Appendix 19.A

Log-Odds and Logit Models: Using Grouped Data

The linear probability model can yield both nonsensical fitted values within the range of sampled *X*'s and nonsensical predictions outside the range of sample *X*'s. This unattractive feature of the linear probability model led econometricians to turn to alternatives to probabilities, to measures of uncertainty that are not limited to the range zero to one. One early alternative to the linear probability model, the *log-odds model*, limits the range of outcomes to values between zero and one and sometimes allows a linear regression specification. *The strategy of the log-odds model is to devise an alternative measure of uncertainty, one that ranges from minus infinity to plus infinity, and to make that alternative measure the dependent variable in a linear regression.* This supplement introduces the alternative measure of uncertainty and shows how the log-odds model can be estimated by ordinary least squares (OLS) using data on outcomes of a binary variable for groups of individuals.

19.A.1 The Log-Odds Model for Grouped Data

The road to understanding the log-odds model begins at the betting window. Oddsmakers don't talk about the probability that a particular horse will win a given race. Instead, they speak about the odds: "The odds are 3–1 in ol' Stewball's favor"; "The odds are 2 to 3 against ol' Stewball." Odds of 3–1 in Stewball's favor claim that Stewball is expected to win 3 times in 4. In contrast, odds of 3 to 1 against Stewball claim Stewball is expected to lose 3 times in 4. Odds of 2 to 3 in favor of Stew-

ball claim that Stewball is expected to win 2 times in 5. In general, if the odds are p–q in Stewball's favor, the probability Stewball will win is $p/(p + q)$. Turned about, if π is the probability that the event will occur, then the odds *for* the event are $[\pi/(1 - \pi)] - 1$: If $\pi = .75$, we have .75/.25 for odds of 3–1; if $\pi = .4$, then we have .40/.60 for odds of $\frac{2}{3} - 1$, which is more often quoted as odds of 2 to 3.

Odds can range from zero (0/1) to infinity (1/0). Odds are a measure of uncertainty that can exceed one without difficulty. Odds do not suffer the restriction of probabilities that they not exceed one. However, like probabilities, odds do still suffer from always being non-negative. If we were to represent the odds O_i, given X_i, as

$$(O_i|X_i) = \frac{\pi(D_i|X_i)}{(1 - \pi(D_i|X_i))} = \beta_0 + \beta_1 X_i,$$

we would face the difficulty that for some values of X, the calculated odds would be negative, which cannot be true.

Econometricians finesse this difficulty by examining the natural logarithm of the odds, which we call the **log-odds**. Log-odds are a measure of uncertainty that ranges from minus infinity to plus infinity. When the odds are less than one, which implies a probability less than 50%, the logarithm of the odds will be negative. For odds greater than one, which implies a probability greater than 50%, the logarithm of the odds will be positive. For a probability of 50%, the odds are 1 and their logarithm is zero. The linear model applied to the log of the odds is called a **log-odds model**. The log-odds model assumes that in the population,

$$\log[odds_i|X_i] = \log\left[\frac{\pi_i|X_i}{(1 - \pi_i|X_i)}\right] = \beta_0 + \beta_1 X_i,$$

in which "log" refers to natural logarithms.

Figure 19.A.1 shows the shape of the probability of success for varying X's in the binomial log-odds model. The probability of success is bounded between zero and one, as must be true of probabilities. Notice that over some ranges of X, the application of a linear approximation could serve quite well, so that in some instances, the linear probability model can serve as a good approximation to the binomial log-odds model. Notice, too, that the log-odds model is not defined if the probability of success is zero or one, because then we are either taking the log of zero or dividing by zero, neither of which is possible.

Estimating the Log-Odds Model with Grouped Data

How are we to estimate the log-odds model? Sometimes, we can use least squares methods. Other times we cannot. We can often consistently estimate the log-odds model as an OLS regression if we have data on groups of individuals. Grouped data, therefore, provide a good starting place for studying the estimation of log-

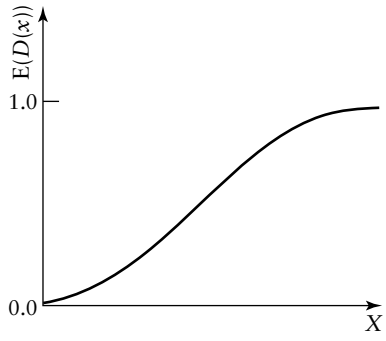

Figure 19.A.1
$E(D|X_i)$ in the Binomial Log-Odds Model.

odds models, because the use of OLS with such data connects these new models of binary outcomes with the linear models of earlier chapters.

Grouped data are data in which each observation is on a group of individuals. With grouped data, we can specify the log-odds model for the DGP as

$$observed\log[odds_i|X_i] = observed\log\left[\frac{p_i|X_i}{(1 - p_i|X_i)}\right] + \varepsilon_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \textbf{19.A.1}$$

in which p_i is the observed proportion of success ($D = 1$) in the i-th group. For example, if the first group contains 100 MBA program applicants with a GMAT of 580, and 23 of these 100 applicants were accepted into the program, $p_1 = 0.23$. Unfortunately, if a group has p equal to one or zero, Equation 19.A.1 is ill defined for that observation, because that would require taking the log of zero or dividing by zero, both of which are impossible. Standard practice is to exclude such groups from a sample when there are few such groups. When there are many such groups, econometricians forgo this approach to estimation and turn to the alternative estimation strategies described later in this chapter.

Because the model in Equation 19.A.1 requires that p_i be less than one and greater than zero, the model requires grouped data; it cannot be applied to observations on individuals. An individual is either employed or not, enters college or doesn't. For individuals, the observed probabilities, the p_i, are always one or zero—in which case the log-odds can't be computed. When we have data on many individuals, though, we can cluster the individuals into groups and estimate Equation 19.A.1 by OLS. However, with a small number of individuals, we are likely to have too many groups in which the observed proportion of successes is zero or one to make grouping feasible. Moreover, even with many individuals, grouping loses information about the variation in the X's if all the members of a

given group do not all have the same X-value. In practice, economists use the mean X for each group as X_i. When we use groups within which X varies, OLS estimation of Equation 19.A.1 is inefficient.

Economists study individual data on binary outcomes and their explanators far more often than they study grouped data. The parameters β_0 and β_1 of the log-odds model that we can consistently estimate with OLS if we have grouped data, we can also consistently estimate with individual data. However, we will learn that the log-odds model with individual data is a model nonlinear in its parameters, which rules out estimating β_0 and β_1 by least squares if we only have observations on individuals, instead of on groups.

Although individual data are more common in studies of binary outcomes, the grouped data model makes a good starting place for studying models of binary outcomes, because the grouped data specification of the log-odds model connects the linear regression models of earlier chapters with the nonlinear limited dependent variable models we need to study individual data on binary outcomes. The next section returns to the football game example that illustrated the linear probability model earlier and uses the grouped data log-odds model to re-estimate the relationship between the probability of the home team winning and the team's point spread.

An Example: The Grouped Log-Odds Model: NFL Point Spreads Revisited

The data set fteams.*** on the textbook Web site (**www.aw-bc.com/murray**) contains the number of home game wins for 30 of the 31 NFL teams in their eight home games of the 2001–2002 regular season. The groups in this data set are the 30 individual teams. The data set also contains the mean point spread for those same teams across their eight games. (The 31st team, the Carolina Panthers, lost all of its games and therefore can't be used in a grouped logit analysis.) The log-odds model for these grouped data is

$$\log\left[\frac{wins_i}{losses_i}\right] = \beta_0 + \beta_1(meanspread)_i + \varepsilon_i.$$

Table 19.A.1 reports the feasible generalized least squares (FGLS) estimates for the grouped log-odds model. Notice that the explanator in this case does not take on the same value for every observation on the group; consequently, the data only approximate the model in Equation 19.A.1, with the mean spread for a team across its games serving as X_i for that team.

In the log-odds model, if a home team (for example, your favorite team, when it plays at home) has a 50% chance of winning when its spread is zero, β_0 is zero, because the odds of winning in that case are 1 to 1, or just 1, and the log of 1 is

Table 19.A.1 What Point Spreads Say About the Probability of Winning in the NFL: II

Weighted least squares logit estimates for grouped data

Source	SS	df	MS
Model	4.85989767	1	4.85989767
Residual	16.5149358	28	.589819134
Total	21.3748334	29	.737063221

Number of obs = 30
F(1, 28) = 8.24
Prob > F = 0.0077
R–squared = 0.2274
Adj R–squared = 0.1998
Root MSE = .768

Logit	Coef.	Std. Err.	t	P > |t|	[95% Conf.	Interval]
meanspread	−.1021239	.0355773	−2.87	0.008	−.1750007	−.029247
_cons	−.0280463	.1605328	−0.17	0.863	−.3568829	−.3007903

zero. In Table 19.A.1, the estimated intercept, −0.028, is statistically insignificantly different from zero, so the estimated probability of your home team winning when its spread is zero is insignificantly different from the 50% probability estimated using the linear probability model.

According to the estimated log-odds model in Table 19.A.1, what was the effect of a one-point increase in the spread on your home team's chances of winning in 2001–2002? Answering this question for the log-odds model is more complicated than answering it for the linear probability model. In the linear probability model, the effect of a change in the spread on the probability of winning is equal to the coefficient on the spread variable. In contrast, because the log-odds model is nonlinear in the probability of winning, the effect of a change in the spread on the probability of winning varies with the level of the spread. To see this dependence, consider what the probability π is in the log-odds model.

In the log-odds model, the probability π is obtained from manipulating

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

as

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0 + \beta_1 X_i)$$

$$\pi_i = \exp(\beta_0 + \beta_1 X_i)(1-\pi_i) = \exp(\beta_0 + \beta_1 X_i) - \exp(\beta_0 + \beta_1 X_i)\pi_i$$

$$\pi_i + \exp(\beta_0 + \beta_1 X_i)\pi_i = \exp(\beta_0 + \beta_1 X_i)$$

$$\pi_i(1 + \exp(\beta_0 + \beta_1 X_i)) = \exp(\beta_0 + \beta_1 X_i),$$

so

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_i)}. \qquad \textbf{19.A.2}$$

Equation 19.A.2 highlights the nonlinear relationship between the probability of winning and X and between the probability of winning and β_1. In the log-odds model, the derivative of the probability with respect to X_i is not equal to β_1, as it would be in the linear probability model, but instead the derivative varies with both β_1 and X_i.

The mean spread in the 2001–2002 season was 5.88 for the home team—that is, on average the spread was against the home team. At a mean spread of 5.88, the estimated log odds based on Table 19.A.1 is

$$\text{estimated } \log[\pi/(1 - \pi)|X = 5.88] = -0.028 - .102 \cdot 5.88 = -0.63.$$

Thus, replacing the actual coefficients in Equation 19.A.2 by their estimated values yields

$$\hat{\pi}/(1 - \hat{\pi}) = e^{-0.63},$$

so

$$\hat{\pi} = \frac{e^{-0.63}}{1 + e^{-0.63}} = 0.348.$$

This probability is very close to the corresponding estimated probability in the linear model, 35.2%. Further from the mean spread, the linear model and the grouped log-odds model are less in accord. At a point spread of 20 points, the linear probability model assigns a senseless negative probability to the home team's winning. The grouped log-odds model estimates an 11% chance of winning for the home team if the spread were 20 points.

Finally, we return to the question, "What was the effect of a one-point increase in the spread on the home team's chances of winning in 2001–2002?" The log-odds model estimated in Table 19.A.1 implies that increasing the point spread from 5.88 (its mean) to 6.88 yields a new estimated probability of 32.5%. Thus, we estimate that a one-point increase in the spread from 5.88 to 6.88 lowers the probability of winning by 2.3 percentage points. This is in close accord with the linear probability model's estimate of 2.5% in Table 19.A.1.

The grouped data logit estimates are not as efficient as estimates based on individual data would be. The lost efficiency stems from averaging the X-values, the spreads, for each team in the sample. In an extreme example, suppose that the average spreads were almost identical for the 30 teams, but that each team's spread

varied considerably from game to game. The average spreads would be highly collinear with the intercept term, so the grouped data logit estimates of the spread's effect on the probability of winning would be very imprecise. In contrast, estimating the model with the observations on individual games would rely on spreads that vary markedly across games, which might yield precise estimates of the spread's effect. *Only when we put individual observations in groups that have identical X-values does grouping not lose information.*

19.A.2 Individual and Group Logits Are the Same

What Are an Estimator's Properties?

The individual logit model assumes that there exists an unobserved variable, Z, that determines the binary outcome of interest. In this model,

$$Z_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

and D_i equals 1 if Z_i exceeds 0. Alternatively, because someone with an expected value of Z equal to, say, 5 will be employed as long as ε_i isn't less than -5, D_i equals 1 if $\varepsilon_i > -(-\beta_0 + \beta_1 X_i) = (-\beta_0 - \beta_1 X_i)$. If we know the cumulative distribution function for ε_i, $F(w)$, we can ascertain the probability that ε_i takes on values less than or equal to w, because that is what $F(w)$ is. In turn, $[1 - F(w)]$ gives the probability that $\varepsilon_i > w$.

The cumulative distribution function, $F(w)$, that underlies the logit model is the logistic distribution, for which

$$F(w) = \frac{\exp(w)}{[1 + \exp(w)]},$$

and for which, therefore,

$$1 - F(w) = \frac{1}{[1 + \exp(w)]}.$$

To see that this is, indeed, the log-odds model in new guise, recall that D_i equals 1 if Z_i exceeds 0, that is, if $\varepsilon_i > (-\beta_0 - \beta_1 X_i)$. Thus, the probability that D_i equals 1 is

$$\Pr[\varepsilon_i > (-\beta_0 - \beta_1 X_i)] = 1 - \Pr[\varepsilon_i < (-\beta_0 - \beta_1 X_i)],$$

which is to say that $[1 - F(-\beta_0 - \beta_1 X_i)]$ is the probability that $D_i = 1$. Therefore, the odds for success are

$$odds = \frac{(\pi_i)}{(1 - \pi_i)} = \frac{[1 - F(w)]}{F(w)} = \frac{1}{\exp(w)} = \exp(-w),$$

with $w = -\beta_0 - \beta_1 X_i$, and the logarithm of the odds is:

$$\log(odds) = \log(\exp(-w)) = -w = \beta_0 + \beta_1 X_i,$$

which is the log-odds model of Equation 19.A.1.

Summary

The linear probability model can predict probabilities that lie outside the range of zero to one. To overcome this limitation of the linear probability model, this supplement introduces a new measure of uncertainty, the natural logarithm of an event's odds that ranges from minus to plus infinity. The log-odds model applies to grouped data. We can efficiently estimate the log-odds model with FGLS when we have grouped data. The log-odds model highlights the connection between the linear regression models of earlier chapters and the nonlinear models that econometricians use to model discrete choice behavior with data from individual choices. The supplement concludes by showing that the logit model that we often use for estimating binary choice models is equivalent to the log-odds model for grouped data.

Concepts for Review

Grouped data

Log-odds

Log-odds model