Structural Econometric Modeling in Industrial Organization and Quantitative Marketing

THEORY AND APPLICATIONS

ALI HORTAÇSU AND JOONHWI JOO

PRINCETON UNIVERSITY PRESS
PRINCETON & OXFORD

CONTENTS

1

Introduction: Structural Econometric Modeling

Structural econometric modeling is a set of approaches that rely extensively on economic theory to explicitly specify and test the relationships among distinct economic phenomena. The terminology defines three parts: structure, econometrics, and model. In what follows, we first discuss what each part of the terminology entails, in reverse order. Then we touch upon the debate around the structural econometric modeling approach against its reduced-form counterpart.

1.1 Model

This section discusses what an economic model is. Then we articulate when a model should be considered as capturing only correlations and when a model can be considered as capturing causality as well. We begin our discussion in a broader context of how models are built and tested in science.

1.1.1 Scientific Model and Economic Model

A scientific model consists of abstractions and simplifications of the real world, selecting and incorporating only the relevant aspects of the world that a researcher is analyzing. Scientific models are most commonly formulated using mathematical language. One of the major strengths of utilizing a model in science comes from its logic of establishing the relations among distinct variables: build a model and test the predictions from that model using real-world data. The main goal of building a model is to specify hypothetical relationship among distinct phenomena, summarized in the form of variables, in a testable form. Once a model is built, predictions from that model are subject to tests using statistical methods applied to real-world data. A statistical test of a scientific model is expressed in terms of testing the null and alternative hypotheses. Very roughly, the probability that the null hypothesis is not true given the data boils down to the p-value. That is, the p-value is gives the probability that a test statistic is obtained just by coincidence, given that (1) the null and alternative hypotheses are set up correctly, and (2) an adequate estimation method is used to compute the p-value. If the real-world data do not support the predictions from a model, the model is rejected. Models that are rejected less often are considered more reliable, and more reliable models are considered to provide more reliable predictions.

Economics stands on the same ground. Economists build economic models and test model predictions using data with econometric methods. An immediate question might arise: what defines a model as an economic model? We suggest that there are two key ingredients of an economic model: (1) optimizing behaviors of (2) the rational agent(s).[1] Economic theory begins from preferences, technology, information, and various equilibrium concepts. As a result of the optimizing behavior of one or multiple rational agents, observable/testable equilibrium outcomes are derived in the form of mathematical statements. Those outcomes are tested using real-world data with appropriate econometric methods.

1.1.2 Predictive Model and Causal Model

A model generally makes testable predictions about correlations between distinct variables. Such correlations can sometimes imply causal relationships between the variables of interest, generally under much more stringent conditions and assumptions. In this subsection, we discuss when a model can be interpreted as implying a causal relationship between distinct variables. We begin our discussion with the following two simple examples. Both examples involve linear models between explanatory and explained variables.

Example 1.1.1. Suppose that one has collected data on the height and weight of a randomly selected group in the population. Let y_i be the weight, and let x_i be the height of each individual. The researcher runs the following regression:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i. \tag{1.1.1}$$

The OLS estimate $\hat{\beta}_2$ turns out to be positive and highly statistically significant. Does this finding imply a causal relationship between height and weight?

1. Recent advances in several fields such as behavioral economics allow for violations of those two key ingredients. For instance, rationality might be bounded or optimization might be imperfect. Although we focus mostly on conventional microeconomic theory here, we do consider advances in behavioral economics as important progress in the profession.

Example 1.1.2. Suppose that one conducted a repeated Hooke's experiment and recorded the results. Let y_i be the length of the spring, and let x_i be the randomly assigned weight of the pendulum. Again, the researcher runs the following regression:

$$y_i = \gamma_1 + \gamma_2 x_i + \epsilon_i. \tag{1.1.2}$$

The OLS estimate $\hat{\gamma}_2$ is positive and highly statistically significant. Does this finding imply a causal relationship between the weight of the pendulum and the length of the spring?

The answer to the first question is definitely no.[2] But the answer to the second question is possibly yes. A positive and highly statistically significant $\hat{\gamma}_2$ estimate may be taken as evidence of a causal relationship—that is, x_i causes y_i. The structures of the two thought experiments seem to be quite similar at a glance; both equations (1.1.1) and (1.1.2) represent a linear model between x_i and y_i,[3] a data set is collected, a simple linear regression is run, and the coefficient estimates have the same sign and are statistically significant. But the implications on the causality can be starkly different. Where does this stark difference come from?

To answer this question, we first remind ourselves what regression reveals and what it does not. The slope coefficient estimate from a simple regression being positive (negative) is equivalent to the in-sample Pearson correlation coefficient between the explanatory variable and the explained variable being positive (negative).[4] If the data used in the regression are randomly sampled from the target population, high statistical significance can be interpreted as the positive (negative) sample correlation revealed from regression implying the positive (negative) population correlation.

What regression per se does not reveal is causality between the explanatory variable(s) and the explained variable. The experimental variations during the data-generating process are what make the correlation evidence of causality. Returning our focus to the two illustrative examples, the data on x_i of the second experiment are generated by a randomized experiment, where the researcher took full control over x_i. By contrast, the data on weight and height are not generated from a randomized experiment. Another possibly exogenous factor, such as good nutrition, is likely to simultaneously affect both height and weight; those exogenous factors are contained in the error term ϵ_i and treated as unobservable to the econometrician in the model considered.

2. If you are not convinced, recall Procrustes, the stretcher, in the *Odyssey*. When Procrustes stretches the guest to fit him in his bed, will the guest's weight increase?

3. We relegate the discussion on the role of ϵ_i to section 1.2.

4. Recall from elementary econometrics that the ordinary least squares (OLS) slope coefficient estimate is the sample covariance of x_i and y_i scaled by the sample variance of x_i.

An experimental variation in the explanatory variable(s) is essential for identifying the corresponding explanatory variable as a cause for change in the explained variable. The intuition behind the importance of experimental variation in establishing causality between two variables can be more easily illustrated in the context of omitted-variable bias in linear regression. Suppose that a causal and linear relationship exists between the vector of explanatory variables (x_i, v_i) and y_i, where v_i is unobserved to a researcher. Furthermore, assume that the correlation between x_i and v_i is nonzero, which is usual. If the sign and magnitude of the causal effect of interest are about variable x_i, a researcher may be tempted to run the following OLS regression:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

and claim $\hat{\beta}$ represents the causal effect of x_i on y_i. This claim is unarguably false unless the correlation between x_i and v_i is zero or the correlation between v_i and y_i is zero.[5] The problem with virtually any observational data is that infinitely many v_i's are possible that are not observed, and the best way to avoid this situation is to have x_i generated by an experiment, and therefore, it has zero correlation with any possible omitted variables.

The linear model in Example 1.1.2, once estimated using experimental data on length and weight as described previously, can be used to predict a *causal* effect of the explanatory variable(s) on the explained variable. A model that has causal interpretation is often referred to as a *causal model*. On the contrary, the linear model in Example 1.1.1, after being estimated using observational data on height and weight, cannot be used to predict a causal relation. However, it does not prevent one from using the model to predict a correlation between the explanatory variable(s) and the explained variable. A model that can only be used to predict the behavior of the explained variable using the explanatory variable is often referred to as a *predictive model*. The usefulness of a causal model is its capability to answer the questions related to *counterfactual* experiments; with only a predictive model, it is generally not possible to answer questions regarding counterfactuals. Counterfactuals are the ultimate goal of building and calibrating a structural econometric model. We will discuss more about counterfactuals in section 1.3.

1.2 Econometrics

Economic (theory) models often do not readily incorporate real-world data without an added stochasticity that is necessary to estimate and/or test the model. The key characteristic that discerns an econometric model from an economic model is

5. See any undergraduate-level econometrics textbook for the reasoning behind this point.

whether the model can directly incorporate relevant data. To incorporate relevant data, additional statistical structure should be added to an economic model. As is often the case, the added statistical structure is imposed in the form of added unobservable (both to the econometrician and/or to economic agents) variable(s) to the economic model of interest. The error terms ϵ_i in Examples 1.1.1 and 1.1.2, respectively, are examples of added unobservables; ϵ_i captures anything other than the assumed linear relationship between x_i and y_i, and it is impossible to rationalize data without the error term. We note that an economic model and an econometric model are sometimes indistinguishable because in some stochastic economic models, the unobservables (to the economic agents) are inherent in the economic model.

Conceptually, econometric models have three kinds of error terms. The first is due to researcher uncertainty, which is sometimes referred to as the "structural error" or "unobserved heterogeneity." This kind of error term is observable to the economic agent, but not to the econometrician. The structural errors affect the decision of the economic agents in the same way that the observables do. The second is driven by agent uncertainty. It is observable to neither the economic agent nor the econometrician. However, the variable may affect the economic agent's decision, often in terms of ex ante expectations. The third is the error term that is added merely for the rationalization of the data or the tractability of estimation. This type of error term may include measurement errors. Distinguishing between these concepts during the estimation is sometimes difficult or even impossible. However, being clear about these conceptual distinctions in the modeling stage is very important because the distinctions may affect the counterfactuals critically.

1.3 Structure

Conducting a counterfactual policy[6] experiment is one of the most important goals of building and calibrating/estimating an econometric model. Through counterfactual policy experiments, a researcher can answer questions related to changes in economic outcomes *caused* by hypothetical changes in a policy that affects economic agents. The key ingredients of an economic model explained in section 1.2, optimizing behaviors of rational agents involved and possible changes in the equilibrium, need to be accounted for during the counterfactual policy experiments; they need to be explicitly formulated in the econometric model to evaluate and quantify the causal effect of a change in policy.

6. The term "policy" is used in a broad sense here. It can be a firm's conduct, government regulation, consumers' choice environment, and so on; it does not necessarily mean public policy.

For a valid counterfactual policy experiment, certain aspects of the corresponding econometric model should be taken as invariant to possible changes in a policy; such invariant aspects are referred to as the *structure* of the model. Structure in a model is a set of restrictions how variables behave. For example, in the simple causal linear model discussed in Example 1.1.2, the key structure imposed is that y_i responds linearly to a change in x_i.[7] The model parameters of the econometric model, (γ_1, γ_2), are set free during the stages of calibration/estimation. Once the model parameters (γ_1, γ_2) are estimated, the parameter estimates are also taken as a part of the structure during predictions and counterfactual experiments.

Economic theory is the main source of the structure in a structural econometric model. The structure of many structural econometric models is nonlinear because most underlying economic models specify nonlinear relationships between the variables of interest up to the set of unknown parameters. By estimating a structural econometric model using real-world data, a researcher can obtain the magnitude of the parameters, in addition to their signs, in the underlying economic model. In turn, the magnitude of the effects resulting from a hypothetical change in a policy can be quantified; in contrast, it is often the case that only signs of the effects from a hypothetical policy change can be identified from the reduced-form counterparts of structural econometric models. However, the ability of quantifying the effects associated with a hypothetical policy change comes with its costs: the nonlinearity from explicitly specifying the possible relationships generally makes the structural econometric approach much more difficult to implement than its reduced-form counterpart.

Formulating and estimating a structural econometric model typically follow the following steps: (1) Formulate a well-defined economic model of the environment under consideration; (2) add a sufficient number of stochastic unobservables to the economic model; (3) identify and estimate the model parameters; and (4) verify the adequacy of the resulting structural economic model as a description of the observed data. In step (2), a researcher should decide whether to fully specify the distribution of the unobservables. Related to steps (2) and (3), estimation of structural econometric models often boils down to obtaining the point-identified, finite-dimensional, and policy-invariant model parameters.[8] A few possibilities

7. The econometric model in equation (1.1.2) is a structural model to the extent that the linearity is taken as coming from a valid theory that specifies the causal linear relationship between x_i and y_i. Note that it is also possible to interpret the econometric model in equation (1.1.2) as an approximation of a possibly nonlinear causal relationship between x_i and y_i. More discussions of the interpretation of the linear models follow in section 1.4.

8. The literature on the partially identified or nonparametric structural econometric models is growing. We study some examples of them in subsequent chapters.

exist for step (4). For example, the researcher can split the sample, estimate the model using only a subset of the sample, and examine the accuracy of the out-of-sample prediction. Another way of validating the structural models is to match the predictions of structural models with the data from a randomized experiment. We think an appropriate model validation is crucial to the credibility of the results from estimating a structural econometric model and conducting counterfactual policy experiments using the estimated structural model. A simple sensitivity analysis alone may not be enough to persuade the audience that the model is a credible and realistic approximation of the world.

1.4 Debate around the Structural Econometric Modeling Approach

Broadly, there are two ends of building an econometric model from an economic model: reduced-form and structural econometric models in a narrow sense.[9] There has been a debate in the literature between the structural and reduced-form approaches in econometric modeling.

Reduced-form econometric models abstract away from rational agents, optimization, and equilibria. They specify the simple relationships between the variables of interest and use relevant estimation methods to back out the parameters. Their econometric specifications are mostly linear, which has a justification that linear functions are a first-order approximation of any smooth functions. The strengths of reduced-form econometric models are their simplicity and relative robustness to the model misspecifications. On the other hand, a structural econometric model begins by explicitly stating the economic model specifications, such as the objective functions, the optimizing variable, the equilibrium concept, the degree of information of the agents and of the econometrician, and the possible source of endogeneity. Then, the model is solved step by step. As a result, the relations between the variables are specified in terms of the moment (in)equalities, likelihoods, or quantile restrictions. Finally, the relevant estimation methods for such specifications are used to back out the model parameters.

By explicitly specifying the economic models, structural econometric modeling enables one to make in-sample and out-of-sample predictions and policy counterfactuals. Specifically, the ability to make out-of-sample causal predictions is one of the greatest strengths of a structural econometric model. For instance,

9. In a wide sense, even the linear instrumental-variable model is a structural econometric model, implicitly imposing a very specific structure on how the instrumental variables affect the outcome variables. This point has been thoroughly investigated by Heckman and Vytlacil (2005).

a reduced-form model of merger identified using retrospective analysis may be enough to predict a merger impact if the analyst is interested in predicting the effect of counterfactual merger with similar attributes to retrospective ones. However, if one is interested in simulating mergers under a different market environment, a linear extrapolation is likely to be a poor fit. Furthermore, the linear shape and even the direction of the merger impact suggested by the reduced-form model may not be valid anymore under some counterfactual policy experiments, subject to the "Lucas critique" (see Lucas 1976). By explicitly specifying and estimating the policy-invariant nonlinear economic relationships between the market environment and the equilibrium outcomes of a merger, structural econometric modeling allows one to make predictions out-of-sample.

A disadvantage of structural econometric modeling is that the predictions or policy counterfactuals can be sensitive to model misspecifications. The possibility of model misspecification is considered one of the greatest weaknesses in the structural econometric modeling approach, especially because structural econometric models generally take sophisticated nonlinear causal relationships between variables, inherited from the underlying economic theory, as given and fixed a priori. Ideally, every ingredient in a structural econometric model could be tested by running carefully designed, randomized experiments, but it is generally very difficult when the subject of study is the economic behavior of individuals or organizations.

Taking either approach does not exclude the other, and much successful research has used one approach to inform work with the other. That said, we view the reduced-form approach and structural approach to econometric modeling as complements with different strengths, not substitutes, as explained previously.

1.5 Outline of This Book

Modern empirical industrial organization and quantitative marketing rely extensively on the structural econometric modeling approach using observational data. The goal of this textbook is to give an overview of how the various streams of literature in empirical industrial organization and quantitative marketing use structural econometric modeling to estimate the model parameters, give economic-model-based predictions, and conduct policy counterfactuals.

This book consists of six chapters and an appendix. We discuss the basics of single-agent static and dynamic discrete choice in chapter 2, which is now a standard baseline modeling framework in empirical industrial organization, quantitative marketing, and many other adjacent fields. In chapter 3, we move on to study demand estimation with market data, where we introduce demand-estimation methods in the product space and characteristics space, respectively. In chapter 4, we focus on strategic interactions of firms in the static and dynamic setup. We then move our focus back to consumers to study the empirical frameworks of consumer search in chapter 5. Finally, we study the theory and empirics of auctions in chapter 6. For completeness, we also summarize basic features of the most commonly used baseline estimation frameworks in the appendix.

The book does not cover many interesting relevant topics, such as production function estimation methods and Bayesian learning models. We refer the readers to relevant survey papers and handbook chapters to learn more about these topics.[10]

Bibliography

Ackerberg, D., C. L. Benkard, S. Berry, & A. Pakes (2007). Chapter 63, "Econometric tools for analyzing market outcomes." In vol. 6A of *Handbook of Econometrics*, 4171–4276. Amsterdam: Elsevier.

Angrist, J. D., & J.-S. Pischke (2010). "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of Economic Perspectives*, 24, 3–30.

Blundell, R. (2010). "Comments on: Michael P. Keane 'Structural vs. atheoretic approaches to econometrics'," *Journal of Econometrics*, 156, 25–26.

Ching, A. T., T. Erdem, & M. P. Keane (2013). "Learning models: An assessment of progress, challenges, and new developments," *Marketing Science*, 32, 913–938.

Ching, A. T., T. Erdem, & M. P. Keane (2017). Chapter 8, "Empirical models of learning dynamics: A survey of recent developments." In *Handbook of Marketing Decision Models*, 223–257. Springer.

Deaton, A. (2010). "Instruments, randomization, and learning about development." *Journal of Economic Literature*, 48, 424–455.

de Loecker, J., & C. Syverson (2021). Chapter 3, "An industrial organization perspective on productivity." In vol. 4 of *Handbook of Industrial Organization*, 141–223. Amsterdam: Elsevier.

Heckman, J. J., & E. Vytlacil (2005). "Structural equations, treatment effects, and econometric policy evaluation." *Econometrica*, 73, 669–738.

Imbens, G. W. (2010). "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature*, 48, 399–423.

Keane, M. P. (2010). "Structural vs. atheoretic approaches to econometrics." *Journal of Econometrics*, 156, 3–20.

Keane, M. P., P. E. Todd, & K. I. Wolpin (2011). Chapter 4, "The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications." In vol. 4a of *Handbook of Labor Economics*, 331–461. Amsterdam: Elsevier.

Laibson, D., & J. A. List (2015). "Principles of (behavioral) economics." *American Economic Review: Papers and Proceedings*, 105, 385–390.

Leamer, E. E. (1983). "Let's take the con out of econometrics." *American Economic Review*, 73, 31–43.

Low, H., & C. Meghir (2017). "The use of structural models in econometrics." *Journal of Economic Perspectives*, 31, 33–58.

Lucas, R. E. (1976). "Econometric policy evaluation: A critique." In *Carnegie-Rochester Conference Series on Public Policy*, vol. 1, 19–46. Amsterdam: Elsevier.

10. For production function estimation methods, see, e.g., de Loecker and Syverson (2021), and for Bayesian learning models, see, e.g., Ching, Erdem, and Keane (2013, 2017).

Nevo, A., & M. D. Whinston (2010). "Taking the dogma out of econometrics: Structural modeling and credible inference." *Journal of Economic Perspectives*, 24, 69–82.

Reiss, P. C., & F. A. Wolak (2007). Chapter 64, "Structural econometric modeling: Rationales and examples from industrial organization." In vol. 6A of *Handbook of Econometrics*, 4277–4415. Amsterdam: Elsevier.

Rust, J. (2010). "Comments on: "Structural vs. atheoretic approaches to econometrics" by Michael Keane." *Journal of Econometrics*, 156, 21–24.

Rust, J. (2014). "The limits of inference with theory: A review of Wolpin (2013)." *Journal of Economic Literature*, 52, 820–850.

Wolpin, K. I. (2013). *The limits of inference without theory*. Cambridge, MA: MIT Press.

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:

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

The next-to-last equality follows from $1-\Pr\left(\epsilon_{i,j}\leq-\mathbf{x}_j'\boldsymbol{\theta}\right)=\Pr\left(\epsilon_{i,j}\leq\mathbf{x}_j'\boldsymbol{\theta}\right)$ 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:

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

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, $\frac{\partial \Pr(y_{i,j}=1|\mathbf{x}_j)}{\partial \mathbf{x}_j^{(l)}}$, 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:

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

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. $\square$

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.

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

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

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

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.

probabilities and the axioms of stochastic choice, which is a set of axioms used to rationalize a stochastic choice.

The simple logit model over the multiple alternatives examined thus far is often referred to as the "conditional logit model" in the literature. There is another class of simple logit model, referred to as the "multinomial logit model." The multinomial logit likelihood is given by

$$\Pr\left(i \text{ Chooses } j\right) = \frac{\exp\left(\mathbf{x}'\boldsymbol{\theta}_j + \alpha_j\right)}{1 + \sum_{k=2}^{J} \exp\left(\mathbf{x}'\boldsymbol{\theta}_k + \alpha_k\right)}.$$

Note that the key differences between the multinomial logit model and the conditional logit model are (1) whether the choice likelihood contains the attribute vector that varies over alternatives and (2) whether the slopes are varying over alternatives.

Anderson, de Palma, and Thisse (1992) presents a classic reference to the theory of static discrete choice, which provides an extensive treatment of establishing the connection between the discrete choice framework and differentiated-product demand framework. We will study some aspects of the discrete choice framework as an aggregate consumer-demand framework in chapter 3. Train (2009) provides a detailed treatment of the microfoundation and estimation of discrete-choice models.

2.3 Single-Agent Dynamic Discrete Choice

Economic agents exhibit forward-looking behavior when the current state variables affect the current and future payoffs and the payoff streams change with respect to accumulation flow on the state variables. In this section, we study single-agent dynamic discrete-choice models, a combination of discrete choice with forward-looking economic agents. We focus on the setups in which the modeled dynamics are stationary, so the value function boils down to a two-period Bellman equation.

We first study the method of Rust (1987)'s bus-engine-replacement problem, which has become the standard baseline dynamic discrete-choice framework applied to discrete-choice data. Then we discuss the key ideas of Hotz and Miller (1993); Hotz et al. (1994), which is based on inverting the conditional choice probabilities. Then we study the nested pseudo-likelihood method by Aguirregabiria and Mira (2002), which is based on finding the dual problem of the value function fixed-point equations in the conditional choice probability space. Finally,