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

$$\frac{\partial \Pr\left(y_{i,j}=1|\mathbf{x}_j\right)}{\partial x_j^{(l)}}=\frac{\partial G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{\partial x_j^{(l)}}$$

$$=g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\theta^{(l)}. \tag{2.1.5}$$

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

diverse contexts of applied microeconomics when multiple mutually exclusive alternatives exist. A common way to derive the logistic choice probabilities is to begin from the additive i.i.d. type I extreme-value-distributed idiosyncratic utility shocks. We present some preliminary results on type I extreme value distribution in section 2.2.1, and then present our main results in the subsections that follow.

2.2.1 Preliminary Results: Type I Extreme Value Distribution and Its Properties

Definition. *(Type I Extreme Value Distribution)* $\epsilon_i \sim T1EV(\alpha)$ if ϵ_i follows the cumulative distribution function

$$\begin{aligned}\Pr(\epsilon_i \leq \epsilon) &= F_\alpha(\epsilon) \\ &= \exp[-\exp[-(\epsilon-\alpha)]].\end{aligned}$$

Note that this distribution is also referred to as a "Gumbel distribution" or "double exponential distribution."[5] When $\alpha = 0$, the expectation of a type I extreme value random variable is the Euler-Mascheroni constant $\gamma \approx 0.5772$. Note that throughout this book, we will take a location shift by $-\gamma \approx -0.5772$ when it represents an econometric error term in order to make it a mean-zero random variable.

Lemma 2.2.1. *(Density Function of Type I Extreme Value Distribution)* Let $F_\alpha(\epsilon)$ be the cumulative distribution function of $T1EV(\alpha)$. Then the probability density function $f_\alpha(\epsilon) = \exp(\alpha-\epsilon) F_\alpha(\epsilon)$.

Lemma 2.2.2. *(Distribution of Maximum over Independently Distributed T1EV Random Variables)* Let $\epsilon_{i,j} \sim T1EV(\alpha_j)$, where $\epsilon_{i,j}$ are independent over j. Let $\alpha = \ln\left[\sum_{j=1}^{J} \exp(\alpha_j)\right]$. Then,

$$\max_j \{\epsilon_{i,j}\} \sim T1EV(\alpha).$$

5. In principle, type I extreme value distribution is a two-parameter distribution, location, and scale. If the scale parameter is denoted by σ, then the cumulative distribution function would be $\Pr(\epsilon_i \leq \epsilon) = \exp[-\exp[-(\epsilon-\alpha)/\sigma]]$. We normalize the scale parameter to 1 because it cannot be identified in general.

Proof. Let $\epsilon \in \mathbb{R}$. We have

$$\begin{aligned}
\Pr\left(\max_{j\in\mathcal{J}}\left\{\epsilon_{i,j}\right\} \leq \epsilon\right) &= \prod_{j=1}^{J} \Pr\left(\epsilon_{i,j} \leq \epsilon\right) \\
&= \prod_{j=1}^{J} \exp\left[-\exp\left[-\left(\epsilon - \alpha_j\right)\right]\right] \\
&= \exp\left[-\sum_{j=1}^{J} \exp\left[-\left(\epsilon - \alpha_j\right)\right]\right] \\
&= \exp\left[-\exp\left(-\epsilon\right)\sum_{j=1}^{J} \exp\left(\alpha_j\right)\right] \\
&= \exp\left[-\exp\left(-\epsilon\right)\exp\left(\alpha\right)\right] \\
&= F_{\alpha}\left(\epsilon\right).
\end{aligned}$$

The first equality follows from the maximum order statistic for a sample size of J. □

Corollary. *(Expectation of Maximum over T1EV Random Variables)* Let $j \in \mathcal{J}$. Let $\epsilon_{i,j} \sim T1EV\left(0\right)$. Let $u_{i,j} = \delta_j + \epsilon_{i,j}$ where δ_j is the additive deterministic component of choice j. Then,

$$\mathbb{E}\left[\max_{j\in\mathcal{J}}\left\{u_{i,j}\right\}\right] = \ln\left[\sum_{j\in\mathcal{J}} \exp\left(\delta_j\right)\right] + \gamma,$$

where the Euler-Mascheroni constant $\gamma \approx 0.5772$.

Lemma 2.2.3. *(Subtraction of Two Independent T1EV Random Variables)* Suppose that $u_{i,j} \sim T1EV\left(\delta_j\right)$ and $u_{i,k} \sim T1EV\left(\delta_k\right)$ where $u_{i,j} \perp\!\!\!\perp u_{i,k}$. Then, $u_{i,j} - u_{i,k} \sim \text{Logistic}\left(\delta_j - \delta_k\right)$.

Proof. Let $F\left(u_j\right)$ be the cumulative distribution function of $T1EV\left(\delta_j\right)$, and let $f_{\delta_j}\left(u_j\right)$ be the corresponding probability density function. By lemma 2.2.1,

$$\begin{aligned}
f_{\delta_k}\left(u_k\right) &= \exp\left[-\left(u_k - \delta_k\right)\right] F_{\delta_k}\left(u_k\right) \\
&= \exp\left[-\left(u_k - \delta_k\right)\right] \exp\left[-\exp\left[-\left(u_k - \delta_k\right)\right]\right].
\end{aligned}$$

Thus,

$$
\begin{aligned}
&\Pr\left(u_j - u_k < u\right) \\
&\quad = \Pr\left(u_j < u_k + u\right) \\
&\quad = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{u_k+u} f_{\delta_j}\left(u_j\right) du_j \right\} f_{\delta_k}\left(u_k\right) du_k \\
&\quad = \int_{-\infty}^{\infty} f_{\delta_k}\left(u_k\right) F_{\delta_j}\left(u_k + u\right) du_k \\
&\quad = \int_{-\infty}^{\infty} \exp\left[-\left(u_k - \delta_k\right)\right] F_{\delta_k}\left(u_k\right) F_{\delta_j}\left(u_k + u\right) du_k \\
&\quad = \int_{-\infty}^{\infty} \exp\left[-\left(u_k - \delta_k\right)\right] \exp\left[-\exp\left[-\left(u_k - \delta_k\right)\right]\right] \\
&\qquad \exp\left[-\exp\left[-\left(u_k + u - \delta_j\right)\right]\right] du_k \\
&\quad = \int_{-\infty}^{\infty} \exp\left[-\left(u_k - \delta_k\right)\right] \exp\left[-\exp\left[-\left(u_k - \delta_k\right)\right]\right. \\
&\qquad \left. - \exp\left[-\left(u_k + u - \delta_j\right)\right]\right] du_k \\
&\quad = \int_{-\infty}^{\infty} \exp\left[-\left(u_k - \delta_k\right)\right] \exp\left[-\exp\left[-\left(u_k - \delta_k\right)\right]\right. \\
&\qquad \left. - \exp\left[-\left(u_k - \delta_k + u + \delta_k - \delta_j\right)\right]\right] du_k \\
&\quad = \int_{-\infty}^{\infty} \exp\left[-\left(u_k - \delta_k\right)\right] \exp\left[-\exp\left[-\left(u_k - \delta_k\right)\right]\right. \\
&\qquad \left. \left\{1 + \exp\left[-\left(u + \delta_k - \delta_j\right)\right]\right\}\right] du_k.
\end{aligned}
$$

Denote $a := \left\{1 + \exp\left[-\left(u + \delta_k - \delta_j\right)\right]\right\}$ for notational simplicity:

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \exp\left[-\left(u_k - \delta_k\right)\right] \exp\left[-a \exp\left[-\left(u_k - \delta_k\right)\right]\right] du_k \\
&\quad = \frac{1}{a} \int_{-\infty}^{\infty} a \exp\left[-\left(u_k - \delta_k\right)\right] \exp\left[-a \exp\left[-\left(u_k - \delta_k\right)\right]\right] du_k \\
&\quad = \frac{1}{1 + \exp\left[-\left(u + \delta_k - \delta_j\right)\right]} \\
&\quad = \frac{\exp\left(u - \left(\delta_j - \delta_k\right)\right)}{\exp\left(u - \left(\delta_j - \delta_k\right)\right) + 1},
\end{aligned}
\tag{2.2.1}
$$

which is a logistic cumulative distribution function with mean $\left(\delta_j - \delta_k\right)$. Note that equation (2.2.1) follows because

$$\exp\left[-a\left(u_k - \delta_k\right)\right]\exp\left[-\exp\left[-a\left(u_k - \delta_k\right)\right]\right]$$

is the probability density function of a type I extreme value distribution with scale parameter a^{-1}, which integrates to 1. □

2.2.2 The Simple Logit Model

Let i denote an individual, and let j denote an alternative where $j \in \mathcal{J} := \left\{1, 2, \ldots, J\right\}$. Consumer i is assumed to choose up to one product in the set of alternatives $\mathcal{J}$. That is, now the consumer's choice is exclusive over the set of alternatives.[6] $\mathcal{J}$ may contain product 0, which is most commonly interpreted as choosing an outside option. The outside option, when included, is often interpreted as representing all other commodities that are not explicitly included in the choice set.

The latent utility is modeled as

$$u_{i,j} = \delta_j + \epsilon_{i,j}, \tag{2.2.2}$$

where δ_j is the utility from the observed product characteristics of product j[7] and $\epsilon_{i,j}$ represents the unobserved idiosyncratic utility shocks. The most common functional form used is the linear utility specification $\delta_j = \mathbf{x}_j'\boldsymbol{\theta}$. When $\mathcal{J}$ contains the outside option, normalizing $\mathbf{x}_0 = \mathbf{0}$ so that the mean utility of an outside option δ_0 is zero is common. Then, the utility levels of all other inside options are defined and identified against the outside option's utility level, normalized as zero.

Analogous with the binary choice model, the probability of individual i choosing product j is

$$\begin{aligned}\Pr\left(i \text{ chooses } j\right) &= \Pr\left(u_{i,j} > u_{i,k}, \forall k \neq j\right) \\ &= \Pr\left(\delta_j + \epsilon_{i,j} > \delta_k + \epsilon_{i,k}, \forall k \neq j\right)\end{aligned}$$

6. The setup and the assumptions on data availability are different from the binary choice model discussed in section 2.1. In the binary choice model, the choice over the set of alternatives was not exclusive—we assumed that the choice data on each alternative are available in the form of {0, 1}. Using the fact that the difference between two i.i.d. type I extreme-value random variables follows the logistic distribution, binary choice can be recast in the form of multiple choice with two alternatives {0, 1}.

7. Note that δ_j may include the unobserved (to the econometrician) attributes of alternative j. We discuss including the unobserved attributes in chapter 3. For now, we take δ_j to be composed only of the observed attributes.

$$= \Pr\left(\delta_j + \epsilon_{i,j} > \max_{k \neq j} \left\{\delta_k + \epsilon_{i,k}\right\}\right)$$

$$= \Pr\left(\mathbf{x}_j'\boldsymbol{\theta} + \epsilon_{i,j} > \max_{k \neq j} \left\{\mathbf{x}_k'\boldsymbol{\theta} + \epsilon_{i,k}\right\}\right), \tag{2.2.3}$$

where the last equality used the functional form $\delta_j = \mathbf{x}_j'\boldsymbol{\theta}$.

For the estimation, we assume the following to derive the closed-form probability of individual i choosing alternative j:

MLM(1) $\forall i, \forall k \neq j, \epsilon_{i,j} \perp\!\!\!\perp \epsilon_{i,k}$.
MLM(2) $\epsilon_{i,j} \sim T1EV\ (0)$.

Note MLM(1) and MLM(2) are often jointly abbreviated as $\epsilon_{i,j} \sim$ *i.i.d.* $T1EV\ (0)$.

Theorem 2.2.1. *(Simple Logit Likelihood over Multiple Choice)* Suppose that *MLM(1)* and *MLM(2)* hold. Then,

$$\Pr\left(i \text{ chooses } j\right) = \frac{\exp\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{\sum_{k \in \mathcal{J}} \exp\left(\mathbf{x}_k'\boldsymbol{\theta}\right)}. \tag{2.2.4}$$

Proof. Let $u_{i,j} = \delta_j + \epsilon_{i,j}, u_{i,-j} = \max_{k \neq j}\left\{\delta_k + \epsilon_{i,k}\right\}$. Let $\delta_{-j} = \ln\left[\sum_{k \neq j} \exp\left(\delta_k\right)\right]$. From lemma 2.2.2, we know that $u_{i,-j} \sim T1EV\left(\delta_{-j}\right)$. Under MLM(1) and MLM(2), we have

$$\begin{aligned}
\Pr\left(\delta_j + \epsilon_{i,j} \geq \max_k \left\{\delta_k + \epsilon_{i,k}\right\}\right) &= \Pr\left(u_{i,j} \geq u_{i,-j}\right) \\
&= \Pr\left(u_{i,-j} - u_{i,j} \leq 0\right) \\
&= \frac{\exp\left(\delta_j\right)}{\exp\left(\delta_{-j}\right) + \exp\left(\delta_j\right)} \qquad (2.2.5) \\
&= \frac{\exp\left(\delta_j\right)}{\sum_{k \in \mathcal{J}} \exp\left(\delta_k\right)}.
\end{aligned}$$

Equation (2.2.5) is derived by applying equation (2.2.1) with $u = 0$. Substituting $\delta_k = \mathbf{x}_k'\boldsymbol{\theta}\ \forall k$, we get the desired result.[8] □

8. Consider the odds ratios of the choice set $\mathcal{J} = \{0, 1\}$ with $\delta_0 = \epsilon_{i,0}$, $\delta_1 = \mathbf{x}_1'\boldsymbol{\theta} + \epsilon_{i,1}$, where $\epsilon_{i,j} \sim$ i.i.d. $T1EV\ (0)$. Lemma 2.2.3 yields that $\epsilon_{i,1} - \epsilon_{i,0}$ follows standard logistic distribution.

It is straightforward from equation (2.2.3) that when $\mathcal{J} = \{0, 1\}$ and $\mathbf{x}_0 = \mathbf{0}$, the equation boils down to the logit likelihood.

Corollary. *(Simple Logit Likelihood with Nonzero Mean Parameter)* Suppose that *MLM(1)* holds. Suppose that *MLM(2)* is replaced with $\epsilon_{i,j} \sim T1EV\left(\alpha_j\right)$. Then,

$$\Pr\left(i \text{ Chooses } j\right) = \frac{\exp\left(\mathbf{x}_j'\boldsymbol{\theta} + \alpha_j\right)}{\sum_{k\in\mathcal{J}} \exp\left(\mathbf{x}_k'\boldsymbol{\theta} + \alpha_k\right)}.$$

The individual choice probability equation (2.2.4) derived from the i.i.d. additive type I extreme-value shocks on the preferences plays a central role in many contexts. The choice probability itself can be used as the likelihood for the maximum-likelihood estimation or it can be equated with data that approximate the individual choice probabilities. We study the models of the latter type in depth in section 3.2.

Now suppose that we have the individual choice data $\left\{y_{i,j}, \mathbf{x}_j\right\}_{i\in\mathcal{I}, j\in\mathcal{J}}$. The likelihood of observing the data is

$$L\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{i\in\mathcal{I}, j\in\mathcal{J}} | \boldsymbol{\theta}\right) = \prod_{i\in\mathcal{I}} \left\{ \prod_{j\in\mathcal{J}} \Pr\left(i \text{ chooses } j\right)^{\mathbf{1}\left(i \text{ chooses } j\right)} \right\},$$

where $\Pr\left(i \text{ chooses } j\right)$ is as in equation (2.2.4). The log-likelihood and score function, which are required for the maximum-likelihood estimation, can be derived as usual. We emphasize that the model parameter $\boldsymbol{\theta}$ can be estimated using maximum likelihood only when δ_j in equation (2.2.2) contains no unknowns or unobservables. A more sophisticated method is required when an unobservable is included. We discuss some of those instances in section 2.3.4, and also in chapter 3.

2.2.3 Independence of Irrelevant Alternatives and the Nested Logit Model

Consider the ratio of simple logit choice probabilities:

$$\frac{\Pr\left(i \text{ chooses } j\right)}{\Pr\left(i \text{ chooses } k\right)} = \frac{\exp\left(\delta_j\right)}{\exp\left(\delta_k\right)} = \frac{\exp\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{\exp\left(\mathbf{x}_k'\boldsymbol{\theta}\right)}. \tag{2.2.6}$$

The ratio in equation (2.2.6), often referred to as the "odds ratio of choices j and k," is constant regardless of the average utility from other choices. The property is

called the "independence of irrelevant alternatives (IIA) property," which is pioneered by Luce (1959). IIA substantially restricts the substitution pattern over the alternatives. We study further how and why IIA may not be desirable in section 3.2.3, in the context of demand estimation. What we want to emphasize at this point is that, given that the mean utility $\delta_j = \mathbf{x}_j'\boldsymbol{\theta}$, the individual choice probability equation (2.2.4) is the only legitimate choice probability equation that satisfies equation (2.2.6) for each alternative in $\mathcal{J}$ and sums to 1. In that sense, the individual choice probability equation (2.2.4) can also be derived from (2.2.6) instead of the additive, idiosyncratic, type I, extreme-value distributed shocks.

One might question why the specific functional form of the ratios between the exponentiated mean utilities are used to characterize the IIA. In principle, any function that satisfies the following three conditions can be used instead of $\exp\left(\delta_j\right)$: (1) the function maps the entire real line onto the positive real numbers, (2) the function is strictly increasing in its domain, and (3) the function does not take δ_k for $k \neq j$ as its argument. The exponential function is the simplest elementary function that satisfies these three conditions. Notably, when the idiosyncratic preference shocks $\epsilon_{i,j}$ are i.i.d. across alternative j, then it would be possible to obtain a different functional form than the $\exp(\cdot)$ used in the characterization in equation (2.2.6). Put another way, the source of IIA is not the shape of type I extreme-value distribution, but the i.i.d. preference shocks across alternatives.

IIA may not be very appealing in the multiple discrete-choice contexts where a third alternative may affect the choice-probability ratios of the two alternatives under consideration. A popular workaround in the literature when individual choice-level data are available is nesting the choice set and modeling the individual's choice in multiple stages. Suppose that the choice set $\mathcal{J}$ can be divided into S disjoint subsets, which we call "modules." Each module is denoted by $\mathcal{B}_s$, where $s \in \{1, 2, \ldots, S\}$. If the joint distribution of $\left\{\epsilon_{i,j}\right\}_{j\in\mathcal{J}}$ takes the form

$$F\left(\left\{\epsilon_j\right\}_{j\in\mathcal{J}}\right) = \exp\left\{-\sum_{s=1}^{S}\alpha_s\left[\sum_{j\in\mathcal{B}_s}\exp\left(-\rho_s^{-1}\epsilon_j\right)\right]^{\rho_s}\right\},$$

it can be shown that the individual choice probabilities have the following closed-form formula:

$$\Pr\left(i \text{ chooses } \mathcal{B}_s\right) = \frac{\alpha_s\left(\sum_{k\in\mathcal{B}_s}\exp\left(\rho_s^{-1}\delta_k\right)\right)^{\rho_s}}{\sum_{\tau=1}^{S}\alpha_\tau\left(\sum_{k\in\mathcal{B}_\tau}\exp\left(\rho_\tau^{-1}\delta_k\right)\right)^{\rho_\tau}} \tag{2.2.7}$$

$$\Pr\left(i \text{ chooses } j | i \text{ chooses } \mathcal{B}_s\right) = \frac{\exp\left(\rho_s^{-1}\delta_j\right)}{\sum_{k\in\mathcal{B}_s}\exp\left(\rho_s^{-1}\delta_k\right)} \tag{2.2.8}$$

$$\Pr\left(i \text{ chooses } j\right) = \Pr\left(i \text{ chooses } j | i \text{ chooses } \mathcal{B}_s\right)\Pr\left(i \text{ chooses } \mathcal{B}_s\right). \tag{2.2.9}$$

$\{\alpha_s, \rho_s\}_{s=1}^{S}$ are the parameters to be estimated. The nesting structures can also be extended to more than two stages in an analogous way.

Given the utility specification $\delta_j = \mathbf{x}_j'\boldsymbol{\theta} + \epsilon_{i,j}$, the nested logit model in equations (2.2.7)–(2.2.9) can be estimated either by the full maximum likelihood or by the two-stage method. Although both methods yield the consistent estimates, the asymptotic variance formulas are different. For further discussions of implementing nested logit with the two-stage methods, see, e.g., McFadden (1981).

Nested logit allows a more complex choice structure than the simple logit, but two major weaknesses remain in the model in equations (2.2.7)–(2.2.9). First, it does not exhibit IIA across modules, but it still exhibits IIA within a module. Next, to implement the nested logit model, the econometrician has to impose prior knowledge on the choice structures, and thus on the composition of modules.

Another possible way to get around of the IIA property is the correlated probit. A benefit of using the correlated probit is that the Gaussian error term naturally allows the correlation of errors across alternatives. However, it leads to greater computational burden than the simple logit because the likelihood has no closed-form solution, and it usually involves evaluating the integral numerically. Furthermore, it is often questioned what variation in data identifies the cross-alternative covariance term of the idiosyncratic preference shocks. These are the major reasons why the somewhat restrictive simple logit model is still the major workhorse in practice.

2.2.4 Discussion

Historically, the development of the simple logit and nested logit models has gone in the opposite direction of our presentation. The type I extreme value distribution is carefully reverse engineered to yield the odds ratio of the form in equation (2.2.4). McFadden (1981) and Cardell (1997) generalized the simple logit model to a broader class called the generalized extreme-value class models.[9] The generalized extreme-value class choice models are often referred to as the "RUM model" to emphasize the connection between the resulting choice

9. The generalized extreme-value class includes nested logit model.