2

Static and Dynamic Discrete Choice

The discrete-choice framework, often referred to as "qualitative response models," has become a major workhorse in diverse contexts of empirical industrial organization and other fields in applied microeconomics. In this chapter, we review the basic theory on the binary choice and multiple discrete-choice models. Then we proceed to study the dynamic discrete-choice models pioneered by Rust (1987); Hotz and Miller (1993), and Hotz et al. (1994) to study how the discrete-choice framework incorporates the forward-looking behavior of an economic agent. Throughout this chapter, we assume that individual choice data on a finite set of alternatives are available. We focus mostly on the fully parametric setup, of which the main goal often boils down to deriving the likelihood function for the maximum-likelihood estimation.

2.1 Binary Choice

2.1.1 Motivation: Linear Probability Model

In this section, we consider the binary choice data of individuals indexed by $i \in \mathcal{I} := \{1, 2, \ldots, I\}$, for alternatives indexed by $j \in \mathcal{J} := \left\{1, 2, \ldots, J\right\}$. Throughout this section, we assume that data for each consumer's choice on each alternative are available where the set of alternatives $\mathcal{J}$ is not exclusive. Let $y_{i,j}$ be a discrete outcome variable with only two possibilities, 0 and 1. If consumer i chooses to buy product j, $y_{i,j} = 1$; otherwise, $y_{i,j} = 0$. For notational convenience, our discussion focuses on a single individual i.

Consider the following linear estimation equation:

$$y_{i,j} = \delta_j + \eta_{i,j},$$

in which $\eta_{i,j}$ is an error term and $\delta_j = \mathbf{x}_j' \boldsymbol{\theta}$, where $\mathbf{x}_j$ is a vector of covariates that shifts the choice probability, observable to the econometrician.[1]

The ordinary least squares (OLS) estimator $\hat{\boldsymbol{\theta}}_{OLS}$ is consistent for $\boldsymbol{\theta}$ and asymptotically normal. The model is called the "linear probability model" because the prediction $\hat{y}_j$ is such that

$$\hat{y}_{i,j} = \hat{\mathbb{E}}\left[y_{i,j}|\mathbf{x}_j\right] = \hat{\Pr}\left(y_{i,j} = 1|\mathbf{x}_j\right) = \mathbf{x}_j' \hat{\boldsymbol{\theta}}_{OLS}.$$

Because this model has the prediction $\hat{y}_{i,j} = \mathbf{x}_j' \hat{\boldsymbol{\theta}}_{OLS}$, the OLS estimator provides us with an easy interpretation of the marginal effects: A one-unit increase in $x_j^{(l)}$ will increase the predicted conditional probability $\hat{\Pr}\left(y_{i,j} = 1|\mathbf{x}_j\right)$ by $\hat{\theta}_{OLS}^{(l)}$. The implied constant marginal effects follow from construction of the linear probability model. An immediate drawback of this approach is that the constant marginal effect assumption is likely invalid in any discrete-choice model. It yields a poor fit when $\hat{y}_{i,j}$ is close to 0 or 1, and it eventually leads the model to predict $\hat{y}_{i,j} > 1$ or $\hat{y}_{i,j} < 0$. It motivates the choice of $G(\cdot)$ to be a legitimate probability distribution with an unrestricted support, as we study in this chapter.[2]

2.1.2 Binary Logit and Binary Probit Model

Let us formalize the setup. Consider a latent utility $y_{i,j}^*$ specified by

$$y_{i,j}^* = \delta_j + \epsilon_{i,j} \qquad \epsilon_{i,j} \sim \text{i.i.d. } G(\cdot) \tag{2.1.1}$$

$$y_{i,j} = \begin{cases} 1 & \text{if } y_{i,j}^* \geq 0 \\ 0 & \text{if } y_{i,j}^* < 0 \end{cases},$$

where δ_j is the mean utility index that represents the observable component of $y_{i,j}^*$ to the econometrician; and $\epsilon_{i,j}$ represents the idiosyncratic shocks on the latent utility that are unobservable to the econometrician, which is fully known to the consumer. We assume that δ_j and $\epsilon_{i,j}$ are orthogonal.

1. Throughout chapter 2, we index the vector of characteristics $\mathbf{x}$ only by j, which implies that $\mathbf{x}$ can vary only with the alternatives, not with individuals. However, this restriction is purely for notational convenience. In principle, $\mathbf{x}$ can be indexed by both i and j.

2. We focus on the parametric methods in this section. Note that there are semiparametric index models that do not require a researcher to specify the shape of $G(\cdot)$. See, for example, Klein and Spady (1993) and Blundell and Powell (2004).

The two parametric specifications of $G(\cdot)$ used most often are standard Gaussian and logistic (0, 1). Assume that $g(.)$, the probability density function of $G(.)$, is symmetric around 0. This is indeed the case for the standard Gaussian and the logistic distribution with location parameter 0. Consider the following conditional probability:

$$\begin{aligned}
\Pr\left(y_{i,j}=1|\mathbf{x}_j\right) &= \Pr\left(\mathbf{x}_j'\boldsymbol{\theta}+\epsilon_{i,j}>0\right)\\
&= \Pr\left(\epsilon_{i,j}>-\mathbf{x}_j'\boldsymbol{\theta}\right)\\
&= 1-\Pr\left(\epsilon_{i,j}\le-\mathbf{x}_j'\boldsymbol{\theta}\right)\\
&= \Pr\left(\epsilon_{i,j}\le\mathbf{x}_j'\boldsymbol{\theta}\right)\\
&= G\left(\mathbf{x}_j'\boldsymbol{\theta}\right).
\end{aligned}$$

The next-to-last equality follows from $1-\Pr\left(\epsilon_{i,j}\le-\mathbf{x}_j'\boldsymbol{\theta}\right)=\Pr\left(\epsilon_{i,j}\le\mathbf{x}_j'\boldsymbol{\theta}\right)$ by the symmetry of $g(.)$. The Probit model assumes $\epsilon_{i,j}\sim$ i.i.d. $\mathcal{N}(0,1)$, and the logit model assumes $\epsilon_{i,j}\sim$ i.i.d. logistic (0, 1).[3] In both models, either the scale of $\boldsymbol{\theta}$ or $G(\cdot)$'s scale parameter σ cannot be identified. To see why, consider the following:

$$\begin{aligned}
\Pr\left(y_{i,j}=1|\mathbf{x}_j\right) &= \Pr\left(\frac{\epsilon_{i,j}}{\sigma}>-\frac{\mathbf{x}_j'\boldsymbol{\theta}}{\sigma}\right)\\
&= \Pr\left(\epsilon_{i,j}\le\mathbf{x}_j'\boldsymbol{\theta}\right).
\end{aligned}$$

So long as $\sigma>0$, any change in σ does not affect the choice probability. The convention is to set $\sigma=1$ rather than adjusting the scale of $\boldsymbol{\theta}$.

Fix i. Suppose that we have the observations $\left\{y_{i,j},\mathbf{x}_j\right\}_{j=1}^{J}$. The likelihood function of the binary choice models for individual i is

$$\begin{aligned}
L_i\left(\left\{y_{i,j},\mathbf{x}_j\right\}_{j=1}^{J}|\boldsymbol{\theta}\right) &= \prod_{j=1}^{J}\left[\Pr\left(y_{i,j}=1|\mathbf{x}_j\right)\right]^{\mathbf{1}(y_{i,j}=1)}\left[\Pr\left(y_{i,j}=0|\mathbf{x}_j\right)\right]^{\mathbf{1}(y_{i,j}=0)}\\
&= \prod_{j=1}^{J}\left[G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]^{y_{i,j}}\left[1-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]^{(1-y_{i,j})}. \qquad (2.1.2)
\end{aligned}$$

3. That is, $\Pr\left(y_{i,j}=1|\mathbf{x}_j\right)=G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)=\frac{\exp\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{1+\exp\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}$.

The log-likelihood function follows by taking the logarithm

$$l_i\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^{J} |\boldsymbol{\theta}\right) = \sum_{j=1}^{J}\left\{y_{i,j} \ln G\left(\mathbf{x}_j'\boldsymbol{\theta}\right) + \left(1 - y_{i,j}\right) \ln\left[1 - G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]\right\}. \quad (2.1.3)$$

Taking derivatives with respect to the parameter vector $\boldsymbol{\theta}$ on equation (2.1.3) yields the sum of the scores:

$$\begin{aligned}\mathbf{s}_i\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^{J} |\boldsymbol{\theta}\right) &= \sum_{j=1}^{J}\left\{\frac{y_{i,j}}{G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)} g\left(\mathbf{x}_j'\boldsymbol{\theta}\right) - \frac{1 - y_{i,j}}{1 - G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)} g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right\}\mathbf{x}_j \\ &= \sum_{j=1}^{J}\left\{\frac{g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\left[1 - G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]}\left[y_{i,j} - G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]\right\}\mathbf{x}_j.\end{aligned} \quad (2.1.4)$$

Setting equation (2.1.4) to $\mathbf{0}$ yields the first-order condition for the maximum-likelihood estimation for the binary choice models. The $\frac{g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\left[1-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]}$ term can be interpreted as a weighting function; and $\left[y_{i,j} - G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]$ is the prediction error, the expectation of which is zero.

In the logit model, the first-order condition $\mathbf{s}_i\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^{J} |\boldsymbol{\theta}\right) = \mathbf{0}$ simplifies further, using the fact that $\forall z \in \mathbb{R}$, $\frac{g(z)}{G(z)[1-G(z)]} = 1$.[4] Combining the fact with

4. For a logistic probability density function $g(\cdot)$ with location parameter 0 and scale parameter 1, the following holds:

$$\begin{aligned}g(z) &= \frac{\exp(z)}{1+\exp(z)} - \frac{\exp(2z)}{[1+\exp(z)]^2} \\ &= \frac{[1+\exp(z)]\exp(z) - \exp(2z)}{[1+\exp(z)]^2} \\ &= \frac{\exp(z) + \exp(2z) - \exp(2z)}{[1+\exp(z)]^2} \\ &= \frac{\exp(z)}{[1+\exp(z)]^2}\end{aligned}$$

and

$$G(z)[1 - G(z)] = \frac{\exp(z)}{1+\exp(z)}\left[\frac{1}{1+\exp(z)}\right].$$

Taking the ratio yields $\frac{g(z)}{G(z)[1-G(z)]} = 1$.

equation (2.1.4), the first-order condition for the maximum likelihood problem with the first-order condition simplifies as

$$\mathbf{s}_i\left(\left\{y_{i,j},\mathbf{x}_j\right\}_{j=1}^{J}|\boldsymbol{\theta}\right)=\mathbf{0}$$
$$\Rightarrow\sum_{j=1}^{J}\left[y_{i,j}-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]\mathbf{x}_j=\mathbf{0}.$$

If $\mathbf{x}_j$ contains 1 in its row, the first-order condition also contains $\bar{y}=\overline{G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}$.

2.1.3 Marginal Effects

The marginal effect of the binary choice model, $\frac{\partial\Pr\left(y_{i,j}=1|\mathbf{x}_j\right)}{\partial\mathbf{x}_j^{(l)}}$, is

$$\begin{aligned}\frac{\partial\Pr\left(y_{i,j}=1|\mathbf{x}_j\right)}{\partial x_j^{(l)}}&=\frac{\partial G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{\partial x_j^{(l)}}\\&=g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\theta^{(l)}.\end{aligned}\tag{2.1.5}$$

Unlike the linear probability model, the marginal effect varies across observations. Heterogeneity in responses exists in this model because of the nonlinearity of $G(\cdot)$. One may report equation (2.1.5) for each observation j in principle. Alternatively, one can consider either (1) the average marginal effect $\frac{1}{J}\sum_{j=1}^{J}g\left(\mathbf{x}_j'\hat{\boldsymbol{\theta}}\right)\hat{\theta}^{(l)}$ or (2) the marginal effect on average (or median) observation $g\left(\bar{\mathbf{x}}'\hat{\boldsymbol{\theta}}\right)\hat{\theta}^{(l)}$. It is acceptable to report either (1) or (2) as the summary measure of marginal effects; the researcher must be transparent about which summary measure is being reported.

2.2 Multiple Choice: Random Utility Maximization Framework

To model a discrete choice over multiple alternatives, we introduce the simple logit model and the nested logit model developed in a series of works by McFadden, (1974, 1978, 1981) and McFadden and Train (2000), among others. The random utility maximization (RUM) framework is the major workhorse in