Chapter 3

The Classical Linear Regression Model

3.1 Textbooks as Catalogs

In chapter 2 we learned that many of the estimating criteria held in high regard by econometricians (such as best unbiasedness and minimum mean square error) are characteristics of an estimator's sampling distribution. These characteristics cannot be determined unless a set of repeated samples can be taken or hypothesized; to take or hypothesize these repeated samples, knowledge of the way in which the observations are generated is necessary. Unfortunately, an estimator does not have the same characteristics for all ways in which the observations can be generated. This means that in some estimating situations a particular estimator has desirable properties but in other estimating situations it does *not* have desirable properties. Because there is no "superestimator" having desirable properties in all situations, for each estimating problem (i.e., for each different way in which the observations can be generated) the econometrician must determine anew which estimator is preferred. An econometrics textbook can be characterized as a catalog of which estimators are most desirable in what estimating situations. Thus, a researcher facing a particular estimating problem simply turns to the catalog to determine which estimator is most appropriate for him or her to employ in that situation. The purpose of this chapter is to explain how this catalog is structured.

The cataloging process described above is centered around a standard estimating situation referred to as the *classical linear regression model* (CLR model). It happens that in this standard situation the ordinary least squares (OLS) estimator is considered the optimal estimator. This model consists of five assumptions concerning the way in which the data are generated. By changing these assumptions in one way or another, different estimating situations are created, in many of which the OLS estimator is no longer considered to be the optimal estimator. Most econometric problems can be characterized as situations in which one (or more) of these five assumptions is violated in a particular way. The catalog works in a straightforward way: the estimating

situation is modeled in the general mold of the CLR model and the researcher pinpoints the way in which this situation differs from the standard situation as described by the CLR model (i.e., finds out which assumption of the CLR model is violated in this problem); he or she then turns to the textbook (catalog) to see whether the OLS estimator retains its desirable properties, and if not what alternative estimator should be used. Because econometricians often are not certain of whether the estimating situation they face is one in which an assumption of the CLR model is violated, the catalog also includes a listing of techniques useful in testing whether or not the CLR model assumptions are violated.

3.2 The Five Assumptions

The CLR model consists of five basic assumptions about the way in which the observations are generated.

1. The *first assumption* of the CLR model is that the dependent variable can be calculated as a linear function of a specific set of independent variables, plus a disturbance term. The unknown coefficients of this linear function form the vector β and are assumed to be constants. Several violations of this assumption, called specification errors, are discussed in chapter 6:
   (a) *Wrong regressors* – the omission of relevant independent variables or the inclusion of irrelevant independent variables.
   (b) *Nonlinearity* – when the relationship between the dependent and independent variables is not linear.
   (c) *Changing parameters* – when the parameters (β) do not remain constant during the period in which data were collected.
2. The *second assumption* of the CLR model is that the expected value of the disturbance term is zero; that is, the mean of the distribution from which the disturbance term is drawn is zero. Violation of this assumption leads to the *biased intercept* problem, discussed in chapter 7.
3. The *third assumption* of the CLR model is that the disturbance terms all have the same variance and are not correlated with one another. Two major econometric problems, discussed in chapter 8, are associated with violations of this assumption:
   (a) *Heteroskedasticity* – when the disturbances do not all have the same variance.
   (b) *Autocorrelated errors* – when the disturbances are correlated with one another.
4. The *fourth assumption* of the CLR model is that the observations on the independent variable can be considered fixed in repeated samples; that is, it is possible to redraw the sample with the same independent variable values. Three important econometric problems, discussed in chapters 10 and 11, correspond to violations of this assumption:
   (a) *Errors in variables* – errors in measuring the independent variables.
   (b) *Autoregression* – using a lagged value of the dependent variable as an independent variable.

(c) *Simultaneous equation estimation* – situations in which the dependent variables are determined by the simultaneous interaction of several relationships.

5. The *fifth assumption* of the CLR model is that the number of observations is greater than the number of independent variables and that there are no exact linear relationships between the independent variables. Although this is viewed as an assumption for the general case, for a specific case it can easily be checked, so that it need not be assumed. The problem of *multicollinearity* (two or more independent variables being approximately linearly related in the sample data) is associated with this assumption. This is discussed in chapter 12.

All this is summarized in Table 3.1, which presents these five assumptions of the CLR model, shows the appearance they take when dressed in mathematical notation, and lists the econometric problems most closely associated with violations of these assumptions. Later chapters in this book comment on the meaning and significance of these assumptions, note implications of their violation for the OLS estimator, discuss ways of determining whether or not they are violated, and suggest new estimators appropriate to situations in which one of these assumptions must be replaced by an alternative assumption. Before we move on to this, however, more must be said about the character of the OLS estimator in the context of the CLR model, because of the central role it plays in the econometrician's "catalog."

Table 3.1 The assumptions of the CLR model.

Assumption	*Mathematical expression*		*Violations*	*Chapter in which discussed*
	Bivariate	*Multivariate*		
1. Dependent variable a linear function of a specific set of independent variables, plus a disturbance	$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, $t = 1, \ldots, N$	$Y = X\beta + \varepsilon$	Wrong regressors Nonlinearity Changing parameters	6
2. Expected value of disturbance term is zero	$E\varepsilon_t = 0$, for all t	$E\varepsilon = 0$	Biased intercept	7
3. Disturbances have uniform variance and are uncorrelated	$E\varepsilon_t\varepsilon_r = 0$, $t \neq r$ $= \sigma^2$, $t = r$	$E\varepsilon\varepsilon' = \sigma^2 I$	Heteroskedasticity Autocorrelated errors	8
4. Observations on independent variables can be considered fixed in repeated samples	x_t fixed in repeated samples	X fixed in repeated samples	Errors in variables Autoregression Simultaneous equations	10 11
5. No exact linear relationships between independent variables and more observations than independent variables	$\sum_{t=1}^{N}(x_t - \bar{x})^2 \neq 0$	Rank of $X = K \leq N$	Perfect multicollinearity	12

The mathematical terminology is explained in the technical notes to this section. The notation is as follows: Y is a vector of observations on the dependent variable; X is a matrix of observations on the independent variables; ε is a vector of disturbances; σ^2 is the variance of the disturbances; I is the identity matrix; K is the number of independent variables; N is the number of observations.

3.3 The OLS Estimator in the CLR Model

The central role of the OLS estimator in the econometrician's catalog is that of a standard against which all other estimators are compared. The reason for this is that the OLS estimator is extraordinarily popular. This popularity stems from the fact that, in the context of the CLR model, the OLS estimator has a large number of desirable properties, making it the overwhelming choice for the "optimal" estimator when the estimating problem is accurately characterized by the CLR model. This is best illustrated by looking at the eight criteria listed in chapter 2 and determining how the OLS estimator rates on these criteria in the context of the CLR model.

1. *Computational cost*. All econometric software packages estimate OLS in a flash, and many popular nonstatistical software packages, such as Excel, do so as well.
2. *Least squares*. Because the OLS estimator is designed to minimize the sum of squared residuals, it is automatically "optimal" on this criterion.
3. *Highest R^2*. Because the OLS estimator is optimal on the least squares criterion, it will automatically be optimal on the highest R^2 criterion.
4. *Unbiasedness*. The assumptions of the CLR model can be used to show that the OLS estimator β^{OLS} is an unbiased estimator of β.
5. *Best unbiasedness*. In the CLR model β^{OLS} is a linear estimator; that is, it can be written as a linear function of the errors. As noted earlier, it is unbiased. Among all linear unbiased estimators of β, it can be shown (in the context of the CLR model) to have the "smallest" variance–covariance matrix. Thus the OLS estimator is the best linear unbiased estimator (BLUE) in the CLR model. If we add the additional assumption that the disturbances are distributed normally (creating the *classical normal linear regression model* [CNLR model]), it can be shown that the OLS estimator is the best unbiased estimator (i.e., best among *all* unbiased estimators, not just linear unbiased estimators).
6. *Mean square error*. It is not the case that the OLS estimator is the minimum mean square error estimator in the CLR model. Even among linear estimators, it is possible that a substantial reduction in variance can be obtained by adopting a slightly biased estimator. This is the OLS estimator's weakest point; chapters 12 and 13 discuss several estimators whose appeal lies in the possibility that they may beat OLS on the mean square error (MSE) criterion.
7. *Asymptotic criteria*. Because the OLS estimator in the CLR model is unbiased, it is also unbiased in samples of infinite size and thus is asymptotically unbiased. It can also be shown that the variance–covariance matrix of β^{OLS} goes to zero as the sample size goes to infinity, so that β^{OLS} is also a consistent estimator of β. Further, in the CNLR model it is asymptotically efficient.
8. *Maximum likelihood*. It is impossible to calculate the maximum likelihood estimator given the assumptions of the CLR model, because these assumptions do not specify the functional form of the distribution of the disturbance terms. However, if the disturbances are assumed to be distributed normally (the CNLR model), then it turns out that β^{MLE} is identical to β^{OLS}.

Thus, whenever the estimating situation can be characterized by the CLR model, the OLS estimator meets practically all of the criteria econometricians consider relevant. It is no wonder, then, that this estimator has become so popular. It is in fact *too* popular: it is often used, without justification, in estimating situations that are not accurately represented by the CLR model. If some of the CLR model assumptions do not hold, many of the desirable properties of the OLS estimator may no longer hold. If the OLS estimator does not have the properties that are thought to be of most importance, an alternative estimator must be found. Before moving to this aspect of our examination of econometrics, however, we will discuss in the next chapter some concepts of and problems in inference, to provide a foundation for later chapters.

General Notes

3.1 Textbooks as Catalogs

- The econometricians' catalog is not viewed favorably by all. Consider the opinion of Worswick (1972, p. 79): "[Econometricians] are not, it seems to me, engaged in forging tools to arrange and measure actual facts so much as making a marvelous array of pretend-tools which would perform wonders if ever a set of facts should turn up in the right form."
- Bibby and Toutenburg (1977, pp. 72–3) note that the CLR model, what they call the general linear model (GLM), can be a trap, a snare, and a delusion. They quote Whitehead as saying: "Seek simplicity ... and distrust it," and go on to explain how use of the linear model can change in undesirable ways the nature of the debate on the phenomenon being examined in the study in question. For example, casting the problem in the mold of the CLR model narrows the question by restricting its terms of reference to a particular model based on a particular set of data; it trivializes the question by focusing attention on apparently meaningful yet potentially trivial questions concerning the values of unknown regression coefficients; and it "technicalizes" the debate, obscuring the real questions at hand, by turning attention to technical statistical matters capable of being understood only by experts.

  They warn users of the GLM by noting that, "it certainly eliminates the complexities of hardheaded thought, especially since so many computer programs exist. For the soft-headed analyst who doesn't want to think too much, an off-the-peg computer package is simplicity itself, especially if it cuts through a mass of complicated data and provides a few easily reportable coefficients. Occam's razor has been used to justify worse barbarities: but razors are dangerous things and should be used carefully."
- If more than one of the CLR model assumptions is violated at the same time, econometricians often find themselves in trouble because their catalogs usually tell them what to do if only *one* of the CLR model assumptions is violated. Much recent econometric research examines situations in which two assumptions of the CLR model are violated simultaneously. These situations will be discussed when appropriate.

3.3 The OLS Estimator in the CLR Model

- The process whereby the OLS estimator is applied to the data at hand is usually referred to by the terminology "running a regression." The dependent variable (the "regressand") is said to be "regressed" on the independent variables ("the regressors") to produce the OLS estimates. This terminology comes from a pioneering empirical study in which it was found that the mean height of children born of parents of a given height tends to "regress" or move towards the population average height. See Maddala (1977, pp. 97–101) for further comment on this and for discussion of the meaning and interpretation of regression

analysis. Regression analysis is the heart and soul of econometrics, as noted by Fiedler (1977, p. 63): "Most economists think of God as working great multiple regressions in the sky." Critics note that the *New Standard Dictionary* defines regression as "The diversion of psychic energy ... into channels of fantasy."

- The result that the OLS estimator in the CLR model is the BLUE is often referred to as the Gauss–Markov theorem.
- The formula for the OLS estimator of a specific element of the β vector usually involves observations on *all* the independent variables (as well as observations on the dependent variable), not just observations on the independent variable corresponding to that particular element of β. This is because, to obtain an accurate estimate of the influence of one independent variable on the dependent variable, the simultaneous influence of other independent variables on the dependent variable must be taken into account. Doing this ensures that the jth element of β^{OLS} reflects the influence of the jth independent variable on the dependent variable, holding all the other independent variables constant. Similarly, the formula for the variance of an element of β^{OLS} also usually involves observations on all the independent variables.
- Because the OLS estimator is so popular, and because it so often plays a role in the formulation of alternative estimators, it is important that its mechanical properties be well understood. The most effective way of expositing these characteristics is through the use of a Venn diagram called the Ballentine. Suppose the CLR model applies, with Y determined by X and an error term. In Figure 3.1 the circle Y represents variation in the dependent variable Y and the circle X represents variation in the independent variable X. The overlap of X with Y, the blue area, represents variation that Y and X have in common in the sense that this variation in Y can be explained by X via an OLS regression. The blue area reflects information employed by the estimating procedure in estimating the slope coefficient β_x; the larger this area, the more information is used to form the estimate and thus the smaller is its variance.

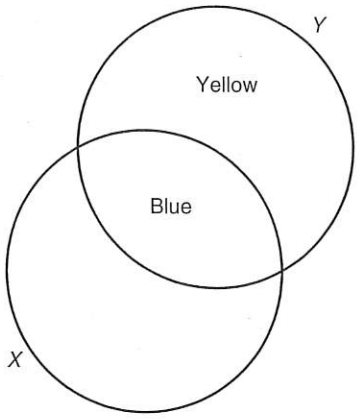

Figure 3.1 Defining the Ballentine Venn diagram.

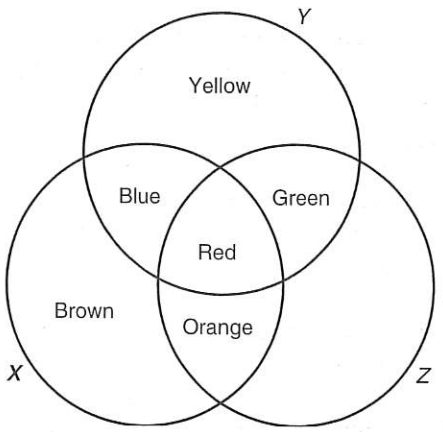

Figure 3.2 Interpreting multiple regression with the Ballentine.

Now consider Figure 3.2, in which a Ballentine for a case of two explanatory variables, X and Z, is portrayed (i.e., now Y is determined by both X and Z). In general, the X and Z circles will overlap, reflecting some collinearity between the two; this is shown in Figure 3.2 by the red-plus-orange area. If Y were regressed on X alone, information in the blue-plus-red area would be used to estimate β_x, and if Y were regressed on Z alone, information in the green-plus-red area would be used to estimate β_z. What happens, though, if Y is regressed on X and Z together?

In the multiple regression of Y on X and Z together, the OLS estimator uses the information in the blue area to estimate β_x and the information in the green area to estimate β_z, *discarding the information in the red area*. The information in the blue area corresponds to variation in Y that matches up uniquely with variation in X; using this information should therefore produce an unbiased estimate of β_x. Similarly, information in the green area corresponds to variation in Y that matches up uniquely with variation in Z; using this information should produce an unbiased estimate of β_z. The information in the red area is not used because it reflects variation in Y that is determined by variation in *both* X and Z, the relative contributions of which are not *a priori* known. In the blue area, for example, variation in Y is all due to variation in X, so matching up this variation in Y with variation in X should allow accurate estimation of β_x. But in the red area, matching up these variations will be misleading because not all variation in Y is due to variation in X.

- Notice that regression Y on X and Z together creates unbiased estimates of β_x and β_z, whereas regressing Y on X and Z separately creates biased estimates of β_x and β_z because this latter method uses the red area. But notice also that, because the former method discards the red area, it uses less information to produce its slope coefficient estimates and thus these estimates will have larger variances. As is invariably the case in econometrics, the price of obtaining unbiased estimates is higher variances.
- Whenever X and Z are orthogonal to one another (have zero collinearity) they do not overlap as in Figure 3.2 and the red area disappears. Because there is no red area in this case, regressing Y on X alone or on Z alone produces the same estimates of β_x and β_z as if Y were regressed on X and Z together. Thus, although in general the OLS estimate of a specific element of the β vector involves observations on *all* the regressors, in the case of orthogonal regressors it involves observations on only one regressor (the one for which it is the slope coefficient estimate).
- Whenever X and Z are highly collinear and therefore overlap a lot, the blue and green areas become very small, implying that when Y is regressed on X and Z together very little information is used to estimate β_x and β_z. This causes the variances of these estimates to be very large. Thus, the impact of multicollinearity is to raise the variances of the OLS estimates. Perfect collinearity causes the X and Z circles to overlap completely; the blue and green areas disappear and estimation is impossible. Multicollinearity is discussed at length in chapter 12.
- In Figure 3.1 the blue area represents the variation in Y explained by X. Thus, R^2 is given as the ratio of the blue area to the entire Y circle. In Figure 3.2 the blue-plus-red-plus-green area represents the variation in Y explained by X and Z together. (Note that the red area is discarded only for the purpose of estimating the coefficients, not for predicting Y; once the coefficients are estimated, all variation in X and Z is used to predict Y.) Thus, the R^2 resulting from the multiple regression is given by the ratio of the blue-plus-red-plus-green area to the entire Y circle. Notice that there is no way of allocating portions of the total R^2 to X and Z because the red area variation is explained by *both*, in a way that cannot be disentangled. Only if X and Z are orthogonal, and the red area disappears, can the total R^2 be allocated unequivocally to X and Z separately.
- The yellow area represents variation in Y attributable to the error term, and thus the magnitude of the yellow area represents the magnitude of σ^2, the variance of the error term. This implies, for example, that if, in the context of Figure 3.2, Y had been regressed on only X, omitting Z, σ^2 would be estimated by the yellow-plus-green area, an overestimate.
- The Ballentine was named by its originators Cohen and Cohen (1975) after a brand of US beer whose logo resembles Figure 3.2. Their use of the Ballentine was confined to the exposition of various concepts related to R^2. Kennedy (1981b) extended its use to the exposition of other aspects of regression. It turns out that the Ballentine can mislead on occasion, particularly when used to exposit R^2 concepts. A limitation of the Ballentine is that it is necessary in certain cases for the red area to represent a negative quantity.

(Suppose the two explanatory variables X and Z each have positive coefficients, but in the data X and Z are negatively correlated: X alone could do a poor job of explaining variation in Y because, for example, the impact of a high value of X is offset by a low value of Z. The red area would have to be negative!) This problem notwithstanding, the use of the Ballentine to exposit bias and variance magnitudes for regression is retained in this book, on the grounds that the benefits of its illustrative power outweigh the danger that it will lead to error. The Ballentine is used here as a metaphoric device illustrating some regression results; it should not be given meaning beyond that.

- An alternative geometric analysis of OLS, using vector geometry, is sometimes used. Davidson and MacKinnon (1993, chapter 1) have a good exposition.

Technical Notes

3.2 The Five Assumptions

- The regression model $y = g(x_1, \ldots, x_k) + \varepsilon$ is really a specification of how the conditional means $E(y \mid x_1, \ldots, x_k)$ are related to each other through x. The population regression function is written as $E(y \mid x_1, \ldots, x_k) = g(x)$; it describes how the average or expected value of y varies with x. Suppose g is a linear function so that the regression function is $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$. Each element of β^{OLS} (β_4^{OLS}, for example) is an estimate of the effect on the conditional expectation of y of a unit change in x_4, with all other x held constant.
- The fourth assumption of the CLR model is that the observations on the explanatory variables can be considered fixed in repeated samples; that is, it is possible to redraw the sample with the same explanatory variable values. This is often weakened to read that the explanatory variables are random but independent of the error term. The examples of violations of this assumption given earlier (errors in variables, autoregression, and simultaneous equations) were all instances in which the explanatory variables were random and *not* independent of the error term.

In many instances the explanatory variables are such that they can be considered fixed in repeated samples, for example, when there is one observation on each of the 50 states so that the sample exhausts the population. But in many instances the observations do not exhaust the population. A sample of a thousand individuals from the Current Population Survey (CPS) is an example. In this latter instance we could ask how would the parameter estimates vary when we draw a set of observations on a new set of a thousand individuals along with a new set of error terms: the nature of the conceptual repeated sample is different! There is no reason to believe that a new draw of a thousand observations from the CPS is related to a new draw of error terms, so the weaker version of the fourth assumption is satisfied. Consequently, the OLS estimator continues to be BLUE (although one might complain that in a sense it is no longer linear). It is straightforward to show that it is unbiased, but a difficulty arises when finding the formula for its variance–covariance matrix. The usual formula is $\sigma^2(X'X)^{-1}$ but when X is stochastic rather than fixed this formula becomes $\sigma^2 E[(X'X)^{-1}]$. The difficulty occurs because $E[(X'X)^{-1}]$ is the expected value of a complicated nonlinear function of a stochastic variable. As seen in the technical notes to section 2.8, the expected value of a nonlinear function is not equal to the nonlinear function of the expected value. Because of this $(X'X)^{-1}$ is a biased estimate of $E[(X'X)^{-1}]$. Econometricians wishing to avoid assuming that the explanatory variables are fixed in repeated samples use two means of dealing with this problem, neither of which is fully satisfactory. First, they may talk in terms of $\sigma^2(X'X)^{-1}$ being the variance of OLS *conditional* on X and so use this traditional formula. But this is just another way of saying that we are holding X constant in repeated samples! Second, they may revert to asymptotic criteria so that although biased, $\sigma^2(X'X)^{-1}$ is a consistent estimate of $\sigma^2 E[(X'X)^{-1}]$, and so continue to use this traditional formula. This is a bit questionable

because in small samples it means that estimation of the variance is biased downward because it does not account for variability coming from the change in explanatory variable observations over repeated samples. Stock and Watson (2007) is a textbook adopting the weaker version of assumption 4, employing the asymptotic approach. They argue that the asymptotic approach is necessary in any event because it is unlikely that errors are distributed normally. (In large samples, the OLS estimator is distributed normally, regardless of how the errors are distributed.)

- In the CLR model, the regression model is specified as $y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k +$ disturbance, a formula that can be written down N times, once for each set of observations on the dependent and independent variables. This gives a large stack of equations, which can be consolidated via matrix notation as $Y = X\beta + \varepsilon$. Here Y is a vector containing the N observations on the dependent variable y; X is a matrix consisting of K columns, each column being a vector of N observations on one of the independent variables; and ε is a vector containing the N unknown disturbances.

3.3 The OLS Estimator in the CLR Model

- The formula for β^{OLS} is $(X'X)^{-1}X'Y$. A proper derivation of this is accomplished by minimizing the sum of squared errors. An easy way of remembering this formula is to premultiply $Y = X\beta + \varepsilon$ by X' to get $X'Y = X'X\beta + X'\varepsilon$, drop the $X'\varepsilon$, and then solve for β.
- The formula for the variance–covariance matrix β^{OLS} is $\sigma^2(X'X)^{-1}$ where σ^2 is the variance of the disturbance term. For the simple case in which the regression function is $y = \beta_1 + \beta_2 x$ this gives the formula $\sigma^2/\Sigma(x - \bar{x})^2$ for the variance of β_2^{OLS}. Note that, if the variation in the regressor values is substantial, the denominator of this expression will be large, tending to make the variance of β^{OLS} small.
- The variance–covariance matrix of β^{OLS} is usually unknown because σ^2 is usually unknown. It is estimated by $s^2(X'X)^{-1}$ where s^2 is an estimator of σ^2. The estimator s^2 is usually given by the formula $\hat{\varepsilon}'\hat{\varepsilon}/(N - K) = \Sigma\hat{\varepsilon}_i^2/(N - K)$ where $\hat{\varepsilon}$ is the estimate of the disturbance vector, calculated as $(Y - \hat{Y})$ where $\hat{Y}$ is $X\beta^{OLS}$. In the CLR model s^2 is the best quadratic unbiased estimator of σ^2; in the CNLR model it is best unbiased.
- By discarding the red area in Figure 3.2, the OLS formula ensures that its estimates of the influence of one independent variable are calculated while controlling for the simultaneous influence of the other independent variables, that is, the interpretation of, say, the jth element of β^{OLS} is as an estimate of the influence of the jth explanatory variable, holding all other explanatory variables constant. That the red area is discarded can be emphasized by noting that the OLS estimate of, say, β_x can be calculated from either the regression of Y on X and Z together or the regression of Y on X "residualized" with respect to Z (i.e., with the influence of Z removed). In Figure 3.2, if we were to regress X on Z we would be able to explain the red-plus-orange area; the residuals from this regression, the blue-plus-brown area, are called X residualized for Z. Now suppose that Y is regressed on X residualized for Z. The overlap of the Y circle with the blue-plus-brown area is the blue area, so exactly the same information is used to estimate β_x in this method as is used when Y is regressed on X and Z together, resulting in an identical estimate of β_x.

  Notice further that, if Y were also residualized for Z, producing the yellow-plus-blue area, regressing the residualized Y on the residualized X would also produce the same estimate of β_x since their overlap is the blue area. An important implication of this result is that, for example, running a regression on data from which a linear time trend has been removed will produce exactly the same coefficient estimates as when a linear time trend is included among the regressors in a regression run on raw data. As another example, consider the removal of a linear seasonal influence; running a regression on linearly deseasonalized data will produce exactly the same coefficient estimates as if the linear seasonal influence were included as an extra regressor in a regression run on raw data.

- A variant of OLS called *stepwise regression* is to be avoided. It consists of regressing Y on each explanatory variable separately and keeping the regression with the highest R^2. (A variant looks for the regressor with the highest t statistic.) This determines the estimate of the slope coefficient of that regression's explanatory variable. Then the residuals from this regression are used as the dependent variable in a new search using the remaining explanatory variables and the procedure is repeated. Suppose that, for the example of Figure 3.2, the regression of Y on X produced a higher R^2 than the regression of Y on Z. Then the estimate of β_x would be formed using the information in the blue-plus-red area. Note that this estimate is biased. Econometricians often denigrate statisticians on the grounds that they espouse such algorithmic searches. Leamer (2007, p.101) expresses this cogently:

  > We don't rely on stepwise regression or any other automated statistical pattern recognition to pull understanding from our data sets because there is currently no way of providing the critical contextual inputs into these algorithms and because an understanding of the context is absolutely critical to making sense of our noisy non-experimental data. The last person you want to analyze an economic data set is a statistician, which is what you get when you run stepwise regression.

- The Ballentine can be used to illustrate several variants of R^2. Consider, for example, the simple R^2 between Y and Z in Figure 3.2. If the area of the Y circle is normalized to be unity, this simple R^2, denoted R^2_{yz}, is given by the red-plus-green area. The *partial* R^2 between Y and Z is defined as reflecting the influence of Z on Y *after* accounting for the influence of X. It is measured by obtaining the R^2 from the regression of Y corrected for X on Z corrected for X, and is denoted $R^2_{yz\cdot x}$. Our earlier use of the Ballentine makes it easy to deduce that in Figure 3.2 it is given as the green area divided by the yellow-plus-green area. The reader might like to verify that it is given by the formula

$$R^2_{yz\cdot x} = (R^2 - R^2_{yx})/(1 - R^2_{yx}).$$

- The OLS estimator has several well-known mechanical properties with which students should become intimately familiar – instructors tend to assume this knowledge after the first lecture or two on OLS. Listed below are the more important of these properties; proofs can be found in most textbooks. The context is $y = \alpha + \beta x + \varepsilon$.
  1. If $\beta = 0$ so that the only regressor is the intercept, y is regressed on a column of ones, producing $\alpha^{OLS} = \bar{y}$, the average of the y observations.
  2. If $\alpha = 0$ so there is no intercept and one explanatory variable, y is regressed on a column of x values, producing $\beta^{OLS} = \Sigma xy/\Sigma x^2$.
  3. If there is an intercept and one explanatory variable $\bar{x}$

$$\begin{aligned}\beta^{OLS} &= \Sigma(x-\bar{x})(y-\bar{y})/\Sigma(x-\bar{x})^2 \\ &= \Sigma(x-\bar{x})y/\Sigma(x-\bar{x})^2.\end{aligned}$$

  4. If observations are expressed as deviations from their means, $y^* = y - \bar{y}$ and $x^* = x - \bar{x}$, then $\beta^{OLS} = \Sigma x^* y^*/\Sigma x^{*2}$. This follows from (3) above. Lower case letters are sometimes reserved to denote deviations from sample means.
  5. The intercept can be estimated as $\bar{y} - \beta^{OLS}\bar{x}$ or, if there are more explanatory variables, as $\bar{y} - \Sigma\beta_i^{OLS}\bar{x}_i$. This comes from the first normal equation, the equation that results from setting the partial derivative of SSE (the sum of squared errors) with respect to α equal to zero (to minimize the SSE).
  6. An implication of (5) is that the sum of the OLS residuals equals zero; in effect the intercept is estimated by the value that causes the sum of the OLS residuals to equal zero.
  7. The predicted, or estimated, y values are calculated as $\hat{y}_i = \alpha^{OLS} + \beta^{OLS}x_i$. An implication of (6) is that the mean of the $\hat{y}$ values equals the mean of the actual y values: $\bar{\hat{y}} = \bar{y}$.
  8. An implication of (5), (6), and (7) above is that the OLS regression line passes through the overall mean of the data points.

9. Adding a constant to a variable, or scaling a variable, has a predictable impact on the OLS estimates. For example, multiplying the x observations by 10 will multiply β^{OLS} by one-tenth, and adding 6 to the y observations will increase α^{OLS} by 6.
10. A linear restriction on the parameters can be incorporated into a regression by eliminating one coefficient from that equation and running the resulting regression using transformed variables. For an example see the general notes to section 4.3.
11. The "variation" in the dependent variable is the "total sum of squares" $SST = \Sigma(y - \bar{y})^2 = y'y - N\bar{y}^2$ where $y'y$ is matrix notation for Σy^2, and N is the sample size.
12. The "variation" explained linearly by the independent variables is the "regression sum of squares," $SSR = \Sigma(\hat{y} - \bar{\hat{y}})^2 = \hat{y}'\hat{y} - N\bar{y}^2$.
13 The sum of squared errors from a regression is $SSE = (y - \hat{y})'(y - \hat{y}) = y'y - \hat{y}'\hat{y} = SST - SSR$. (Note that textbook notation varies. Some authors use SSE for "explained sum of squares" and *SSR* for "sum of squared residuals," creating results that look to be the opposite of those given here.)
14 SSE is often calculated by $\Sigma y^2 - \alpha^{OLS}\Sigma y - \beta^{OLS}\Sigma xy$, or in the more general matrix notation $y'y - \beta^{OLS'}X'y$.
15 The coefficient of determination, $R^2 = SSR/SST = 1 - SSE/SST$ is maximized by OLS because OLS minimizes SSE. R^2 is the squared correlation coefficient between y and $\hat{y}$; it is the fraction of the "variation" in y that is explained linearly by the explanatory variables.
16 When no intercept is included, it is possible for R^2 to lie outside the zero to one range. See the general notes to section 2.4.
17 Minimizing with some extra help cannot make the minimization less successful. Thus SSE decreases (or in unusual cases remains unchanged) when an additional explanatory variable is added; R^2 must therefore rise (or remain unchanged).
18 Because the explanatory variable(s) is (are) given as much credit as possible for explaining changes in y, and the error as little credit as possible, ε^{OLS} is uncorrelated with the explanatory variable(s) and thus with $\hat{y}$ (because $\hat{y}$ is a linear function of the explanatory variable(s)).
19 The estimated coefficient of the ith regressor can be obtained by regressing y on this regressor "residualized" for the other regressors (the residuals from a regression of the ith regressor on all the other regressors). The same result is obtained if the "residualized" y is used as the regressand, instead of y. These results were explained earlier in these technical notes with the help of the Ballentine.