GLS estimates, with the weights given by the corresponding area under the posterior distribution of r. The greater is the number of subsets, the more closely will the numerical integration approximate the "true" Bayesian estimator. How many is enough? Kennedy and Simons (1991) suggest for this example that only 40 are required for this estimator to perform well. For a textbook exposition of these formulas, see Judge et al. (1985, pp. 291-3).

14
Dummy Variables

14.1 Introduction

Explanatory variables are often qualitative in nature (e.g., wartime versus peace-time, male versus female, east versus west versus south), so that some proxy must be constructed to represent them in a regression. Dummy variables are used for this purpose. A dummy variable is an artificial variable constructed such that it takes the value unity whenever the qualitative phenomenon it represents occurs, and zero otherwise. Once created, these proxies, or "dummies" as they are called, are used in the CLR model just like any other explanatory variable, yielding standard OLS results.

The exposition below is in terms of an example designed to illustrate the roles dummy variables can play, give insight to how their coefficients are estimated in a regression, and clarify the interpretation of these coefficient estimates.

Consider data on the incomes of doctors, professors and lawyers, exhibited in figure 14.1 (where the data have been ordered so as to group observations into the professions), and suppose it is postulated that an individual's income depends on his or her profession, a qualitative variable. We may write this model as

$$Y = \alpha_D D_D + \alpha_P D_P + \alpha_L D_L + \varepsilon \qquad (1)$$

where *D*D is a dummy variable taking the value one whenever the observation in question is a doctor, and zero otherwise; *D*P and *D*L are dummy variables defined in like fashion for professors and lawyers. Notice that the equation in essence states that an individual's income is given by the coefficient of his or her related dummy variable plus an error term. (For a professor, for example, *D*D and *D*L are zero and *D*P is one, so (1) becomes $Y = aP + e$.)

From the structure of equation (1) and the configuration of figure 14.1, the logical estimate of aD is the average of all doctors' incomes, of aP the average of all professors' incomes, and of aL the average of all lawyers' incomes. It is reassuring, then, that if *Y* is regressed on these three dummy variables, these are exactly the estimates that result.

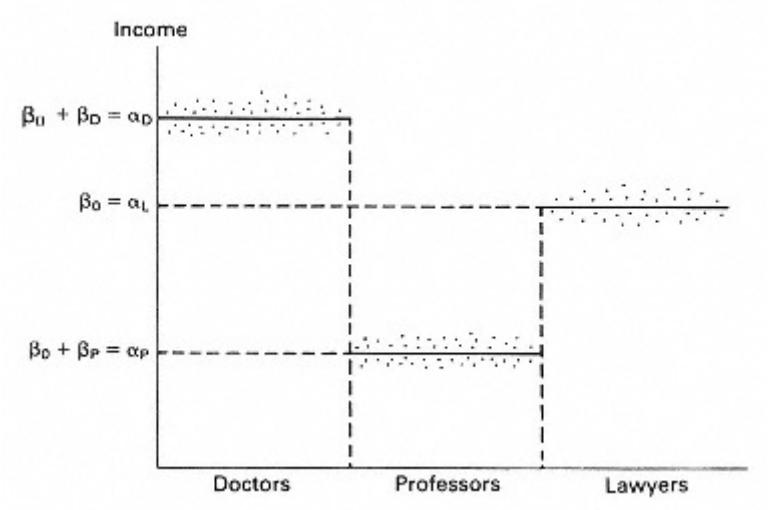

Figure 14.1
A step function example of using dummy variables

14.2 Interpretation

Equation (1) as structured does not contain an intercept. If it did, perfect multicollinearity would result (the intercept variable, a column

of ones, would equal the sum of the three dummy variables) and the regression could not be run. Nonetheless, more often than not, equations with dummy variables do contain an intercept. This is accomplished by omitting one of the dummies to avoid perfect multicollinearity.

Suppose *D*L is dropped, for example, creating

$$Y = \beta_0 + \beta_D D_D + \beta_P D_P + \varepsilon. \qquad (2)$$

In this case, for a lawyer *D*D and *D*P are zero, so a lawyer's expected income is given by the intercept b0. Thus the logical estimate of the intercept is the average of all lawyers' incomes. A doctor's expected income is given by equation (2) as b0 + bD; thus the logical estimate of bD is the difference between the doctors' average income and the lawyers' average income. Similarly, the logical estimate of bP is the difference between the professors' average income and the lawyers' average income. Once again, it is reassuring that, when regression (2) is undertaken (i.e., regressing *Y* on an intercept and the dummy variables *D*D and *D*P), exactly these results are obtained. The crucial difference is that with an intercept included the interpretation of the dummy variable coefficients changes dramatically.

With no intercept, the dummy variable coefficients reflect the expected

income for the respective professions. With an intercept included, the omitted category (profession) becomes a base or benchmark to which the others are compared. The dummy variable coefficients for the remaining categories measure the extent to which they differ from this base. This base in the example above is the lawyer profession. Thus the coefficient BD, for example, gives the *difference* between the expected income of a doctor and the expected income of a lawyer.

Most researchers find the equation with an intercept more convenient because it allows them to address more easily the questions in which they usually have the most interest, namely whether or not the categorization makes a difference and if so by how much. If the categorization does make a difference, by how much is measured

directly by the dummy variable coefficient estimates. Testing whether or not the categorization is relevant can be done by running a *t* test of a dummy variable coefficient against zero (or, to be more general, an *F* test on the appropriate set of dummy variable coefficient estimates).

14.3 Adding Another Qualitative Variable

Suppose now the data in figure 14.1 are rearranged slightly to form figure 14.2, from which it appears that gender may have a role to play in determining income. This issue is usually broached in one of two ways. The most common way is to include in equations (1) and (2) a new dummy variable *D*F for gender to create

$$Y = \alpha_D^* D_D + \alpha_P^* D_P + \alpha_L^* D_L + \alpha_F^* D_F + \varepsilon \qquad (1^*)$$

$$Y = \beta_0^* + \beta_D^* D_D + \beta_P^* D_P + \beta_F^* D_F + \varepsilon \qquad (2^*)$$

where *D*F takes the value 1 for a female and 0 for a male. Notice that no dummy variable *D*M representing males is added; if such a dummy were added perfect multicollinearity would result, in equation (1*) because *D*D + *D*P + *D*L = *D*F + *D*M and in equation (2*) because *D*F + *D*M is a column of ones, identical to the implicit intercept variable. The interpretation of both α_F^* and β_F^* is as the extent to which being female changes income, regardless of profession. α_D^*, α_P^* and α_D^* are interpreted as expected income of a male in the relevant profession; a similar reinterpretation is required for the coefficients of equation (2*).

The second way of broaching this issue is to scrap the old dummy variables and create new dummy variables, one for each category illustrated in figure 14.2. This produces

$$Y = \alpha_{FD} D_{FD} + \alpha_{MD} D_{MD} + \alpha_{FP} D_{FP} + \alpha_{MP} D_{MP} + \alpha_{FL} D_{FL} + \alpha_{ML} D_{ML} + \varepsilon \qquad (1')$$

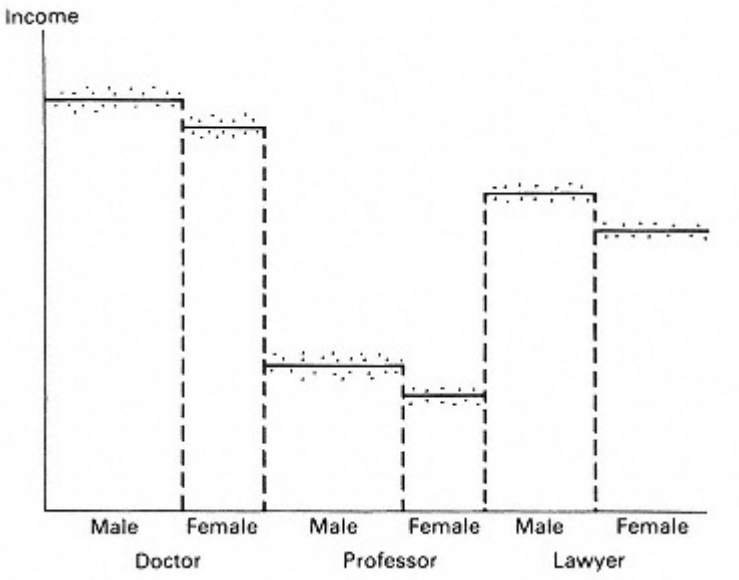

Figure 14.2
Adding gender as an additional dummy variable

and

$$Y = \beta_0^* + \beta_{FD} D_{FD} + \beta_{MD} D_{MD} + \beta_{FP} D_{FP} + \beta_{MP} D_{MP} + \beta_{FL} D_{FL} + \varepsilon. \qquad (2')$$

The interpretation of the coefficients is straightforward: aFD, for example, is the expected income of a female doctor, and bFD is the extent to which the expected income of a female doctor differs from that of a male lawyer.

The key difference between these two methods is that the former method forces the difference in income between male and female to be the same for all professions whereas the latter does not. The latter method allows for what are called interaction effects. In the former method a female doctor's expected income is the sum of two parts, one attributable to being a doctor and the other attributable to being a female; there is no role for any special effect that the combination or interaction of doctor and female might have.

14.4 Interacting with Quantitative Variables

All the foregoing examples are somewhat unrealistic in that they are regressions in which all the regressors are dummy variables. In general, however, quantitative variables determine the dependent variable as well as qualitative variables. For example, income in an earlier example may also be determined by years of experience, *E,* so that we might have

$$Y = \gamma_0 + \gamma_D D_D + \gamma_P D_P + \gamma_E E + \varepsilon. \qquad (3)$$

In this case the coefficient gD must be interpreted as reflecting the difference between doctors' and lawyers' expected incomes, taking account of years of experience (i.e., assuming equal years of experience).

Equation (3) is in essence a model in which income is expressed as a linear function of experience, with a different intercept for each profession. (On a graph of income against experience, this would be reflected by three parallel lines, one for each profession.) The most common use of dummy variables is to effect an intercept shift of this nature. But in many contexts it may be that the slope coefficient gE could differ for different professions, either in addition to or in place of a different intercept. (This is also viewed as an interaction effect.)

This case is handled by adding special dummies to account for slope differences. Equation (3) becomes

$$Y = \gamma_0^* + \gamma_D^* D_D + \gamma_P^* D_P + \gamma_E^* E + \gamma_{ED}^* (D_D E) + \gamma_{EP}^* (D_P E) + \varepsilon. \qquad (4)$$

Here *(DDE* is a variable formed as the "product" of *D*D and *E;* it consists of the value of *E* for each observation on a doctor, and 0 elsewhere. The special "product" dummy *(DPE)* is formed in similar fashion. The expression (4) for observations on a lawyer is $\gamma_0^* + \gamma_E^* E + \varepsilon$, so γ_0^* and γ_E^* are the intercept and slope coefficients relevant to lawyers. The expression (4) for observations on a doctor is $\gamma_0^* + \gamma_D^* + (\gamma_E^* + \gamma_{ED}^*)E + \varepsilon$, so the interpretation of γ_D^* is as the difference between the doctors' and the lawyers' intercepts and the

interpretation of γ_{ED}^{*} is as the difference between the doctors' and the lawyers' slope coefficients. Thus this special "product" dummy variable can allow for changes in slope coefficients from one data set to another and thereby capture a different kind of interaction effect.

Equation (4) is such that each profession has its own intercept and its own slope. (On a graph of income against experience, the three lines, one for each profession, need not be parallel.) Because of this there will be no difference between the estimates resulting from running this regression and the estimates resulting from running three separate regressions, each using just the data for a particular profession. Thus in this case using dummy variables is of no value. The dummy variable technique is of value whenever restrictions of some kind are imposed on

the model in question. Equation (3) reflects such a restriction; the slope coefficient gE is postulated to be the same for all professions. By running equation (3) as a single regression, this restriction is imposed and more efficient estimates of all parameters result. As another example, suppose that years of education were also an explanatory variable but that it is known to have the same slope coefficient in each profession. Then adding the extra explanatory variable years of education to equation (4) and performing a single regression produces more efficient estimates of all parameters than would be the case if three separate regressions were run. (It should be noted that running a single, constrained regression incorporates the additional assumption of a common error variance.)

14.5 Observation-Specific Dummies

An observation-specific dummy is a dummy variable that takes on the value one for a specific observation and zero for all other observations. Since its use is mainly in time series data, it is called a period-specific dummy in the discussion below. When a regression is run with a period-specific dummy the computer can ignore the specific observation - the OLS estimates can be calculated using all the other observations and then the coefficient for the period-specific dummy is estimated as the value that makes that period's error equal to zero. In this way *SSE* is minimized. This has several useful implications:

(1) The coefficient estimate for the period-specific dummy is the forecast error for that period, and the estimated variance of this coefficient estimate is the estimate of the variance of the forecast error, an estimate that is otherwise quite awkward to calculate - see chapter 18.

(2) If the value of the dependent variable for the period in question is coded as zero instead of its actual value (which may not be known, if we are trying to forecast it) then the estimated coefficient of the period-specific dummy is the forecast of that period's dependent variable.

(3) By testing the estimated coefficient of the period-specific dummy against zero, using a *t* test, we can test whether or not that observation is "consistent" with the estimated relationship. An *F* test would be used to test if several observations could be considered consistent with the estimated equation. In this case each observation would have its own period-specific dummy. Such tests are sometimes called post-sample predictive tests. This is described in the technical notes as a variant of the Chow test. The "rainbow" test (general notes, section 6.3) is also a variant of this approach, as are some tests for outliers.

14.6 Fixed and Random Effects Models

Dummy variables are sometimes used in the context of panel, or longitudinal, data - observations on a cross-section of individuals or firms, say, over time. In

this context it is often assumed that the intercept varies across the N cross-sectional units and/or across the T time periods. In the general case $(N - 1) + (T - 1)$ dummies can be used for this, with computational short-cuts available to avoid having to run a regression with all these extra variables. This way of analyzing panel data is called the *fixed effects* model. The dummy variable coefficients reflect ignorance - they are inserted merely for the purpose of measuring shifts in the regression line arising from unknown variables. Some researchers feel that this type of ignorance should be treated in a fashion similar to the general ignorance represented by the error term, and have accordingly proposed the *random effects, variance components, or error components* model.

In the random effects model there is an overall intercept and an error term with two components: e*it* + *ui*. The e*it* is the traditional error term unique to each observation. The *ui* is an error term representing the extent to which the intercept of the *i*th cross-sectional unit differs from the overall intercept. (Sometimes a third error is included, representing the extent to which the *t*th time period's intercept differs from the overall intercept.) This composite error term is seen to have a particular type of nonsphericalness that can be estimated, allowing the use of EGLS for estimation. (EGLS is explained in chapter 8.)

Which of the fixed effects and the random effects models is better? This depends on the context of the data and for what the results are to be used. If the data exhaust the population (say observations on all firms producing automobiles), then the fixed effects approach, which produces results conditional on the units in the data set, is reasonable. If the data are a drawing of observations from a large population (say a thousand individuals in a city many times that size), and we wish to draw inferences regarding other members of that population, the fixed effects model is no longer reasonable; in this context, use of the random effects model has the advantage that it saves a lot of degrees of freedom.

The random effects model has a major drawback, however: it assumes that the random error associated with each cross-section unit is uncorrelated with the other regressors, something that is not likely to be the case. Suppose, for example, that wages are being regressed on schooling for a large set of individuals, and that a missing variable, ability, is thought to affect the intercept; since schooling and ability are likely to be correlated, modeling this as a random effect will create correlation between the error and the regressor schooling (whereas modeling it as a fixed effect will not). The result is bias in the coefficient estimates from the random effects model. This may explain why the slope estimates from the fixed and random effects models are often so different.

A Hausman test (discussed in chapters 9 and 10) for correlation between the error and the regressors can be used to check for whether the random effects model is appropriate. Under the null hypothesis of no correlation between the error and the regressors, the random effects model is applicable and its EGLS estimator is consistent and efficient. Under the alternative it is inconsistent. The OLS estimator of the fixed effects model is consistent under both the null and the alternative. Consequently, the difference between the variance-covariance

matrices of the OLS and EGLS estimators is the variance-covariance matrix of the difference between the two estimators, allowing calculation of a chi-square test statistic to test this difference against zero.

General Notes

14.1 Introduction

The terminology "dummy variable" has invited irreverent remarks. One of the best is due to Machlup (1974, p. 892): "Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex."

Care must be taken in evaluating models containing dummy variables designed to capture structural shifts or seasonal factors, since these dummies could play a major role in generating a high R2, hiding the fact that the independent variables have little explanatory power.

Dummy variables representing more than two categories could represent categories that have no natural order (as in dummies for red, green and blue), but could represent those with some inherent order (as in low, medium and high income level). The latter are referred to as ordinal dummies; see Terza (1987) for a suggestion of how estimation can take account of the ordinal character of such dummies.

Regressions using microeconomic data often include dummies representing aggregates, such as regional, industry or occupation dummies. Moulton (1990) notes that within these aggregates errors are likely to be correlated and that ignoring this leads to downward-biased standard errors.

For the semi-logarithmic functional form In $Y = a + bx + dD + e$, the coefficient b is interpreted as the percentage impact on Y per unit change in x, but the coefficient d cannot be interpreted as the percentage image on Y of a change in the dummy variable D from zero to one status. The correct expression for this percentage impact is ed - 1. See Halvorsen and Palmquist (1980) and Kennedy (1981a).

Dummy variable coefficients are interpreted as showing the extent to which behavior in one category deviates from some base (the "omitted" category). Whenever there exist more than two categories, the presentation of these results can be awkward, especially when laymen are involved; a more relevant, easily understood base might make the presentation of these results more effective. For example, suppose household energy consumption is determined by income and the region in which the household lives. Rather than, say, using the South as a base and comparing household energy consumption in the North East, North Central and West to consumption in the South, it may be more effective, as a means of presenting these results to laymen, to calculate dummy variable coefficients in such a way as to compare consumption in each region with the national average. A simple adjustment permits this. See Suits (1984) and Kennedy (1986).

Goodman and Dubin (1990) note that alternative specifications containing different dummy variable specifications may not be nested, implying that a non-nested testing procedure should be employed to analyze their relative merits.

14.4 Interacting with Quantitative Variables

Dummy variables play an important role in structuring Chow tests for testing if there has been a change in a parameter value from one data set to another. Suppose Y is a linear function of X and Z and the question at hand is whether the coefficients are the same in period 1 as in period 2. A dummy variable D is formed such that D takes the value zero for observations in period 1 and the value one for observations in period 2. "Product" dummy variables DX and DZ are also formed (i.e., DX takes the value X in period 2 and is 0 otherwise). Then the equation

$$Y = \beta_0 + \alpha_0 D + \beta_1 X + \alpha_1 (DX) + \beta_2 Z + \alpha_2 (DZ) + \varepsilon \qquad (1)$$

is formed.

Running regression (1) as is allows the intercept and slope coefficients to differ from period 1 to period 2. This produces *SSE* unrestricted. Running regression (1) forcing a0, a1, and a2 to be 0 forces the intercept and slope-coefficients to be identical in both periods. An *F*

test, structured in the usual way, can be used to test whether or not the vector with elements a0, a1, and a2 is equal to the zero vector. The resulting F statistic is

$$\frac{[SSE(\text{constrained}) - SSE(\text{unconstrained})]/K}{SSE(\text{unconstrained})/(T_1 + T_2 - 2K)}$$

where K is the number of parameters, $T1$ is the number of observations in the first period and $T2$ is the number of observations in the second period. If there were more than two periods and we wished to test for equality across all periods, this methodology can be generalized by adding extra dummies in the obvious way.

Whenever the entire set of parameters is being tested for equality between two data sets the SSE unconstrained can be obtained by summing the SSEs from the two separate regressions and the SSE constrained can be obtained from a single regression using all the data; the Chow test often appears in textbooks in this guise. In general, including dummy variables to allow the intercept and all slopes to differ between two data sets produces the same coefficient estimates as those obtained by running separate regressions, but estimated variances differ because the former method constrains the estimated variance to be the same in both equations.

The advantage of the dummy variable variant of the Chow test is that it can easily be modified to test subsets of the coefficients. Suppose, for example, that it is known that, in equation (1) above, b2 changed from period 1 to period 2 and that it is desired to test whether or not the other parameters (b0 and b1) changed. Running regression (1) as is gives the unrestricted *SSE* for the required *F* statistic, and running (1) without *D* and *DX* gives the restricted *SSE*. The required degrees of freedom are 2 for the numerator and *T* - 6 for the denominator, where *T* is the total number of observations.

Notice that a slightly different form of this test must be used if, instead of knowing (or assuming) that b2 had changed from period 1 to period 2, we knew (or assumed) that it had *not* changed. Then running regression (1) without *DZ* gives the unrestricted *SSE* and running regression (2) without *D, DX* and *DZ* gives the restricted *SSE*. The degrees of freedom are 2 for the numerator and *T* - 5 for the denominator.

Using dummies to capture a change in intercept or slope coefficients, as described above, allows the line being estimated to be discontinuous. (Try drawing a graph of the curve - at the point of change it "jumps.") Forcing continuity creates what is called a *piecewise linear model;* dummy variables can be used to force this continuity, as explained, for example, in Pindyck and Rubinfeld (1981, pp. 126-7). This model is a special case of a *spline function,* in which the linearity assumption is dropped. For an exposition see Suits et al. (1978). Poirier (1976) has an extended discussion of this technique and its applications in economics.

A popular use of dummy variables is for seasonal adjustment. Setting dummies up to represent the seasons and then including these variables along with the other regressors eliminates seasonal influences in so far as, in a linear model, these seasonal influences affect the intercept term (or, in a log-linear model, these seasonal influences can be captured as seasonal percentage impacts on the dependent variable). Should the slope coefficients be affected by seasonal factors, a more extensive de-seasonalizing procedure would be required, employing "product" dummy variables. Johnston (1984, pp. 234-9) has a good discussion of using dummies to de-seasonalize. It must be noted that much more elaborate methods of de-seasonalizing data exist. For a survey see Pierce (1980). See also Raveh (1984) and Bell and Hillmer (1984). Robb (1980) and Gersovitz and MacKinnon (1978) suggest innovative approaches to seasonal factors. See also Judge et al. (1985, pp. 258-62) and Darnell (1994, pp. 359-63) for discussion of the issues involved.

14.5 Observation-specific Dummies

Salkever (1976) introduced the use of observation-specific dummies for facilitating estimation; see Kennedy (1990) for an exposition. Pagan and Nicholls (1984) suggest several extensions, for example to the context of autocorrelated errors.

The Chow test as described earlier cannot be performed whenever there are too few observation in one of the data sets to run a regression. In this case an alternative (and less-powerful) version of the Chow test is employed, involving the use of observation-specific dummies. Suppose that the number of observations $T2$ in the second time period is too small to run a regression. $T2$ observation-specific dummy variables are

formed, one for each observation in the second period. Each dummy has a value of 1 for its particular observation and 0 elsewhere. Regressing on the K independent variables plus the $T2$ dummies over the $T1 + T2$ observations gives the unrestricted regression, identical to the regression using the K independent variables and $T1$ observations. (This identity arises because the coefficient of each dummy variable takes on whatever value is necessary to create a perfect fit, and thus a zero residual, for that observation.)

The restricted version comes from restricting each of the $T2$ dummy variable coefficients to be zero, yielding a regression identical to one using the K independent variables and $T1 + T2$ observations. The F statistic thus becomes:

$$\frac{[SSE(\text{constrained}) - SSE(\text{unconstrained})]/T_2}{SSE(\text{unconstrained})/(T_1 - K)}$$

This statistic can be shown to be equivalent to testing whether or not the second period's set of observations falls within the prediction confidence interval formed by using the regression from the first period's observations. This dummy-variable approach, introduced in the first edition of this book, has been formalized by Dufour (1980).

14.6 Fixed and Random Effects Models

Baltagi (1995, pp. 17) is an excellent reference for the econometrics of panel data, offering descriptions of major panel data sets such as the Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys of Labor Market Experience (NLS), and discussion of the benefits and limitations/problems of panel data. Estimation with panel data allows us to control for individual heterogeneity, alleviate aggregation bias, improve efficiency by using data with more variability and less collinearity, estimate and test more complicated behavioral models, and examine adjustment dynamics. Furthermore, this type of data allows examination of some issues that otherwise could not be broached. For example, in a specific cross-section a high percentage of people may be unemployed, but from that alone we cannot tell if this

percentage is an average or if the same people are unemployed in every period. As a second example, consider the problem of separating economies of scale from technological change. Cross-sectional data provide information on the former, while time series data mix the two. In both these examples, panel data allow the researcher to resolve these issues.

Fixed and random effects models are usually employed when the number of cross-sectional units is large and the number of time periods over which those units are observed is small. When the reverse is the case, several alternative models are common, differing in the assumptions they make regarding the error variance-covariance matrix. The simplest case assumes that each cross-section unit has an error with a different variance, so a simple correction for heteroskedasticity is employed. A slightly more complicated case is to assume also contemporaneous correlation between the errors of different cross-section units. A further complication would be to allow for errors to be autocorrelated across time in some way. All these models require EGLS estimation, which Beck and Katz (1995) find in practice performs very poorly in this context because the error variance-covariance matrix is poorly estimated. They recommend using OLS with its variance-covariance matrix estimated by (X'X)-1X'WX(X'X)-1 where W is an estimate of the error variance-covariance matrix. Baltagi (1986) uses a Monte Carlo study to compare these types of estimators to random effects estimators, concluding that the loss in efficiency is less severe when employing incorrectly the random effects estimator than when the alternatives are employed incorrectly.

Greene (1997, chapter 14) has an excellent textbook exposition of estimation with panel data, including examples, computational simplifications, relationships among various estimators, and relevant test statistics. Baltagi and Griffin (1984) have a good discussion of the issues. Judge et al. (1985) and Dielman (1983) also have useful surveys.

Gumpertz and Pantula (1989) suggest using the mean of the parameter estimates from OLS estimation (on each cross-sectional unit separately) for inference in the random effects model.

Robertson and Symons (1992) suggest that if the slope parameters are not the same for all observations in panel data estimation, but estimation forces equality, serious bias problems arise. If one is not certain whether the coefficients are identical, Maddala (1991) recommends shrinking the separate estimates towards some common estimate.

Technical Notes

Analysis of variance is a statistical technique designed to determine whether or not a particular classification of the data is meaningful. The total variation in the dependent variable (the sum of squared differences between each observation and the overall mean) can be expressed as the sum of the variation between classes (the sum of the squared differences between the mean of each class and the overall mean, each times the number of observations in that class) and the variation within each class (the sum of the squared difference between each observation and its class mean). This decomposition is used to structure an F test to test the hypothesis that the between-class variation is large relative to the within-class variation, which implies that the classification is meaningful, i.e., that there is a significant variation in the dependent variable between classes.

If dummy variables are used to capture these classifications and a regression is run, the dummy variable coefficients turn out to be the class means, the between-class variation is the regression's "explained" variation, the within-class variation is the regression's "unexplained" variation, and the analysis of variance F test is equivalent to testing whether or not the dummy variable coefficients are significantly different from one another. The main advantage of the dummy variable regression approach is that it provides estimates of the magnitudes of class variation influences on the dependent variables (as well as testing whether the classification is meaningful).

Analysis of covariance is an extension of analysis of variance to handle cases in which there are some uncontrolled variables that could not be standardized between classes. These cases can be analyzed by using dummy variables to capture the classifications and regressing the dependent variable on these dummies and the uncontrollable variables. The analysis of covariance F tests are equivalent to testing whether the coefficients of the dummies are significantly different from one another. These tests can be interpreted in terms of changes in the residual sums of squares caused by adding the dummy variables. Johnston (1972, pp. 192207) has a good discussion.

In light of the above, it can be concluded that anyone comfortable with regression analysis and dummy variables can eschew analysis of variance and covariance techniques.

15
Qualitative Dependent Variables

15.1 Dichotomous Dependent Variables

When the dependent variable is qualitative in nature and must be represented by a dummy variable, special estimating problems arise. Examples are the problem of explaining whether or not an individual will buy a car, whether an individual will be in or out of the labor force, whether an individual will use public transportation or drive to work, or whether an individual will vote yes or no on a referendum.

If the dependent variable is set up as a 01 dummy variable (for example, the dependent variable is set equal to 1 for those buying cars and equal to 0 for those not buying cars) and regressed on the explanatory variables, we would expect the predicted values of the dependent variable to fall mainly within the interval between 0 and 1, as illustrated in figure 15.1. This suggests that the predicted value of the dependent variable could be interpreted as the probability that that individual will buy a car, given that individual's characteristics (i.e., the values of the explanatory variables). This is in fact the accepted convention. In figure 15.1 the dots represent the sample observations; most of the high values of the explanatory variable x correspond to a dependent dummy variable value of unity (implying that a car was bought), whereas most of the low values of x correspond to a dependent dummy variable value of zero (implying that no car was bought). Notice that for extremely low values of x the regression line yields a negative estimated probability of buying a car, while for extremely high values of x the estimated probability is greater than 1. As should be clear from this diagram, R2 is likely to be very low for this kind of regression, suggesting that R2 should not be used as an estimation criterion in this context.

An obvious drawback to this approach is that it is quite possible, as illustrated in figure 15.1, to have estimated probabilities outside the 01 range. This embarrassment could be avoided by converting estimated probabilities lying outside the 01 range to either 0 or 1 as appropriate. This defines the *linear probability model*. Although this model is often used because of its computational ease,

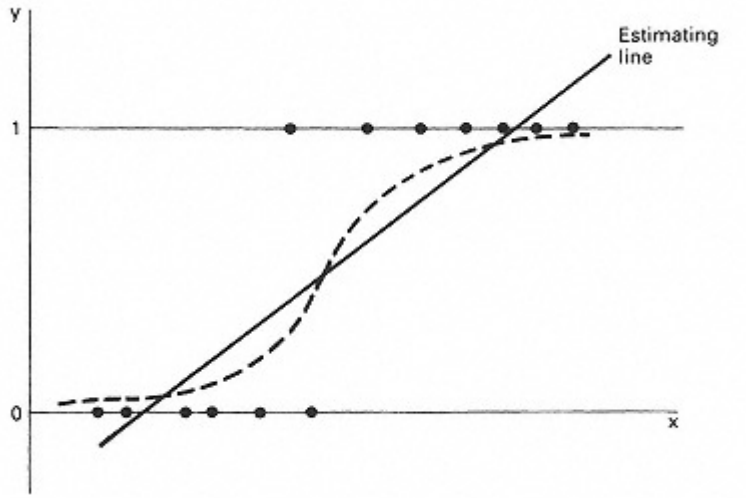

Figure 15.1
The linear probability model

many researchers feel uncomfortable with it because outcomes are sometimes predicted with certainty when it is quite possible that they may not occur.

What is needed is some means of squeezing the estimated probabilities inside the 01 interval without actually creating probability estimates of 0 or 1, as shown by the dashed line in figure 15.1. Many possible functions of this nature are available, the two most popular being the cumulative normal function and the logistic function. Using the cumulative normal function for this purpose creates the *probit* model; using the logistic function creates the *logit* model. These two functions are very similar, and in today's software environment the choice

between them is a matter of taste because both are so easy to estimate. Logit is more common, perhaps for historical reasons - its lower computational cost made it more common before modern software eliminated this advantage.

A novel feature of these models, relative to the traditional regression model, is that the stochastic ingredient is no longer represented by an error term. This is because the stochastic element in this model is inherent in the modeling itself - the logit equation, for example, provides the expression for the probability that an event will occur. For each observation the occurrence or non-occurrence of that event comes about through a chance mechanism determined by this probability, rather than by a draw from a bowl of error terms.

Estimation is almost always undertaken by maximum likelihood. For the logit case, for example, the logit function provides the probability that the event will occur and one minus this function provides the probability that it will not occur. The likelihood is thus the product of logit functions for all observations for which the event occurred multiplied by the product of one-minus-the-logit-func-

tions for all observations for which the event did not occur. This is formalized in the technical notes.

15.2 Polychotomous Dependent Variables

The preceding section addressed the problem of binary, or dichotomous, variables, for which there are only two choice categories. Categorical variables that can be classified into many categories are called polychotomous variables. For example, a commuter may be presented with a choice of commuting to work by subway, by bus or by private car, so there are three choices. Estimation in this context is undertaken by means of a generalization of the logit or probit models, called, respectively, the multinomial logit and the multinomial probit models. These generalizations are motivated by employing the random utility model.

In the random utility model the utility to a consumer of an alternative is specified as a linear function of the characteristics of the consumer and the attributes of the alternative, plus an error term. The probability that

a particular consumer will choose a particular alternative is given by the probability that the utility of that alternative to that consumer is greater than the utility to that consumer of all other available alternatives. This makes good sense to an economist. The consumer picks the alternative that maximizes his or her utility. The multinomial logit and multinomial probit models follow from assumptions made concerning the nature of the error term in this random utility model.

If the random utility error terms are assumed to be independently and identically distributed as a log Weibull distribution, the *multinomial logit* model results. The great advantage of this model is its computational ease; the probability of an individual selecting a given alternative is easily expressed (as described in the technical notes), and a likelihood function can be formed and maximized in straightforward fashion. The disadvantage of this model is that it is characterized by what is called the *independence of irrelevant alternatives* property. Suppose a new alternative, almost identical to an existing alternative, is added to the set of choices. One would expect that as a result the probability from this model of choosing the duplicated alternative would be cut in half and the probabilities of choosing the other alternatives would be unaffected. Unfortunately, this is not the case, implying that the multinomial logit model will be inappropriate whenever two or more of the alternatives are close substitutes.

If the random utility error terms are assumed to be distributed multivariate-normally, the *multinomial probit* model results. This model allows the error terms to be correlated across alternatives, thereby permitting it to circumvent the independence of irrelevant alternatives dilemma. Its disadvantage is its high

computational cost, which becomes prohibitively high when there are more than four alternatives.

15.3 Ordered Logit/Probit

For some polychotomous dependent variables there is a natural order. Bond ratings, for example, are expressed in terms of categories (triple A, double A, etc.) which could be viewed as resulting from a continuous, unobserved measure called "creditworthiness"; students' letter grades for an economics course may be generated by their

instructor's assessment of their "level of understanding" of the course material; the reaction of patients to a drug dose could be categorized as no reaction, slight reaction, severe reaction, and death, corresponding to a conceptual continuous measure called "degree of allergic reaction."

For these examples, using multinomial probit or logit would not be efficient because no account would be taken of the extra information implicit in the ordinal nature of the dependent variable. Nor would ordinary least squares be appropriate, because the coding of the dependent variable in these cases, usually as 0, 1,2,3, etc., reflects only a ranking: the difference between a 1 and a 2 cannot be treated as equivalent to the difference between a 2 and a 3, for example.

The *ordered logit* or *probit* model is used for this case. Consider the example of bond ratings, for which the unobserved continuous measure, creditworthiness, is specified to be a linear function (with parameter vector b, say) of explanatory variables. Each bond rating corresponds to a specific range of the creditworthiness index, with higher ratings corresponding to a higher range of the creditworthiness values. Suppose, for example, that a firm's current bond rating is A. If its creditworthiness were to grow, it would eventually exceed the creditworthiness value that marks the boundary between the A and double A categories, and the firm would then experience an increase in its bond rating. Estimation is undertaken by maximum likelihood, with b being estimated in conjunction with estimation of the unknown boundary values defining the ranges of the creditworthiness index. For further discussion see the technical notes.

15.4 Count Data

Very often data take the form of non-negative integer values such as number of children, recreational trips, bankruptcies, or patents. To exploit this feature of the data, estimation is undertaken using a count-data model, the most common example of which is a Poisson model. In this model the Poisson distribution provides the probability of the number of event occurrences and the Poisson parameter corresponding to the expected number of occurrences is modeled as a function of explanatory variables. Estimation is undertaken by maximum likelihood.

The Poisson model embodies some strong assumptions, such as that the prob-

ability of an occurrence is constant at any point in time and that the variance of the number of occurrences equals the expected number of occurrences. Both are thought to be unreasonable since contagious processes typically cause occurrences to influence the probability of future occurrences, and the variance of the number of occurrences usually exceeds the expected number of occurrences. Generalizations of the Poisson model, discussed in the technical notes, are employed to deal with these shortcomings.

General Notes

Maddala (1993) is an extensive reference on qualitative dependent variables and modelling options. Amemiya (1981) is a classic survey article for qualitative choice. Fry et al. (1993) discuss economic motivations for models with qualitative dependent variables. Winkelmann and Zimmermann (1995) is a good survey of count-data modeling; Winkelmann (1997) is a comprehensive reference. LIMDEP is the software of choice for estimating models discussed in this chapter.

15.1 Dichotomous Dependent Variables

Although estimation of the dichotomous or binary dependent variable is almost always by maximum likelihood, on occasion one sees an alternative procedure, popular before computer software made maximum likelihood so simple to perform. This case occurs when there is a very large data set, large enough that observations can be grouped into sets of several observations on identical individuals. If there are enough observations in each group, a reliable estimate of the probability of an observation in that group experiencing the event can be produced by calculating the percentage of observations in that group experiencing the event. (Alternatively, the data may be available only in aggregated form.) This estimated probability can be used in two ways to provide estimates. First, it can be used as the dependent variable in a regression on the group characteristics to estimate a linear probability model. Second, the log of the ratio of this probability to one minus this probability (the log-odds ratio) can be used as the dependent variable in a regression on the group characteristics to estimate a logit function. (The technical notes show how this comes about.) In both cases there is heteroskedasticity that should be adjusted for.

The role of the error term in qualitative dependent variable models is not obvious. In traditional regression models the dependent variable is written as a linear function of several explanatory variables plus an error, so that for example we have y = Xb + e. For qualitative dependent variables, however, the probability of obtaining the dependent variable value is written as a logit or probit function of these explanatory variables, without an error term appearing, so that for example.

$$\text{prob}(y = 1) = \text{logit}(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

An error term is not necessary to provide a stochastic ingredient for this model because for each observation the value of the dependent variable is generated via a chance mechanism embodying the probability provided by the logit equation.

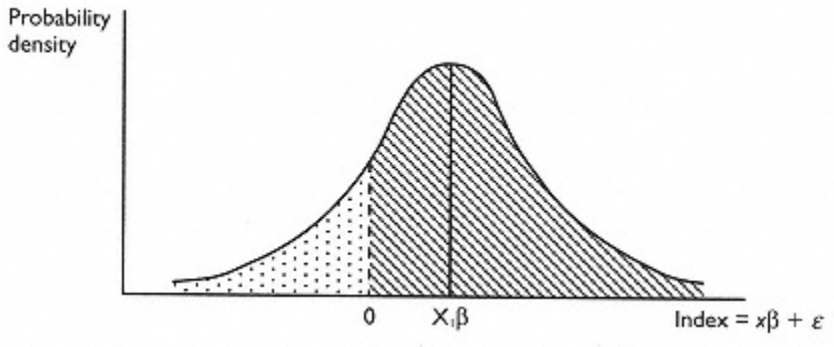

Figure 15.2
Explaining probit and logit

Despite this, most researchers conceptualize an underlying model that does contain an error term. An unobserved (latent) index is specified as a linear function of explanatory variables plus an error term (i.e., Xb + e). If this index exceeds a critical value (normalized to be zero, assuming an intercept appears in Xb) then y = 1, otherwise y = 0. More formally.

$$\text{prob}(y = 1) = \text{prob}(X\beta + \varepsilon > 0) = \text{prob}(\varepsilon > -X\beta)$$

which is a cumulative density. If e is distributed normally this is the cumulative density of a normal distribution (normalized to have variance one, which scales the coefficient estimates) and we have the probit model; if e is distributed such that its cumulative density is a logistic function, we have the logit model.

Thinking of this model in terms of its underlying latent index can be advantageous for several reasons. First, it provides a means of interpreting outcomes in terms of the theoretically attractive random utility model, as described later in the technical notes. Second, it facilitates the exposition of ordered logit/probit, discussed later in this chapter. Third, it is consistent with the modeling of sample selection problems, presented in chapter 16. And fourth, it allows the development of R2 measures applicable to this context.

Figure 15.2 illustrates the exposition given above. Suppose we are modeling the decision to buy a car, so that the latent index Xb + e is referred to as a "buying index," and if an individual's buying index exceeds zero, he or she buys. An individual with characteristics given by the row vector X1 has buying index X1b + e, so the density of buying indices for such people is shown in figure 15.2 centered at X1b. Some such individuals need little encouragement to buy a car and so have high, positive error terms producing high index values, whereas other seemingly identical individuals hate buying cars and so have large, negative error terms producing low index values. The probability that such a person buys is the probability that his or her index value exceeds zero, given by the lined area to the right of zero. If e is distributed normally this is the cumulative density of e from minus X1b to infinity, equal to the cumulative density from minus infinity to plus X1b. This is just the probit model. For the logit

model this area/probability is given by $\frac{e^{X_i\beta}}{1 + e^{X_i\beta}}$

For an individual with a different row vector of characteristics the lined area would be of a different size. The likelihood function is formed by multiplying together expressions for the probability of each individual in the sample doing what he or she did (buy or not buy). Expressions measuring the lined area are used for buyers and expressions for the dotted area (one minus the expression for the lined area) are used for those not buying.

Cameron (1988) shows how to undertake logit estimation in the context of "referendum" survey data when people are asked to answer yes or no to a choice question, such as willingness to pay for a project, with the payment varying across respondents.

An estimated b value in a logit or a probit does not estimate the change in the probability of $y = 1$ due to a unit change in the relevant explanatory variable. This probability change is given by the partial derivative of the expression for prob$(y = 1)$ with respect to b, which is not equal to b. For the logit, for example, it is [prob$(y = 1)$][1 - prob$(y = 1)$]b, which is usually reported by estimating it at the mean values of the explanatory variables. It should be noted that this formula can give misleading estimates of probability changes in contexts in which an explanatory variable is postulated to change by an amount that is not infinitesimal. Estimation using the difference between the estimated prob$(y = 1)$ before and after the change is safer. See Caudill and Jackson (1989).

There is no universally-accepted goodness-of-fit measure (pseudo-R2) for probit, logit, or count-data models. Veall and Zimmermann (1996) is a good survey of alternative measures and their relative attributes. They recommend the measure of McKelvey and Zavoina (1975), a pseudo-R2 which is close to what the OLS R2 would be using the underlying latent index implicit in the model. Most computer packages provide a table giving the number of $y = 1$ values correctly and incorrectly predicted, and the number of $y = 0$ values correctly and incorrectly predicted, where an observation is predicted as $y = 1$ if the estimated prob$(y = 1)$ exceeds one-half. It is tempting to use the percentage of correct predictions as a measure of goodness of fit. This temptation should be resisted: a naive predictor, for example that every $y = 1$, could do well on this criterion. A better measure along these lines is the sum of the fraction of zeros correctly predicted plus the fraction of ones correctly predicted, a number which should exceed unity if the prediction method is of value. See McIntosh and Dorfman (1992). It should be noted that a feature of logit is that the number of $y = 1$

predictions it makes is equal to the number of $y = 1$ observations in the data.

One use of logit models is to classify observations. Suppose a logit analysis has been done on the dichotomous choice of public versus private transportation. Given the characteristics of a new individual, the probabilities that he or she will choose public or private transportation are estimated from the estimated logit function, and he or she is classified to whichever transportation mode has the higher estimated probability.

The main competitor to logit for classification is *discriminant analysis*. In this technique it is assumed that the individual's characteristics can be viewed as being distributed multivariate-normally, with a different mean vector (but the same variance-covariance matrix) associated with the two transportation modes. The original data are used to estimate the two mean vectors and the joint variance-covariance matrix. Given a new individual's characteristics these estimates can be used to estimate the height of the density function for each transportation mode; the new observation is classified to the transportation mode with the higher estimated density (since it is "more likely" to have come from that category).

Most studies, such as Press and Wilson (1978), have concluded that logit is superior to discriminant analysis for classification, primarily because the assumption of multivariate-normally distributed characteristics is not reasonable, especially when some characteristics are qualitative in nature (i.e., they are represented by dummy variables). Recently, linear programming techniques for classification have been providing strong competition for logit. See Freed and Glover (1982). Kennedy (1991b) provides a graphical comparison of these three classification techniques.

By adding an error to the traditional logit or probit specifications so that, for example,

$$\text{prob}(y = 1) = \frac{e^{X\beta+\varepsilon}}{1 + e^{X\beta+\varepsilon}}$$

it is possible to model unobserved differences between individuals beyond those captured by the error in the latent index. Although this unobserved heterogeneity, as it is called, is important in some contexts, such as count-data models or duration models, Allison (1987) finds that in logit and probit models it is a problem only in special cases.

Unfortunately, logit and probit models are sensitive to misspecifications. In particular, in contrast to OLS in the CLR model, estimators will be inconsistent if an explanatory variable (even an orthogonal variable) is omitted or if there is heteroskedasticity. Davidson and MacKinnon (1993, pp. 523-8) suggest a computationally attractive way of using a modified Gauss-Newton regression to test for various specification errors. Murphy (1994) shows how a heteroskedasticity test of Davidson and MacKinnon (1984) can be applied to the multinomial case. Landwehr et al. (1984) suggest some graphical means of assessing logit models. Grogger (1990) exposits a Hausman-type specification test for exogeneity in probit, logit and Poisson regression models. Lechner (1991) is a good exposition of specification testing in the context of logit models. Pagan and Vella (1989) is a classic paper showing that many difficult-to-derive LM tests for specification in qualitative and limited dependent variable models can more easily be undertaken as conditional moment tests. Greene, Knapp and Seaks (1995) show how a Box-Cox transformation can be used to allow a more flexible functional form for the independent variables in a probit model. Fader, Lattin and Little (1992) address the problem of estimating nonlinearities within the multinomial logit model.

15.2 Polychotomous Dependent Variables

There are three ways of structuring the deterministic part of the random utility model.

(1) Specify that the utility of an alternative to an individual is a linear function of that individual's *n* characteristics, with a different set of parameters for each alternative. In this case *n* coefficients must be estimated for each of the alternatives (less one - as shown in the example in the technical notes, one alternative serves as a base). Given characteristics of an individual - say, income, sex, and geographic location - with this specification one could estimate the probabilities of that individual choosing, say, each type of commuter mode.

(2) Specify that the utility of an alternative to an individual is a linear function of the *m* attributes of that alternative, as seen through the eyes of that individual. In this case, *m* coefficients, identical for all individuals, must be estimated. Given how the characteristics of an alternative relate to an individual - say, commuting time, cost, and convenience - it would be possible to estimate the probabilities of that individual choosing each type of commuter mode. If the researcher wanted to capture inherent differences between the alternatives that are the same for all individuals, dummy variables for all but one alternative would be included.

(3) Specify a combination of (1) and (2) above, namely a linear function of both the attributes of the alternatives as they affect the individuals and the characteristics of the individuals, with a different set of parameters for the individuals' characteristics for each alternative (less one) plus one set of parameters for the alternatives' attributes.

Specification (1) above is called the multinomial logit/probit, specification (2) is called the conditional logit/probit, and specification (3) is called the mixed logit/probit model. Frequently, as is the case in this book, the multinomial terminology is used to refer to all three.

The independence-of-irrelevant-alternatives problem arises from the fact that in the multinomial logit model the *relative* probability of choosing two existing alternatives is unaffected by the presence of additional alternatives. As an example, suppose a commuter is twice as likely to commute by subway as by bus and three times as likely to commute by private car as by bus, so that the probabilities of commuting by bus, subway and private car are 1/6, 2/6 and 3/6, respectively. Now suppose an extra bus service is added, differing from the existing bus service only in the color of the buses. One would expect the probabilities of commuting by new bus, old bus, subway and private car to be 1/12, 1/12, 2/6 and 3/6, respectively. Instead, the multinomial logit model produces probabilities 1/7, 1/7, 2/7 and 3/7, to preserve the relative probabilities.

One way of circumventing the independence of irrelevant alternatives problem is to estimate using a sequential logit/probit model. In this model people are assumed to make decisions sequentially. For example, rather than choosing between an imported car, an imported truck, a domestic car and a domestic truck, which creates an IIA problem,

people are assumed first to make a decision between a car and a truck and then, conditional on that decision, to choose between an imported and a domestic model. Van Ophem and Schram (1997) examine a model that nests sequential and multinomial logit models and so allows testing one against the other.

Hausman and McFadden (1984) develop tests for the independence of irrelevant alternatives (IIA) assumption. One test is based on the idea that if a category is dropped then if the IIA assumption is true the estimated coefficients should not change. A second test is based on the fact that under IIA a multinomial logit is a special case of a sequential logit. Zhang and Hoffman (1993) have a good exposition of these methods of testing for IIA, recommending a procedure due to Small and Hsiao (1985).

A flexibility of the multinomial probit model is that the coefficients of the individual characteristics in the random utility model can be stochastic, varying (normally) over individuals to reflect individual's different tastes. This can be incorporated in the multinomial probit model through the covariances of the error terms; it cannot be made part of the multinomial logit model because the covariance between the error

terms must be zero. This restriction of the multinomial logit model is what gives rise to the IIA problem. Using the multinomial probit to circumvent this problem involves very high computational cost, however, because multiple (one less than the number of categories) integrals must be calculated, which is why the multinomial logit is used so often despite the IIA problem. Chintagunta (1992) has introduced a computationally feasible means of estimation for the multinomial probit model. Keane (1992) discusses the computational problems associated with calculating the multinomial probit. He notes that although exclusion restrictions (some explanatory variables do not affect the utility of some options) are not required for estimation, in practice estimation is questionable without them.

15.3 Ordered Logit/Probit

Surveys often ask respondents to select a range rather than provide a specific value, for example indicating that income lies in one of several specified ranges. Is the measurement error avoided by asking respondents to select categories worth the loss in information associated with foregoing a continuous measure? By comparing OLS and ordered logit with a unique data set, Dunn (1993) concludes that it is better to avoid gathering categorical data. Stewart (1983), Stern (1991), Caudill and Jackson (1993) and Bhat (1994) suggest ways of estimating in this context.

Murphy (1996) suggests an artificial regression for testing for omitted variables, heteroskedasticity, functional form and asymmetry in ordered logit models.

15.4 Count Data

Many individual decisions are made in a two-stage process in which first a decision is made to, say, purchase a good, and then a decision is made on the number of purchases. This can lead to more or fewer zeros in the data than that predicted by the Poisson model. The hurdle Poisson model is used to deal with this problem, in which a dichotomous model capturing the first stage is combined with a Poisson model for the second stage. This approach has been extended by Terza and Wilson (1990) to allow a choice of several different types of trips, say, in conjunction with choice of number of trips.

In some applications zero values are unobserved because, for example, only people at a recreational site were interviewed about the number of trips they made per year to that site. In this case a truncated count-data model is employed in which the formula for the Poisson distribution is rescaled by dividing by one minus the probability of zero occurrences. Interestingly, the logit model results from truncating above one to produce two categories, zero and one. Caudill and Mixon (1995) examine the related case in which observations are censored (i.e., the explanatory variables are observed but the count is known only to be beyond some limit) rather than truncated (no observations at all beyond the limit).

15.1 Dichotomous Dependent Variables

If the linear probability model is formulated as $Y = X\mathrm{b} + \mathrm{e}$ where Y is interpreted as the probability of buying a car, the heteroskedastic nature of the error term is easily derived by noting that if the individual buys a car (probability Xb) the error term takes the value $(1 - X\mathrm{b})$ and that if the individual does not buy a car (probability $(1 - X\mathrm{b})$) the error term takes the value $-X\mathrm{b}$.

The logistic function is given as $f(\mathrm{q}) = \mathrm{eq}/(1 + \mathrm{eq})$. It varies from zero to one as q varies from $-\infty$ to $+\infty$, and looks very much like the cumulative normal distribution. Note that it is much easier to calculate than the cumulative normal, which requires evaluating an integral. Suppose q is replaced with an index xb, a linear function of (for example) several characteristics of a potential buyer. Then the logistic model specifies that the probability of buying is given by

$$\text{prob(buy)} = \frac{e^{x\beta}}{1 + e^{x\beta}}$$

This in turn implies that the probability of not buying is

$$\text{prob(not buy)} = 1 - \text{prob(buy)} = \frac{1}{1 + e^{x\beta}}$$

The likelihood function is formed as

$$L = \prod_i \frac{e^{x\beta}}{1 + e^{x\beta}} \prod_j \frac{1}{1 + e^{x\beta}}$$

where i refers to those who bought and j refers to those who did not buy.

Maximizing this likelihood with respect to the vector b produces the MLE of b. For the nth individual, then, the probability of buying is estimated as

$$\frac{e^{x_i\beta^{(1)}}}{1 + e^{x_i\beta^{(1)}}}$$

The formulae given above for the logit model imply that

$$\frac{\text{prob(buy)}}{\text{prob(not buy)}} = e^{x\beta}$$

so that the log-odds ratio is

$$\ln\left[\frac{\text{prob(buy)}}{\text{prob(not buy)}}\right] = x\beta.$$

This is the rationale behind the grouping method described earlier.

The logic of discriminant analysis described earlier is formalized by the *linear discriminant rule,* namely classify an individual with characteristics given by the vector *x* to category 1 if

$$(\mu_1 - \mu_2)'\Sigma^{-1}x > (1/2)(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$$

where the m*i* are the estimated mean vectors of the characteristics vectors of individuals in category *i,* and S is their estimated common variance-covariance matrix. This is easily derived from the formula for the multivariate normal distribution. This rule can be modified for cases in which the prior probabilities are unequal or the misclassification costs are unequal. For example, if the cost of erroneously classifying an observation to category 1 were three times the cost of erroneously classifying an observation to category 2, the 1/2 in the linear discriminant rule would be replaced by 3/2.

15.2 Polychotomous Dependent Variables

The log Weibull distribution, also known as the type I extreme-value distribution, has the convenient property that the cumulative density of the difference between any two random variables with this distribution

is given by the logistic function. Suppose, for example, that the utility of option A to an individual with a row vector of characteristics *x*0 is *x*0bA + eA and of option B is *x*0bB + eB where eA and eB are drawn independently from a log Weibull distribution. This individual will choose option A if

$$x_0\beta_B + \varepsilon_B < x_0\beta_A + \varepsilon_A$$

or, alternatively, if

$$\varepsilon_B - \varepsilon_A < x_0(\beta_A - \beta_B).$$

The probability that this is the case is given by the cumulative density of eB - eA to the point *x*0(bA - bB). Since the cumulative density of eB - eA is given by the logistic function we have

$$\text{prob(choose option A)} = \frac{e^{x_0(\beta_A - \beta_B)}}{1 + e^{x_0(\beta_A - \beta_B)}}$$

This shows, for the binary case, the relationship between the random utility function and the logit model. A similar result for the polychotomous case can be derived (see Maddala, 1993, pp. 5961), producing the multinomial logit model, a generalization of the binary logit. Notice that both bA and bB cannot be estimated; one category serves as a base and the estimated coefficients (bA - bB) reflect the difference between their utility function coefficients.

The type I extreme-value (or log Weibull) distribution has density *f(x)* = exp*(-x* - e-*x),* with cumulative density *F(x* < *a)* = exp(-e-*a*). Its mode is at zero, but its mean is 0.577. Consider the random utility model with the utility of the *i*th option to the *j*the individual given by *Ui* = *Xjbi* + e*i(i* = 1, 2) with the e*i* distributed

as independent log Weibulls. The probability that the *j*th individual chooses option 1 is

$$\text{prob}[\varepsilon_2 < \varepsilon_1 + X_j(\beta_1 - \beta_2)] = \int \text{prob}(\varepsilon_1)\text{prob}[\varepsilon_2 < \varepsilon_1 + X_j(\beta_1 - \beta_2)|\varepsilon_1]d\varepsilon_1.$$

By exploiting the fact that the integral of the density above is unity, this can be shown to be the logit $\{1 + \exp[Xj(b2 - b1)]\}$-1.

A proper derivation of the multinomial logit is based on the random utility model. The resulting generalization of the binary logit can be illustrated in less rigorous fashion by specifying that the ratio of the probability of taking the *k*th alternative to the probability of taking some "base" alternative is given by $e^{x\beta_k}$ where b*k* is a vector of parameters relevant for the *k*th alternative. This is a direct generalization of the earlier result that prob(buy)/prob(not buy) = e*x*b. Note that this ratio is unaffected by the presence of other alternatives; this reflects the independence of irrelevant alternatives phenomenon. Note also that the coefficient estimates change if the "base" alternative is changed (as they should, because they estimate something different); if different computer packages normalize differently in this respect, they will not produce identical estimates.

As an example of how this generalization operates, suppose there are three alternatives A, B and C, representing commuting alone (A), by bus (B), and by carpool (C). The model is specified as

$$\frac{\text{prob(A)}}{\text{prob(C)}} = e^{x\beta_A} \text{ and } \frac{\text{prob(B)}}{\text{prob(C)}} = e^{x\beta_B}$$

Here carpooling is chosen as the "standard" or base alternative; only two such ratios are necessary since the remaining ratio, prob(A)/prob(B), can be derived from the other two. Using the fact that the sum of the probabilities of the three alternatives must be unity, a little algebra reveals that

$$\text{prob(A)} = \frac{e^{x\beta_A}}{1 + e^{x\beta_A} + e^{x\beta_B}}$$

$$\text{prob(B)} = \frac{e^{x\beta_B}}{1 + e^{x\beta_A} + e^{x\beta_B}}$$

$$\text{prob(C)} = \frac{1}{1 + e^{x\beta_A} + e^{x\beta_B}} .$$

The likelihood function then becomes

$$L = \prod_i \frac{e^{x_i\beta_A}}{1 + e^{x_i\beta_A} + e^{x_i\beta_B}} \prod_j \frac{e^{x_j\beta_B}}{1 + e^{x_j\beta_A} + e^{x_j\beta_B}} \prod_k \frac{1}{1 + e^{x_k\beta_A} + e^{x_k\beta_B}}$$

where the subscripts *i,j,* and *k* refer to those commuting alone, by bus and by carpool, respectively. This expression, when maximized with respect to bA and bB, yields β_A^{MLE} and β_B^{MLE}. For any particular individual, his or her characteristics can be used, along with β_A^{MLE} and β_B^{MLE}, to estimate prob(A), the probability that that person will commute to work alone, prob(B), the probability that that person will commute to work by bus, and prob(C), the probability that he or she will carpool it. Extension of this procedure to more than three alternatives is straightforward.

The commuter example can be used to describe more fully the independence-of-irrelevant-alternatives phenomenon. Suppose we were to use the same data to estimate a logit model expanded to discriminate between commuting on a red bus (RB) versus commuting on a blue bus (BB). In the original example these two alternatives had been lumped together. Now there are four alternatives, A, RB, BB and C. Assuming everyone is indifferent between blue and red buses, it would seem logical that, when estimated, the expanded model should be such that for any individual each of the estimated probabilities of commuting alone, taking the bus (either red or blue) and carpooling it should remain

unchanged, with the probability of riding the bus broken in half to estimate each of the two bus line alternatives. Unfortunately, this is not the case: adding an irrelevant alternative changes the probabilities assigned to all categories.

The key to understanding why this comes about is to recognize that the number of people in the data set who commute by bus, relative to the number of people in the data set who, say, carpool it, is irrelevant from the point of view of calculating the estimate of bB. It is the differences in these people's characteristics that determine the estimate of bB. If the people riding the bus are now arbitrarily divided into two categories, those riding red buses and those riding blue buses, there will be a change in the number of people in a bus category relative to the carpool category, but there will be no change in the nature of the differences in the characteristics of people in the bus categories versus people in the carpool category. Consequently, the estimate of bRB (where $\text{prob(RB)/prob(C)} = e^{x\beta_{RB}}$) will be virtually the same as the original estimate of bB, as will the estimate of bBB.

For the *n*th individual, before the introduction of the irrelevant alternative, the probability of commuting alone is estimated as

$$\text{prob(A)} = \frac{e^{x_n\beta_A^{MLE}}}{1 + e^{x_n\beta_A^{MLE}} + e^{x_n\beta_B^{MLE}}}$$

a probability we would hope would remain unchanged when the irrelevant alternative is introduced. But it does change; by setting $\beta_B^{MLE} = \beta_{RB}^{MLE} = \beta_{BB}^{MLE}$ it becomes approximately

$$\text{prob(A)} = \frac{e^{x_n\beta_A^{MLE}}}{1 + 2e^{x_n\beta_A^{MLE}} + e^{x_n\beta_B^{MLE}}}$$

Because of this problem, the multivariate logit methodology can be used only when the categories involved are all quite different from one another.

15.3 Ordered Logit/Probit

Ordered probit specifies that, for example, $y^* = a + bx + e$ is an unobservable index of "creditworthiness," and we observe $y = B$ if $y^* \leq d1$, $y = A$ if $d1 \leq y^* \leq d2$, $y = AA$ if $d2 \leq y^* \leq d3$ and $y = AAA$ if $d3 \leq y^*$. The ds are unknown "threshold" parameters that must be estimated along with a and b. If an intercept is included in the equation for y^*, as it is here, it is customary to normalize by setting d1 equal to zero.

Estimation proceeds by maximum likelihood. The probability of obtaining an observation with $y = AA$, for example, is equal to

$$\begin{aligned} &\text{prob}(\delta_2 \leq y^* = \alpha + \beta x + \varepsilon \leq \delta_3) \\ &= \text{prob}(\delta_2 - \alpha - \beta x \leq \varepsilon \leq \delta_3 - \alpha - \beta x). \end{aligned}$$

A likelihood function can be formed, and thus estimation undertaken, once a density for e is known. The ordered probit model results from assuming that e is distributed normally. (The ordered logit model results from assuming that the cumulative density of e is the logistic function; in practice the two formulations yield very similar results.) The usual normalization is that e has mean zero and variance one; selecting a variance of four, say, would simply double the estimated values of the coefficients.

Application of ordered probit has become more frequent since it has been built into computer packages, such as LIMDEP. Greene (1990, pp. 7036) has a good textbook presentation; Becker and Kennedy (1992) have a graphical exposition. Note that if a change in an x value increases the creditworthiness index, the probability of having rating AAA definitely increases, the probability of having rating B definitely decreases, but the probabilities of being in the intermediate categories could move in either direction.

15.4 Count Data

In the Poisson model the probability of y number of occurrences of an event is given by e-lly/y! for y a non-negative integer. The mean and variance of this distribution are both l, typically specified to be l = exp(xb) where x is a row vector of explanatory variables. Choosing the exponential function has the advantage that it assures non-negativity.

Like the logit and probit models, in the Poisson model the formula for the probability of an occurrence is a deterministic function of the explanatory variables - it is not allowed to differ between otherwise-identical individuals. In the case of logit and probit, relaxation of this assumption can be achieved by introducing "unobserved heterogeneity" in the form of an error term, adding an extra stochastic ingredient. Unlike the case of logit and probit, however, in the Poisson model this addition makes a substantive difference to the model, allowing the variance of the number of occurrences to exceed the expected number of occurrences, thereby creating a model consistent with the almost universal tendency to observe such overdispersion.

A popular way of introducing unobserved heterogeneity into the Poisson model is to specify l as exp(xb + e) where e is an error distributed as a gamma distribution.

This leads to a negative binomial distribution for the number of occurrences, with mean l and variance l + a-1l2 where a is the common parameter of the gamma distribution. By assuming a to be different functions of l, different generalizations of this compound Poisson model are created.

An alternative way of modeling count data to produce overdispersion is to relax the assumption of the Poisson model that the probability of an occurrence is constant at any moment of time and instead allow this probability to vary with the time since the last occurrence. See Winkelmann (1995) and Butler and Worral (1991).

16
Limited Dependent Variables

16.1 Introduction

Dependent variables are sometimes limited in their range. For example, data from the negative income tax experiment are such that income lies at or below some threshold level for all observations. As another example, data on household expenditure on automobiles has a lot of observations at 0, corresponding to households who choose not to buy a car. As a last example, data on wage rates may be obtainable only for those for whom their wage exceeds their reservation wage, others choosing not to work. If the dependent variable is limited in some way, OLS estimates are biased, even asymptotically.

The upper half of figure 16.1 illustrates why this is the case (ignore for now the lower half of this diagram). The relationship $y = a + bx + e$ is being estimated, where e is a normally distributed error and observations with y values greater than k are not known. This could happen because y is the demand for tickets to hockey games and the arena on some occasions is sold out so that for these games all we know is that the demand for tickets is greater than k, the capacity of the arena. These unknown y values are denoted by small circles to distinguish them from known data points, designated by dots. Notice that for high values of x the known (dotted) observations below the (unconditional) expectation $E(y) = a + bx$ are not fully balanced off by observations above $E(y) = a + bx$, because some of these observations (the circled ones) are missing. This causes the resulting OLS regression line to be too flat, as shown by the dashed line.

Samples with limited dependent variables are classified into two general categories, censored and truncated regression models, depending on whether or not the values of x for the missing y data are known.

(1) *Censored sample* In this case some observations on the dependent variable, corresponding to known values of the independent variable(s), are not observable. In figure 16.1, for example, the y values corresponding to the circled data points are not known, but their corresponding x values are known. In a study of the determinants of wages, for example, you may have data on the explanatory variables for people who were not working, as

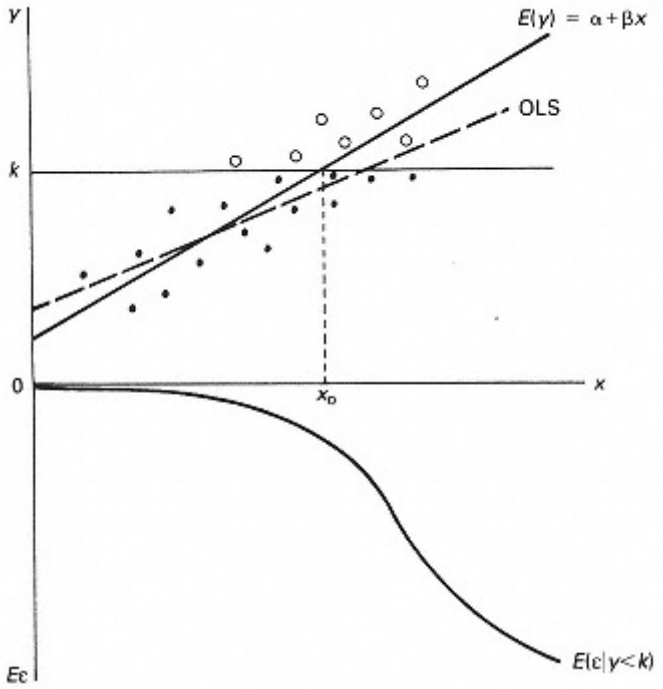

Figure 16.1
A limited dependent variable model

well as for those who were working, but for the former there is no observed wage.

(2) *Truncated sample* In this case values of the independent variable(s) are known only when the dependent variable is observed. In the example of the negative income tax experiment noted earlier, no data of any kind are available for those above the income threshold; they were not part of the sample.

The dependent variable can be limited in a variety of different ways, giving rise to several alternative models. The easiest of these models is the Tobit model for censored data.

16.2 The Tobit Model

A common feature of microeconomic data is that observations on the dependent variable that lie in a certain range are translated into (or reported as) a single

variable. In the demand for hockey game tickets example all demands above the capacity of the arena are translated into k, the arena capacity. This problem is analyzed using a Tobit model, named after James Tobin who was the first to analyze this type of data in a regression context.

How should estimation be undertaken? Our discussion earlier indicated that omitting the limit observations creates bias. Ignoring these observations would in any case be throwing away information, which is not advisable. How should they be included? It should be obvious from inspection of figure 16.1 that including the limit observations as though they were ordinary observations also creates bias. The solution to this dilemma is to employ maximum likelihood estimation.

The likelihood consists of the product of expressions for the "probability" of obtaining each observation. For each non-limit observation this expression is just the height of the appropriate density function representing the probability of getting that particular observation. For each limit observation, however, all we know is that the actual observation is above k. The probability for a limit observation therefore must be the probability of getting an observation above k, which would be the integral above k of the appropriate density function. In Tobin's original article (Tobin, 1958) durable goods purchases as a fraction of disposable income were modeled as a function of age and the ratio of liquid assets to disposable income. There were several limit observations at zero, corresponding to people who bought no durable goods, which entered the likelihood function as integrals from minus infinity to zero. The bottom line here is that the likelihood function becomes a mixture of densities and cumulative densities; fortunately, modern computer packages handle this with ease.

This estimation procedure for the Tobit model applies to the case of censored data. If the data are truncated, so that for example the limit observations are missing completely, the Tobit model no longer applies and an alternative maximum likelihood estimation procedure must be employed, described in the technical notes.

16.3 Sample Selection

The Tobit model is a special case of a more general model incorporating what is called *sample selection*. In these models there is a second equation, called the selection equation, which determines whether an observation makes it into the sample. This causes the sample to be non-random, drawn from a subpopulation of a wider population. For example, observations on hours worked are available only on those for whom their wage exceeds their reservation wage. The main problem here is that often the researcher wishes to draw conclusions about the wider population, not just the subpopulation from which the data is taken. If this is the case, to avoid *sample selection bias* estimation must take the sample selection phenomenon into account.

In the Tobit model, the sample selection equation is the same as the equation being estimated, with a fixed, known limit determining what observations get into the sample. Many cases do not fit this sample mold. For example, the decision to purchase a consumer durable may in part depend on whether desired expenditure exceeds a threshold value equal to the cost of the cheapest acceptable durable available. This threshold value will be unique to each individual, depending on each individual's characteristics, and will incorporate a random error. In this case the limit is unknown, varies from person to person, and is stochastic.

Unlike the Tobit model, these extended models have likelihood functions that are difficult to derive and are not always found in push-button form in econometrics packages. Consequently, practitioners are eager to find a practical alternative to maximum likelihood. The Heckman two-step estimation procedure, a second-best alternative to maximum likelihood, is very popular in this context.

The rationale of the Heckman method can be explained with the help of figure 16.1. Consider the value x0. For the corresponding y to be observed, the related error must be zero or negative, since if it were positive y would exceed k and would therefore be unobserved. This implies that for x0 the expected value of the error term is negative. Now consider values of x less than x0. For y to be observed the error can take on small positive values, in addition to being negative or zero, so

the expected value of the error becomes less negative. When x is greater than x0 the opposite occurs. As x becomes larger and larger, for y to be observed the error must lie below a larger and larger negative number. The expected value of the error term becomes more and more negative, as shown in the bottom half of figure 16.1.

The implication of this is that the error term is correlated with the explanatory variable, causing bias even asymptotically. If the expected value of the error term were known it could be included in the regression as an extra explanatory variable, removing that part of the error which is correlated with the explanatory variables and thereby avoiding the bias. The first stage of the Heckman procedure estimates the expected value of the error and the second stage reruns the regression with the estimated expected error as an extra explanatory variable. The details of finding estimates of the expected value of the error term are explained in the technical notes. It requires observations on the explanatory variables for the limit observations, so the Heckman procedure only works with censored data.

16.4 Duration Models

Economic analysis often focuses on the length of time a person or firm stays in a specific state before leaving that state. A popular example is the state of unemployment - what determines the duration of unemployment spells? Duration models are used to investigate empirically this issue.

Typically the data available for duration analysis consists of two types of observations. For the first type, the length of the unemployment spell is known (an individual found work after five weeks, for example). For the second type, the length of the unemployment spell is unknown because at the time of gathering data the individual was in the middle of an unemployment spell (an individual was still looking for work after five weeks, for example). In the latter case the observations are censored, implying that an estimation technique similar to that used for limited dependent variables should be employed.

Models in this context are formalized by specifying a probability density function for the duration of the unemployment spell. This is a function

of time t (measured from when the individual first became unemployed) providing the "probability" that an unemployment spell will be of length/duration t. Explanatory variables such as age, education, gender, and unemployment insurance eligibility, are included in this formula as well, to incorporate additional determinants of this probability. Maximum likelihood estimation can be used. The likelihood ingredient for each completed unemployment spell in the data is given by this duration density formula. The likelihood ingredient for each uncompleted unemployment spell in the data is given by an appropriate cumulation of this duration density giving the probability of getting an observation at least as great as the observed uncompleted spell. Thus the likelihood function becomes a mixture of densities and cumulative densities, just as in the Tobit analysis earlier.

Although the duration density function introduced above is the essential ingredient in duration models in that it is used to produce the likelihood function, discussion of duration models usually is undertaken in terms of a different function, the hazard function. This function gives the probability of leaving unemployment at time t given that the unemployment spell has lasted to time t; it is a conditional rather than an unconditional density function. The hazard function is the basis for discussion because it is usually the phenomenon of most interest to economists: What is the probability that someone who is unemployed will leave that state during this week?

The hazard function can be derived mathematically from the duration density function, so introduction of the hazard function does not change the nature of the model. But because interest and economic theory focus on the hazard function, it makes sense to choose a duration density specification that produces a hazard function that behaves as we believe it should. This explains why the duration densities used in duration models do not take a familiar form such as the normal distribution - they must be chosen so as to produce suitable hazard functions.

Some special cases of hazard functions are illustrated in Figure 16.2. The flat hazard, associated with the exponential duration density, says that the probability of leaving the unemployment state is the same, no matter how long one has been unemployed. The rising and falling hazards, associated with Weibull duration densities (with different Weibull parameter values giving rise to these two different hazards), says that the probability of leaving the unemployment state

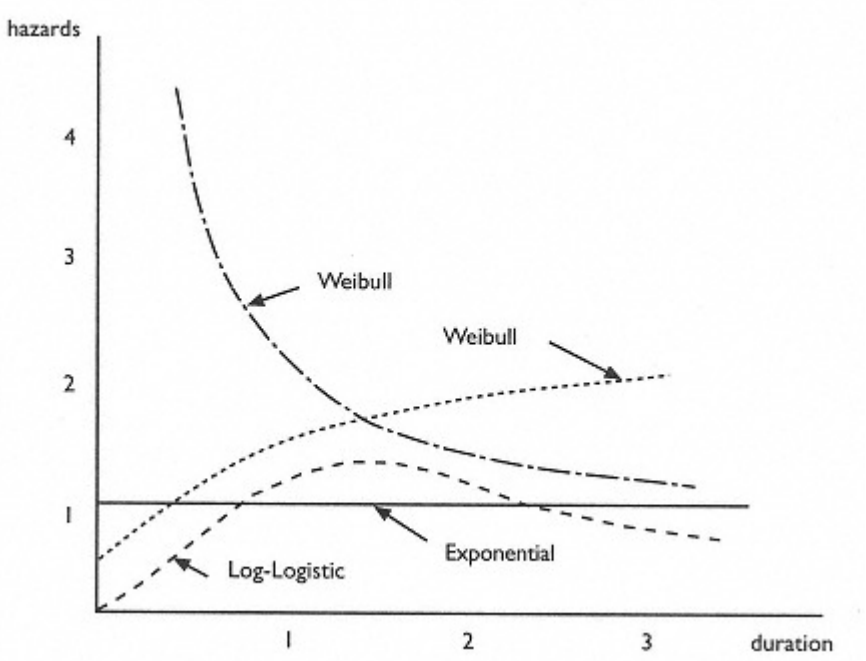

Figure 16.2
Examples of hazard functions associated with different duration densities

increases or decreases, respectively, as the unemployment spell lengthens. The hazard associated with the log-logistic duration density at first rises and then falls.

Explanatory variables such as age and gender enter by affecting the level and/or shape of these basic hazard functions. Estimation is simplified if a change in an explanatory variable simply shifts the basic hazard up or down. As explained in the technical notes, this produces what is called a proportional hazards model.

General Notes

16.1 Introduction

Maddala (1983) is an extensive reference on limited dependent variables and modeling options. Amemiya (1984) is a classic survey article. LIMDEP is the software of choice for estimation. Limited dependent variable modeling is prominent in the analysis of disequilibrium and switching phenomena; Maddala (1986) is a survey.

A major problem with limited dependent variable models is that estimation is quite sensitive to specification errors such as omission of a relevant explanatory variable

(even if orthogonal), heteroskedasticity, and non-normal errors. Maddala (1995) is a survey of specification tests in this context. Pagan and Vella (1989) have advocated use of conditional moment tests in this context; Greene (1997, pp. 972-4) is a good textbook exposition. Selection bias can be tested by performing the Heckman two-stage procedure and testing against zero the coefficient of the expected error term. Greene (1997, p. 970) exposits a test for Tobit versus the more general model in which a second equation determines whether y is observed. Volume 34 (1,2) of the *Journal of Econometrics* is devoted to specification tests in limited dependent variable models. Because of this sensitivity to specification errors, attention has focused recently on the development of robust estimators for this context. (Robust estimation is discussed in chapter 19.) Volume 32 (1) of the *Journal of Econometrics* is devoted to robust methods for limited dependent variables.

Heteroskedasticity of known form can be dealt with by building it into the likelihood function. Izadi (1992) suggests dealing with non-normal errors by assuming the errors come from the Pearson family of distributions of which the normal is a special case. These solutions require dealing with awkward likelihood functions, some of which are programmed into LIMDEP. Greene (1997, pp. 968-9) shows how an LM test can avoid this difficulty.

16.2 The Tobit Model

The Tobit model was introduced by Tobin (1958) to model a limit of zero expenditure. Garcia and Labeaga (1996) survey alternative approaches to modeling zero expenditures. Veall and Zimmermann (1996) survey goodness-of-fit measures (pseu-do-R2s) for Tobit and

duration models. Lankford and Wyckoff (1991) show how the Tobit model can be generalized to incorporate a Box-Cox functional form. Greene (1981) finds that the Tobit maximum likelihood estimates can be approximated quite well by dividing the OLS estimates by the proportion of nonlimit observations in the sample.

The estimated coefficients from censored and truncated models must be interpreted with care. Suppose we are estimating an equation explaining desired expenditure but that whenever it is negative we observe zero expenditure. McDonald and Moffit (1980) show that although the expected change in desired expenditure due to a unit change in an explanatory variable is the coefficient of that explanatory variable, the expected change in actual expenditure is not; for the latter the required calculation must account for the probability of being above the limit and changes therein. To be specific, they show that the expected actual change is the change in expected expenditure of those above the limit times the probability of being above the limit, plus the expected expenditure of those above the limit times the change in the probability of being above the limit. Note that this illustrates how Tobit contains the elements of regression (expected expenditure, and changes therein, of those above the limit) and the elements of probit (the probability, and changes therein, of being above the limit). They discuss and illustrate the implications of this for the use and interpretation of results of studies employing this type of model. For example, we could be interested in how much of the work disincentive of a negative income tax takes the form of a reduction in the probability of working versus a reduction in hours worked. In other cases, however, interest may focus on the untruncated population, in which case the

Tobit coefficients themselves are the relevant results since the Tobit index reflects the underlying population.

16.3 Sample Selection

The Heckman two-stage estimator was introduced in Heckman (1976). It is inferior to maximum likelihood because although it is consistent it is inefficient. Further, in "solving" the omitted variable problem the Heckman procedure introduces a measurement error problem, since an estimate of the expected value of the error term is employed in the second stage. In small samples it is not clear that the Heckman

procedure is to be recommended. Monte Carlo studies such as Stolzenberg and Relles (1990), Hartman (1991), Zuehlke and Zeman (1991) and Nawata (1993) find that on a MSE criterion, relative to subsample OLS the Heckman procedure does not perform well when the errors are not distributed normally, the sample size is small, the amount of censoring is small, the correlation between the errors of the regression and selection equations is small, and the degree of collinearity between the explanatory variables in the regression and selection equations is high. It appears that the Heckman procedure can often do more harm than good, and that subsample OLS is surprisingly efficient, and more robust to non-normality. Nawata (1994) and Nawata and Hagase (1996) recommend using maximum likelihood, and discuss computational considerations.

Limited dependent variable models can arise in a variety of forms. Suppose for example that we have

$$y = \alpha + \beta x + \varepsilon$$
$$p = \gamma + \delta z + u$$

with y being observed only if $y \geq p$. The likelihood function for this model is discussed by Maddala (1983, pp. 174-7). For example, suppose y represents wages of females and p represents the reservation wage. Consider individuals with high e values, so that their actual wage happens to be particularly high. Their reservation wage is more likely to be exceeded and such people are likely to be employed. Individuals with low e values, on the other hand, are more likely to have actual wage below reservation wage and such people are likely not to be employed. Thus using a sample of employed women to estimate the wage function will contain a disproportionate number of observations with high e values, biasing the estimators.

The preceding example is only one of several possible variants. One alternative is to specify that instead of y being observed when $y \geq p$, it is observed when $p \geq 0$. Bias arises in this case from the fact that often the two errors, e and u, are correlated; see Maddala (1983, p. 231) for the likelihood function for this case. Suppose, for example, that y represents earnings and p represents the decision to emigrate. There may be an unobservable element of u, call it "energy," that also affects earnings, i.e. energy is also an element of e, so that u and e are correlated. Immigrants as a group will have a disproportionate number of people with high energy, so using observations on immigrants to estimate the earnings function creates biased estimators of the earnings

function relevant to the population at large, or relevant to the population of the country from which they emigrated.

An important variant of this last example is a context in which a researcher is interest-

ed in the impact of a treatment or program of some kind. Greene (1993, pp. 713-14) has a good example of an equation determining earnings as a function of several explanatory variables plus a dummy representing whether or not an individual has a college education. An estimation problem arises because individuals self-select themselves into the college education category on the basis of the expected benefit to them of a college education, biasing upward the coefficient estimate for this dummy. A selection equation must be recognized, with an error term correlated with the error term in the earnings equation.

16.4 Duration Models

Duration modeling goes by many different names in the literature. To biologists it is *survival* analysis because it was originally developed to analyze time until death. Engineers, interested in the breakdown of machines, call it *reliability* or *failure time* analysis. Sociologists refer to it as *event history analysis.* The literature in this area can be quite technical, a notable exception being Allison (1984). Kiefer (1988) and Lancaster (1990) are expositions aimed at economists, the latter quite advanced. Goldstein et al. (1989) review software; the econometrics package with the most extensive duration model estimating routines is LIMDEP.

The exposition earlier was couched in terms of a *continuous-time* analysis in which knowledge of the exact time of duration was available. Although this may be reasonable for some types of economic data, for example strike durations measured in days, often this knowledge is not available. Unemployment duration, for example, is frequently measured in weeks, with no knowledge of when during the week of departure a particular individual left the unemployment state. In this case all those leaving the unemployment state during that week are grouped into a single discrete-time measure. Whenever the length of time of these discrete units of measurement is relatively large, analysis is undertaken via a *discrete-time* duration model, sometimes called a

grouped-data duration model. For a variety of reasons, explained in the technical notes, estimation via a discrete-time duration model is a very attractive alternative to estimation using a continuous-time duration model, and so is becoming more and more the method of choice amongst economists.

Technical Notes

16.1 Introduction

The likelihood functions for censored and truncated samples are quite different. This can be illustrated with the help of figure 16.3, which graphs the density function of the error e from figure 16.1. Consider a particular value *x*3 of *x*. For *y*3 to be observable, e3 must lie to the left of *k* - a - b*x*3; for *y*3 unobservable, e3 must lie to the right of *k* - a - b*x*3. This result follows from the discussion of *E*e above.

Suppose first we have a censored sample. If *x*3 corresponds to an observable *y,* then there will be a specific e3 and the likelihood for that observation is given by *L*3 in figure 16.3, the height of the density function for e at e3. But if *x*3 corresponds to an unobservable (i.e., missing) value of *y,* we have no specific e3; all we know is that e3

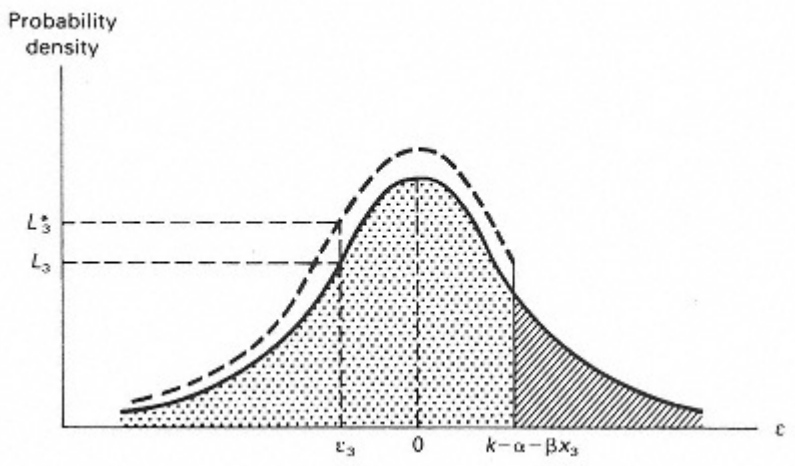

Figure 16.3
Explaining the likelihood for censored and truncated models

must lie to the right of *k* - a - b*x*3. The likelihood of this observation is thus the probability that e3 exceeds *k* - a - b*x*3, given by the lined area in figure 16.3, and calculated as 1 minus the density function cumulated to the point *k* - a - b*x*3. The likelihood for each observation in the sample may be calculated in one of these two ways, depending on whether the *y* value is observed or unobserved. Multiplying together all of these likelihood expressions, some of which are densities and some of which are cumulative densities, creates the likelihood for the censored sample.

Suppose now we have a truncated sample. For every possible value of *x*3 in the sample the associated error must come from the left of *k* - a - b*x*3 in figure 16.3. Consequently the lined area should not be viewed as part of the density of e3. Because of this, e3 can be viewed as being drawn from the truncated normal distribution given by the dashed curve in figure 16.3. This dashed curve is obtained by dividing the height of the original normal distribution by the dotted area, forcing the area under the dashed curve to equal 1. Thus the likelihood of the observation *y*3 is given in figure 16.3 by . Note that is a complicated function of the data, consisting of the height of the normal density function at the observation *(y*3, *x*3), divided by that density function cumulated to the point *k* - a - b*x*3. Each observation will give rise to a different dashed curve from which the likelihood of that observation can be calculated. Multiplying together all these likelihood expression creates the likelihood function for the entire sample.

16.3 Sample Selection

How does one go about estimating Ee to implement the Heckman two-step procedure? Consider once again the example of figure 16.1 as reflected in its supplementary graph figure 16.3. For any value *x*3 of *x,* the corresponding error term e3 for an observed *y*3 has in effect been drawn from the truncated normal distribution shown in

figure 16.3 as the dashed curve, cut off at the point k - a - bx3. Thus Ee is the expected value of this truncated normal distribution. A standard formula for the calculation of Ee can be used if it is known how many standard deviations k - a - bx3 represents. Estimation of $(k$ - a - bx3)/s, where s2 is the variance of the normal untrucated distribution, therefore allows estimation of Ee.

In a censored sample the data on y can be interpreted as dichotomous, with y taking the value 1 if observed and 0 if unobserved. Then a probit analysis can be done on these data, generating for x3, say, an estimate of the probability that y3 is observed. (Note: this cannot be done for a truncated sample, since the x values for the unobserved y values are also missing - this explains why the Heckman two-step method can be used only with censored samples.) Given an estimate of this probability, the dotted area in figure 16.3, it is easy to find the corresponding number of standard deviations of the standard normal giving rise to that probability, yielding the required estimate of $(k$ - a - bx3)/s.

The standard formula for the expected value of a truncated distribution is $E(e|e \leq a)$ = m + sl(q) where q is the number of standard deviations, $(a$ - m)/s, of a from the mean m of e, and l(q) is -f(q)/F(q), the inverse of the "Mills ratio," where f is the density function for the standard normal and f is its cumulative density function. Here m is zero and the estimate of $(k$ - a - bx3)/s is an estimate of q the inverse of the Mills ratio is estimated and used as an extra regressor, reducing the bias (and eliminating it asymptotically). For discussion of the interpretation of this extra regressor's coefficient estimate see Dolton and Makepeace (1987).

For this example, maximum likelihood estimation is not costly, so the two-step method is not used. However, the principles illustrated are employed to generate an estimate of the expected value of the error for more difficult cases such as the immigration example discussed earlier in the general notes to section 16.3. In this example the expected value of the error e in the earnings equation is nonzero because it is correlated with the error u that determines the decision to emigrate. The expected value of e is rsl(q) where r is the correlation between e and u. Consequently, when the inverse Mills ratio l(q) is added as a regressor for the second step of the Heckman method, its coefficient estimator estimates rs.

16.4 Duration Models

Often the first step in undertaking estimation in a continuous-time duration model is to plot a preliminary version of the hazard function by calculating the fraction of observations leaving the unemployment state during successive discrete time intervals. The measure for the fifth week, for example, is the number of observations leaving unemployment during the fifth week divided by the number of observations which could have left unemployment during that week. This picture can sometimes help determine the basic shape of the hazard function, facilitating the development of an appropriate specification. The two most popular ways of calculating this preliminary sketch of the hazard (or related survivor) function are called the life table and Kaplan-Meier methods.

A popular continuous-time duration model specification is the *proportional hazards* model. In this model the hazard function is composed of two separate parts, multiplied together. The first part is exclusively a function of duration time. It is called the *baseline hazard* and is usually written as l0(t). The second part is a function of explanatory

variables other than time and is traditionally chosen to take the form exp(x'b) where x is a vector of observations on an individual's characteristics (which may vary with time) and b is a parameter vector. The hazard function is then written as

$$\lambda(t) = \lambda_0(t)\exp(x'\beta)$$

The key thing is that time itself is separated from the explanatory variables so that the hazard is obtained simply by shifting the baseline hazard as the explanatory variables change (i.e., for all individuals the hazard function is proportional to the baseline hazard function). The reason for its popularity is that estimation can be undertaken by maximizing a much simpler function, called the "partial likelihood," instead of the full likelihood, with little loss in estimation efficiency. Furthermore, it happens that the baseline hazard cancels out of the partial likelihood formula, so that this estimation method has the tremendous advantage of being insensitive to the specification of the

baseline hazard. This advantage is offset by the fact that the baseline hazard and thus the full hazard function is not estimated. This disadvantage is not of consequence if interest focuses exclusively on the influence of the explanatory variables, as it often does.

Two ways of testing for the appropriateness of the proportional hazard model are popular. First, different categories of the explanatory variables should give rise to hazard functions that are proportional, so plotting an estimated hazard function for males, say, should produce a function roughly parallel to the estimated hazard function for females. In the second method, an extra explanatory variable, measured as an interaction of time with one of the existing explanatory variables, is added to the specification. Upon estimation this variable should have an estimated coefficient insignificantly different from zero if the proportional hazards specification is correct. An LR test can be used.

To obtain a sense of the algebra of continuous-time duration models, suppose f(t) is the duration density, reflecting the probability that a spell of unemployment has duration t. The hazard function is then l(t) = f(t)/[1 - F(t)] where F(t) is the cumulative density of t. The expression [1 - F(t)] is called the survivor function since it gives the probability of an individual surviving in the unemployment state at least to time t. Each observation on a completed spell is entered into the likelihood function as f(t) and each observation on an uncompleted spell is entered as [1 - F(t)].

A popular density function to use for f(t) is the exponential f(t) = de-dt where the parameter d is greater than zero. For this case F(t) = 1 - e-dt and the hazard function is a constant l(t) = d. Other distributions for f(t) give rise to hazard functions that are functions of t. For example, the Weibull distribution is a generalization of the exponential with

$$f(t) = \gamma\alpha t^{\alpha-1}\exp(-\gamma t^{\alpha})$$

and corresponding hazard

$$\lambda(t) = \gamma\alpha t^{\alpha-1}$$

where the two Weibull parameters g and a are positive. Note that if a = 1 this distrib-

ution becomes the exponential. If a < 1 the hazard function is increasing, and if a < 1 it is decreasing. These were illustrated in figure 16.2.

Explanatory variables are incorporated into duration models by specifying how they affect the hazard function, usually introduced in ways that are computationally tractable. For the exponential distribution, for example, the parameter d is modeled as ex'b. Since d-1 in this model is the mean duration, this is specifying that the mean duration is determined by explanatory variables according to the formula e-x'b. In the likelihood function d is replaced by e-x'b and maximization is done with respect to the b vector.

Duration models assume that individuals with identical values of the explanatory variables have exactly the same probability of leaving the state of unemployment, in the same way that probit and logit models assume that probabilities are deterministic. But we know that observationally similar people differ because of unobserved characteristics or just plain randomness; this is the reason why specifications of behavior in OLS regressions include an error term. This unobserved difference among individuals causes problems for duration models. Suppose there are two types of people with an unobservable difference in their "spunk." Those with a lot of spunk are very active in seeking a job and so spend less time in the unemployment state than those with less spunk. Consequently, over time those with less spunk come to be over-represented in the set of those still unemployed, biasing downward the hazard function. This *unobserved heterogeneity* problem is addressed by adding a multiplicative error term with mean unity to the hazard function, complicating still further the likelihood function (this error must be integrated out of the likelihood expression). A computationally tractable, and thus frequently employed density for this error is the gamma density. Heckman and Singer (1984) contend that a discrete error distribution with only a few possible values for this error works well and facilitates computation.

Estimation in a discrete-time model is much simpler because a complicated likelihood maximization problem is replaced with a familiar logit estimation problem for which standard software programs are available. This is accomplished by viewing each individual as contributing not one but several observations to a giant logit likelihood function. In the first time period each individual either stays or leaves

the state of unemployment, so a logit likelihood could be structured, with appropriate explanatory variables, to capture this. Now consider all the individuals who have not yet left the unemployment state and who have not become censored, namely all the individuals for whom it is possible to leave the unemployment state during the second time period. In the second time period each of these individuals either stays or leaves the state of unemployment, so a second logit likelihood, with the same explanatory variables (whose values could be different if they vary with time), can be structured to capture this. Similar logit likelihoods can be formulated for each of the remaining time periods, with the number of observations contributing to these likelihoods diminishing as individuals are censored or leave the unemployment state. A giant likelihood can then be formed by multiplying together all these separate-period likelihoods. Each individual contributes several terms to this giant likelihood, one term for each time period for which that individual was at risk of leaving the unemployment state.

A baseline hazard can be built into this specification by including a function of time among the explanatory variables. Alternatively, we could allow the intercept in each of the separate-period logit formulations to be different. If there are a total of k time periods, k dummy variables, one for each period (taking the value one for that period

and zero for all other periods) are entered as additional explanatory variables in the logit specification in place of the intercept. These dummy variables allow each duration length to contribute to the intercept of the logit specification separately, thereby modeling a completely unrestricted baseline hazard.

This discrete-time estimation procedure for duration models has become popular for several reasons. First, although most economic decisions are not made at discrete times, the data we have available usually report events as having occurred during some discrete time period rather than at a specific time. Second, the partial likelihood approach becomes quite difficult whenever more than one observation experiences the event during a measurement period, a common phenomenon in economic data. Third, it avoids having to deduce and program a complicated likelihood function. Not all specifications have software available for their estimation. Fourth, it permits an easy nonparametric way of

estimating the baseline hazard. And fifth, it provides a good approximation to continuous-time duration models. For a good economist-oriented exposition of discrete-time estimation see Jenkins (1995). This does not mean that more complicated maximum likelihood estimation is not employed; a popular proportional hazards approach that allows the baseline hazard to be flexible is that of Meyer (1990).

17
Time Series Econometrics

17.1 Introduction

Until not so long ago econometricians analyzed time series data in a way that was quite different from the methods employed by time series analysts (statisticians specializing in time series analysis). Econometricians tended to formulate a traditional regression model to represent the behavior of time series data, and to worry about things like simultaneity and autocorrelated errors, paying little attention to the specification of the dynamic structure of the time series. Furthermore, they assumed that the fact that most time series economic data are "non-stationary" (because they grow over time and so do not have a fixed, "stationary" mean) did not affect their empirical analyses. Time series analysts, on the other hand, tended to ignore the role of econometric "explanatory variables," and modeled time series behavior in terms of a sophisticated extrapolation mechanism. They circumvented the stationarity problem by working with data that were differenced a sufficient number of times to render them stationary.

Neither group paid much attention to the other until the appearance of two types of disquieting (for econometricians) studies. The first set of studies claimed that forecasts using the econometricians' methodology were inferior to those made using the time series analysts' approach; the second type claimed that running regressions on non-stationary data can give rise to misleading (or "spurious") values of R2, DW and t statistics, causing economists erroneously to conclude that a meaningful relationship exists among the regression variables. Although technically none of the CLR model assumptions is violated, inference using OLS is