–Schwarz inequality (Theorem D.13), $E[|xy|] \leq \{E[x^2]\}^{1/2}\{E[y^2]\}^{1/2}$ will be helpful.) Is this assumption sufficient to establish asymptotic normality?
17. Consider a data set consisting of n observations, n_c complete and n_m incomplete for which the dependent variable, y_i, is missing. Data on the independent variables, $\mathbf{x}_i$, are complete for all n observations, $\mathbf{X}_c$ and $\mathbf{X}_m$. We wish to use the data to estimate the parameters of the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Consider the following the imputation strategy: Step 1: Linearly regress $\mathbf{y}_c$ on $\mathbf{X}_c$ and compute $\mathbf{b}_c$. Step 2: Use $\mathbf{X}_m$ to predict the missing $\mathbf{y}_m$ with $\mathbf{X}_m\mathbf{b}_c$. Then regress the full sample of observations, $(\mathbf{y}_c, \mathbf{X}_m\mathbf{b}_c)$, on the full sample of regressors, $(\mathbf{X}_c, \mathbf{X}_m)$.
    a. Show that the first and second step least squares coefficient vectors are identical.
    b. Is the second step coefficient estimator unbiased?
    c. Show that the sum of squared residuals is the same at both steps.
    d. Show that the second step estimator of σ^2 is biased downward.

14

insert next page msp 4/14

Applications

1. Data on U.S. gasoline consumption for the years 1953 to 2004 are given in Table F2.2. Note, the consumption data appear as total expenditure. To obtain the per capita quantity variable, divide GASEXP by GASP times Pop. The other variables do not need transformation.
    a. Compute the multiple regression of per capita consumption of gasoline on per capita income, the price of gasoline, all the other prices and a time trend. Report all results. Do the signs of the estimates agree with your expectations?
    b. Test the hypothesis that at least in regard to demand for gasoline, consumers do not differentiate between changes in the prices of new and used cars.
    c. Estimate the own price elasticity of demand, the income elasticity, and the cross-price elasticity with respect to changes in the price of public transportation. Do the computations at the 2004 point in the data.
    d. Reestimate the regression in logarithms so that the coefficients are direct estimates of the elasticities. (Do not use the log of the time trend.) How do your estimates compare with the results in the previous question? Which specification do you prefer?
    e. Compute the simple correlations of the price variables. Would you conclude that multicollinearity is a "problem" for the regression in part a or part d?

Insert on msp 4-73 where indicated

15 18. In (4-13), we find that when superfluous variables $\mathbf{X}_2$ are added to the regression of $\mathbf{y}$ on $\mathbf{X}_1$ the least squares coefficient estimator is an unbiased estimator of the true parameter vector, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \mathbf{0}')'$. Show that in this long regression, $\mathbf{e}'\mathbf{e}/(n-K_1-K_2)$ is also unbiased, as estimator of σ^2.

16 19. In Section 4.7.~~1~~3, we consider regressing $\mathbf{y}$ on a set of principal components, rather than the original data. For simplicity, assume that $\mathbf{X}$ does not contain a constant term, and that the K variables are measured in deviations from the means and are "standardized" by dividing by the respective standard deviations. We consider regression of $\mathbf{y}$ on L principal components, $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$, where $L < K$. Let $\mathbf{d}$ denote the coefficient vector. The regression model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In the discussion, it is claimed that $E[\mathbf{d}] = \mathbf{C}_L'\boldsymbol{\beta}$. Prove the claim.

17 20. Example 4.9 presents a regression model that is used to predict the auction prices of Monet paintings. The most expensive painting in the sample sold for \$33.0135M (log = 17.3124). The height and width of this painting were 35" and 39.4", respectively. Use these data and the model to form prediction intervals for the log of the price, then the price for this painting.

end of insert

f. Notice that the price index for gasoline is normalized to 100 in 2000, whereas the other price indices are anchored at 1983 (roughly). If you were to renormalize the indices so that they were all 100.00 in 2004, then how would the results of the regression in part a change? How would the results of the regression in part d change?

g. This exercise is based on the model that you estimated in part d. We are interested in investigating the change in the gasoline market that occurred in 1973. First, compute the average values of log of per capita gasoline consumption in the years 1953–1973 and 1974–2004 and report the values and the difference. If we divide the sample into these two groups of observations, then we can decompose the change in the expected value of the log of consumption into a change attributable to change in the regressors and a change attributable to a change in the model coefficients, as shown in Section 4.7.3. Using the Oaxaca–Blinder approach described there, compute the decomposition by partitioning the sample and computing separate regressions. Using your results, compute a confidence interval for the part of the change that can be attributed to structural change in the market, that is, change in the regression coefficients.

4.5.3

2. Christensen and Greene (1976) estimated a generalized Cobb–Douglas cost function for electricity generation of the form

$$\ln C = \alpha + \beta \ln Q + \gamma \left[\tfrac{1}{2}(\ln Q)^2\right] + \delta_k \ln P_k + \delta_l \ln P_l + \delta_f \ln P_f + \varepsilon.$$

P_k, P_l, and P_f indicate unit prices of capital, labor, and fuel, respectively, Q is output and C is total cost. To conform to the underlying theory of production, it is necessary to impose the restriction that the cost function be homogeneous of degree one in the three prices. This is done with the restriction $\delta_k + \delta_l + \delta_f = 1$, or $\delta_f = 1 - \delta_k - \delta_l$. Inserting this result in the cost function and rearranging produces the estimating equation,

$$\ln(C/P_f) = \alpha + \beta \ln Q + \gamma \left[\tfrac{1}{2}(\ln Q)^2\right] + \delta_k \ln(P_k/P_f) + \delta_l \ln(P_l/P_f) + \varepsilon.$$

6.6

The purpose of the generalization was to produce a U-shaped average total cost curve. [See Example 6.3 for discussion of Nerlove's (1963) predecessor to this study.] We are interested in the **efficient scale,** which is the output at which the cost curve reaches its minimum. That is the point at which $(\partial \ln C/\partial \ln Q)_{|Q=Q^*} = 1$ or $Q^* = \exp[(1-\beta)/\gamma]$.

F4.4

a. Data on 158 firms extracted from Christensen and Greene's study are given in Table F4.3. Using all 158 observations, compute the estimates of the parameters in the cost function and the estimate of the asymptotic covariance matrix.

b. Note that the cost function does not provide a direct estimate of δ_f. Compute this estimate from your regression results, and estimate the asymptotic standard error.

c. Compute an estimate of Q^* using your regression results, then form a confidence interval for the estimated efficient scale.

d. Examine the raw data and determine where in the sample the efficient scale lies. That is, determine how many firms in the sample have reached this scale, and whether, in your opinion, this scale is large in relation to the sizes of firms in

4-76 / End 4

the sample. Christensen and Greene approached this question by computing the proportion of total output in the sample that was produced by firms that had not yet reached efficient scale. (Note, there is some double counting in the data set—more than 20 of the largest "firms" in the sample we are using for this exercise are holding companies and power pools that are aggregates of other firms in the sample. We will ignore that complication for the purpose of our numerical exercise.)