CHAPTER 15

Categorical Dependent Variables

15.0 What We Need to Know When We Finish This Chapter

1. **Section 15.1:** The values for *discrete dependent variables* represent categories rather than quantities. We try to predict the probability that

the outcome will be in a particular category, rather than the quantity of the outcome.

2. **Section 15.4:** The underlying continuous variable that we would like to observe is y_i^*. The categorical representation of this variable that we can actually measure is y_i. y_i serves as the equivalent of the "dependent variable" in our analysis here.
3. **Equation (15.20), section 15.4:** The true probability of observing the sample is

$$\text{P}(\text{observing the sample})=\left[\prod_{y_i=0}\text{P}\left(\frac{\varepsilon_i}{\sigma}<-\left(\frac{\alpha}{\sigma}+\frac{\beta}{\sigma}x_i\right)\right)\right]\left[\prod_{y_i=1}\text{P}\left(\frac{\varepsilon_i}{\sigma}\geq-\left(\frac{\alpha}{\sigma}+\frac{\beta}{\sigma}x_i\right)\right)\right].$$

   This probability depends on the population parameters β/σ and α/σ.
4. **Equation (15.21), section 15.4:** Our objective is to choose, based on the sample, an estimator b of β/σ and an estimator a of α/σ. With these choices, the *likelihood* is our estimate of the true probability of observing the sample:

$$\text{L}(y_1, y_2, \ldots, y_n, a, b)=\prod_{y_i=0}P\left(\frac{\varepsilon_i}{\sigma}<-(a+bx_i)\right)\prod_{y_i=1}P\left(\frac{\varepsilon_i}{\sigma}\geq-(a+bx_i)\right).$$

5. **Section 15.4:** Estimator values that maximize the likelihood function, such as b and a here, are *maximum likelihood (ML) estimators*. We obtain them through *maximum likelihood estimation*. When y_i is categorical and ε_i/σ is a standard normal random variable, the maximum likelihood procedure is known as *probit estimation*.
6. **Section 15.5:** In practice, we maximize the log of the likelihood function rather than the likelihood function itself. The derivatives of the log-likelihood function are usually complicated nonlinear functions of the estimators, so they can't be solved explicitly. Instead, we use an *iterative procedure* in which we choose a sequence of estimator values. This sequence systematically increases the value of the log-likelihood function until it converges to its maximum. The estimator values that accomplish this are our ML estimates.
7. **Equation (15.37), section 15.6:** The effect of a change in x_i on the probability of observing the outcome of interest depends on the initial value of x_i. For purposes of comparability, we usually calculate this

effect beginning at the average value of x_i, $\bar{x}$:

$$P\left(\frac{\varepsilon_i}{\sigma} \geq -\left(a+b\left(\bar{x}+\Delta x\right)\right)\right) - P\left(\frac{\varepsilon_i}{\sigma} \geq -\left(a+b\bar{x}\right)\right).$$

The effects of a change in x_i on this probability beginning from a different initial value will be in the same direction but of different magnitude.

8. **Section 15.6:** ML estimators are consistent. In addition, they are *best consistent*, meaning that they have the smallest variances among all consistent estimators. Their distribution is asymptotically normal, which means we can make confidence intervals and hypothesis tests with them as we did in chapter 7.
9. **Section 15.7:** The unobserved dependent variable y_i^* is often taken to represent a continuous *latent propensity* to engage in the activity under examination. The observed variable y_i is the expression of that propensity through the choice of whether or not to engage. Under this interpretation, probit analysis is applicable to many dependent variables that do not, at first, appear to have underlying continuous representations.
10. **Section 15.8:** *Selection problems* arise when the value of the disturbance term helps to determine whether the outcome is observed. When this happens, bias can occur because the disturbance may be correlated with the explanatory variables for those observations that can be included in the sample. A *sample selection model*, which estimates a probit equation predicting inclusion in the sample *simultaneously* with the regression of interest, may mitigate these biases. However, the probit must contain an effective explanatory variable that does not belong in the regression.
11. **Section 15.9:** When we take ε_i/σ as a logistic random variable rather than a standard normal random variable, the estimation procedure of sections 15.4 and 15.5 is called *logit estimation*. Ordinary least squares regression with the categorical y_i as the dependent variable is called the *linear probability model* but is generally not a good way to investigate the determinants of y_i.