SPECIFICATION, ESTIMATION AND TESTING OF TREATMENT EFFECTS IN MULTINOMIAL OUTCOME MODELS: ACCOMMODATING ENDOGENEITY AND INTER-CATEGORY COVARIANCE

Shichao Tang

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Economics,
Indiana University

September 2018

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Joseph V. Terza, Ph.D., Chair

_____

Paul Carlin, Ph.D.

June 18, 2018

_____

Hsien-Chang Lin, Ph.D.

_____

Gwendolyn Morrison, Ph.D.

_____

Boyoung Seo, Ph.D.

ACKNOWLEDGEMENT

Thirteen years ago, I left my hometown, a small town located in the central part of China, and went to Beijing for college. I never thought one day I would have the chance to go to the United States to study, not to mention getting a Ph.D. degree here. I am extremely grateful for the support and help I have received along this fantastic journey. I would like to first thank my parents for their unconditional support in the past five years. They have always been encouraging me to pursue my dreams.

I am also deeply indebted to my advisor, Dr. Joseph V. Terza, who also served as my dissertation chair. Dr. Terza has provided tremendous help and advice on my dissertation. He is very intelligent and always cares about his students. I feel blessed to have Dr. Terza as my advisor.

I would like to thank my other dissertation committee members: Dr. Paul Carlin, Dr. Hsien-Chang Lin, Dr. Gwendolyn Morrison, and Dr. Boyoung Seo. I appreciate their comments and feedback on my dissertation. I would particularly like to thank Dr. Carlin for his encouragement throughout the past five years. I am also particularly thankful for Dr. Hsien-Chang Lin who provided me with insightful understanding on public health research and shared with me his research experience.

I am also grateful to other faculty members in the Department of Economics, especially Dr. Anne Royalty, our Ph.D. program director, who puts together a fantastic Ph.D. program and cultivates a great research environment. I also would like to thank Dr. Steven Russell, the chair, for his tremendous work on assigning us to different graduate assistant positions. My sincere thanks also go to Dr. Mark Ottoni-Wilhelm, Dr. Vidhura Tennekoon, Dr. Henry Mak, Dr. Jaesoo Kim, Dr. Sumedha Gupta, Dr. Subir Chakrabarti,

Dr. Jisong Wu, and Dr. Yaa Akosa Antwi for providing me with solid and rigorous training in Economics. I am also thankful to Terri Crews and Dana Ward for their help with different administrative affairs.

I would like to thank Dr. Kosali Simon who taught my health economics class in Bloomington when I was pursuing my M.P.A. degree and inspired me to get interested in health economics.

I am extremely grateful to my dear friends in both the United States and China. I cannot imagine how my life would look without my friends' help and support.

Shichao Tang

SPECIFICATION, ESTIMATION AND TESTING OF TREATMENT EFFECTS IN MULTINOMIAL OUTCOME MODELS: ACCOMMODATING ENDOGENEITY AND INTER-CATEGORY COVARIANCE

In this dissertation, a potential outcomes (PO) based framework is developed for causally interpretable treatment effect parameters in the multinomial dependent variable regression framework. The specification of the relevant data generating process (DGP) is also derived. This new framework simultaneously accounts for the potential endogeneity of the treatment and loosens inter-category covariance restrictions on the multinomial outcome model (e.g., the independence from irrelevant alternatives restriction). Corresponding consistent estimators for the "deep parameters" of the DGP and the treatment effect parameters are developed and implemented (in Stata). A novel approach is proposed for assessing the inter-category covariance flexibility afforded by a particular multinomial modeling specification [e.g. multinomial logit (MNL), multinomial probit (MNP), and nested multinomial logit (NMNL)] in the context of our general framework. This assessment technique can serve as a useful tool for model selection. The new modeling/estimation approach developed in this dissertation is quite general. I focus here, however, on the NMNL model because, among the three modeling specifications under consideration (MNL, MNP and NMNL), it is the only one that is both computationally feasible and is relatively unrestrictive with regard to inter-category covariance. Moreover, as a logical starting point, I restrict my analyses to the simplest version of the model – the trinomial (three-category) NMNL with an endogenous treatment (ET) variable conditioned on individual-specific covariates only. To identify potential computational issues and to

assess the statistical accuracy of my proposed NMNL-ET estimator and its implementation (in Stata), I conducted a thorough simulation analysis. I found that conventional optimization techniques are, in this context, generally fraught with convergence problems. To overcome this, I implement a systematic line search algorithm that successfully resolves this issue. The simulation results suggest that it is important to accommodate both endogeneity and inter-category covariance simultaneously in model design and estimation. As an illustration and as a basis for comparing alternative parametric specifications with respect to ease of implementation, computational efficiency and statistical performance, the proposed model and estimation method are used to analyze the impact of substance abuse/dependence on the employment status using the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) data.

Joseph V. Terza, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction, Background, Significance, and Summary

The objective of most empirical economic research is to provide solid evidence that can be used to evaluate past, current and future policy. Essential to this goal is rigorous specification and accurate estimation of parameters that characterize the key causal relationships. In this dissertation, I focus on the specification, estimation and testing of treatment effects in multinomial outcome models. The relevant treatment effect specifications and estimation method are developed in a potential outcomes (PO) framework so as to ensure the causal interpretability of the targeted effect parameters and their estimates. Based on that PO (structural) model, I show how, under certain conditions, possible endogeneity of the treatment variable can be accommodated in a multinomial outcomes model of the relevant data generating process (DGP). Moreover, within this endogeneity-corrected multinomial outcomes DGP framework, a new modeling strategy that allows for inter-category covariance flexibility (ICF) is developed and evaluated. Conventional modeling approaches that do not accommodate ICF are subject to misspecification bias and yield counterintuitive predictions – such models are often plagued by the so-called independence from irrelevant alternatives (IIA) restriction. To my knowledge, the modeling approach that is introduced in this dissertation is the first to accommodate both the endogeneity of treatment and ICF, simultaneously.

The remainder of the dissertation is organized as follows. I begin, in Chapter 2, by rigorously defining the ultimate estimation objective (the *average treatment effect - ATE*) in a well-defined PO framework for models with multinomial outcomes and binary treatment variables. Therein, I also derive a corresponding consistent ATE estimator whose formulation follows from a given specification for the multinomial PO. It is seen

that, in addition to the specification of the multinomial PO, implementation of this estimator requires a consistent estimate of the "deep" parameters of the model. Chapter 3 focuses on the details of development of this consistent estimator of the deep parameters of the relevant PO. This novel estimation method is designed to accommodate both the endogeneity of the treatment variable and ICF. A variety of versions of the model and estimator corresponding to different multinomial likelihood specifications are detailed. An important component of this discussion is a comparison of the alternatives with respect to the degree to which each of them accommodates ICF. To make such assessments, a new standardized measure of ICF is proposed. In Chapter 4, a thorough simulation study of the nested logit model with an endogenous treatment is conducted. The simulation results suggest that it is important to accommodate both endogeneity and inter-category covariance simultaneously in model design and estimation. In Chapter 5, I apply the nested logit model with endogeneity in the context of real data analysis – the effect of substance abuse (binary) on multinomial employment status (out of the labor force, unemployed, or employed). The database used for this purpose is the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Chapter 6 summarizes and concludes.

Chapter 2: Fully Parametric Multinomial Outcome (FPMO) Models and Causal Inference

In this dissertation, the goal is to specify and estimate the causal treatment effect in an FPMO model while accommodating *both* endogeneity *and* ICF within the same model specification. This chapter begins with a general review of the potential outcome (PO) framework as discussed by Terza (2018).[1] Relevant specific concepts within the PO framework are detailed. These concepts are then used to specify the *average treatment effect* (ATE) in the multinomial outcome (MO) context. It is important to cast the MO ATE in the PO framework because, by doing so, the conditions under which it is causally interpretable (CI) can be made explicit. As Terza (2018) makes clear, the essential condition here is that, conditional on the relevant vector of control variables, the multinomial PO and the observed version of the causal treatment variable are stochastically independent. I note here that some of the requisite regression control variables are unobservable. (This is the part of the specification that embodies the endogeneity of treatment.) This chapter continues by proposing a stylized consistent sample analog estimator of the MO ATE. This estimator is a sample analog in the sense that mathematical expectation and deep population parameters are replaced by mathematical expectation and consistent parameter estimates, respectively. The design of this estimator, at this point,

[1] The main focus of this dissertation is comparison of specifications and estimators for the "deep" parameters (e.g., random utility parameters) of a multinomial outcome model while simultaneously allowing for endogeneity of treatment and inter-category covariance flexibility. As will be seen, inclusion of these two modeling features adds considerable analytic and computational complication to the discussion. For this reason, I have tried to keep other features of the model as simple as possible without substantial loss of generality. For instance: 1) I only explicitly consider the simplest version of the treatment effect – the ATE – and leave discussions of other versions like the *local average treatment effect, marginal treatment effect, average treatment effect on the treated* and *average treatment effect on the untreated* (see Heckman et al., 2003 and Heckman and Vltlacil, 2005) to future research; 2) I focus on the trinomial case; and 3) I only consider model specifications involving observation-specific regressors that do not vary across alternatives (e.g. person-specific demographic and socio-economic characteristics).

assumes resolution of two key practical issues: a) the essential fact that requisite elements of the relevant vector of regression controls are unobservable; and b) consistent estimation of the deep population parameters. These are the main topics covered in Chapter 3. Chapter 2 concludes by explicitly specifying the generic form of the probability mass function (pmf) of the observed version of the MO conditional on the causal variable and the regression controls, including the unobservables, (under the conditional mean independence condition noted above). This pmf formulation is a key component of the relevant data generating process (DGP) and, therefore, is essential to the discussion in Chapter 3 whose focus is resolving aforementioned issues (a) and (b).

2.1 Specifying the Treatment Effect of Interest in the Potential Outcomes Framework[2]

The focus here is rigorous specification and accurate estimation of the ATE in a FPMO modeling context characterizing the causal relationship between a *binary policy variable of interest* ($\mathbf{X_p}$), which to some degree is (or can be brought) under the control of a policy maker, and a specified multinomial *outcome of policy interest* ($\mathbf{Y}$).[3] I first draw the distinction between two versions of the $\mathbf{X_p}$:[4]

[2] See Terza (2018) for a detailed and more general discussion of the PO framework.

[3] $\mathbf{X_p}$ and $\mathbf{Y}$ are to be taken as global replacements for the phrases "policy variable of interest" and "outcome of policy interest," respectively.

[4] Henceforth I will adhere to the following notational conventions: 1) uppercase letters for random variables (e.g., A); 2) lowercase letters for particular values in the support of the random variable in question (e.g., a) and 3) uppercase letters with an "i" subscript for the sampled version of the random variable in question (e.g. A_i with i = 1, …, n indicating the ith observation from a sample of size n).

$X_p \equiv$ the binary random variable representing the observable (*factual*) version of the distribution of the $\mathbf{X_p}$ (The sampled values of the policy variable are drawn from the distribution of X_p.)

and

$X_p^* \equiv$ the binary random variable upon which a policy relevant counterfactual is based (e.g., if the distribution of the $\mathbf{X_p}$ were that of X_p^*, what would be the distribution of the $\mathbf{Y}$?).[5] Note X_p^* is, by design, independent of all other variates germane to the present discussion.

Likewise I distinguish two versions of the $\mathbf{Y}$:

$Y \equiv [Y_1 \ldots Y_J] \equiv$ the random vector representing the observable factual version of the distribution of the $\mathbf{Y}$ (The sampled values of the outcome are drawn from the distribution of Y.). Each of the individual elements of Y are binary variables and Y is a mutually exclusive and collectively exhaustive outcome

and

$Y_{X_p^*} \equiv [Y_{X_p^*1} \ldots Y_{X_p^*J}] \equiv$ the random vector representing the distribution of *potential outcomes*, defined as the distribution of values of the $\mathbf{Y}$ that would have manifested for a particular X_p^*. Stated as a relevant counterfactual. If the distribution of the $\mathbf{X_p}$ were X_p^*

[5] To clarify, I use the Free Dictionary definition of the term *counterfactual* – a conditional statement in which the first clause is a past tense subjunctive statement expressing something contrary to fact. When I use the term *counterfactual* as an adjective in describing a random variable I mean to convey that it has been, can be, or will be used as a key component of a relevant counterfactual.

then the distribution of the $\mathbf{Y}$ would be that of $Y_{X_p^*}$. $Y_{X_p^*}$ denotes a multinomial outcome (i.e., a mutually exclusive and collectively exhaustive categorical outcome), defined such that, for the jth category, $Y_{X_p^* j} = 1$ and all other elements of $Y_{X_p^*}$ are equal to zero.

Throughout the remainder of the discussion, I focus on the *average treatment effect* (ATE) defined in the above PO framework as

$$ATE_j = E[Y_{0j}] - E[Y_{1j}] \qquad (1)$$

The ATE in (1) is configured so as to conform to the illustrative example (detailed in Chapter 5) in which the $\mathbf{Y}$ is employment status [ES] (out of the labor force, unemployed, or employed) and the $\mathbf{X_p}$ is substance abuse [SA] (current abuse and/or dependence on alcohol, marijuana, cocaine, heroin, sedatives, tranquilizers, opioids, amphetamines, solvents, hallucinogens, or other drugs). In the context of this illustration, the causally interpretable (CI) parameter in (1) measures the potential impact on ES of a fully effective drug treatment policy designed to prevent and eliminate SA. The effect of the prevention component of the treatment policy is measured as the resultant change in the likelihood of being in a particular ES category. For instance, in the illustrative example, the prevention effect is represented by the increase in the probability of being in the "employed" state as a result of "preventing" the non-abusers from becoming abusers (under the assumption that they would have become abusers). The effect of the elimination component of the treatment policy is measured likewise. For instance, in the illustrative example, the

elimination effect is represented by the increase in the probability of being in the "employed" state as a result of convincing all abusers to become non-abusers.[6]

2.2 FPMO Models in the Potential Outcomes Framework

Parameters like (1) cannot be directly estimated from data because Y_{1j} and Y_{0j} are counterfactual and, therefore, do not represent observable statistical populations from which samples can be drawn. Let us suppose, however, that the potential outcome $Y_{X_p^*}$ (defined in the previous section) has the following conditional probability mass function (pmf) given a vector of covariates C

$$\text{pmf}(Y_{X_p^*} \mid C) = f_{(Y_{X_p^*} \mid C)}(X_p^*, C; \tau) = \prod_{j=1}^{J} \pi_j(X_p^*, C; \tau)^{Y_{X_p^* j}} \qquad (2)$$

where $\pi_j(X_p^*, C; \tau) = \Pr(Y_{X_p^* j} = 1 \mid C)$ is the jth category probability (of has known form) and τ is a vector of unknown parameters.[7] Henceforth, I refer to the vector τ as comprising the "deep" parameters of the model. It follows from (2) that all of the conditional moments of $Y_{X_p^*}$ are of known form. In particular, for j = 1, ..., J the conditional mean of $Y_{X_p^* j}$ is

$$E[Y_{X_p^* j} \mid C] = \Pr(Y_{X_p^* j} = 1 \mid C) = \pi_j(X_p^*, C; \tau). \qquad (j = 1, ..., J) \qquad (3)$$

---

[6] I realize that the hypothetical policy that I describe in this section is somewhat stylized and that the ATE as defined in (1) is best viewed as an upper bound on the relevant policy effect. As mentioned in footnote 1, throughout the dissertation, in order to keep this aspect of the discussion as simple as possible (without substantial loss of generality), I maintain (1) as policy-relevant parameter of interest. The focus of the present research is the development and comparison of methods for simultaneously dealing with possible endogeneity of the policy variable and a flexible covariance structure for the multinomial outcome categories.

[7] $g_{(A|C)}(A, B, C; \psi)$ denotes the pmf of A conditional on C, written as a function of A, B, C with parameter vector ψ.

Using the law of iterated expectations, combining (1) with (3)

$$ATE_j = E[\pi_j(0, C; \tau)] - E[\pi_j(1, C; \tau)] \quad (4)$$

To this point in the discussion, not much about the role and composition of the vector of covariates C has been said. For the remainder of the dissertation I assume that it can be partitioned as

$$C = [X_o \quad X_u] \quad (5)$$

where X_o is a vector of observable controls (confounders – loosely speaking, variables that are correlated with both Y and X_p) and X_u is a scalar composite of all relevant unobservable confounders.[8] Note that X_u is included so as to account for the potential endogeneity of X_p. The variate X_u, in fact, embodies the endogeneity of X_p because it comprises all unboservables that are related to both Y and X_p. Given this specification of C in (5), (4) can be rewritten as

$$ATE_j = E\left[E_{X_u}[\pi_j(0, X_o, X_u; \tau) - \pi_j(1, X_o, X_u; \tau)]\right] \quad (6)$$

2.3 Estimating the Treatment Effect of Interest

X_u cannot be observed but, as seen later in this dissertation, I will formalize its aforementioned relationship with X_p and, given a consistent estimate of τ (say $\hat{\tau}$), this will be sufficient for consistent estimation of (6) using

[8] See Terza (2018) for a more rigorous definition of the term "confounder."

$$\widehat{\text{ATE}_j} = \sum_{i=1}^{n} \frac{1}{n} \left\{ E_{X_u} \left[\pi_j(0, X_o, X_u; \hat{\tau}) - \pi_j(1, X_o, X_u; \hat{\tau}) \right] \right\} \quad (7)$$

Asymptotic standard errors for the $\widehat{\text{ATE}_j}$ (j = 1, ..., J) can be obtained using the approach of Terza (2016a,b).[9] Note that because (7) is consistent for (1), which is CI, results obtained from (7) are also CI. As this discussion makes clear, two items are needed for the implementation of (7): 1) formalization of the relationship between X_u and X_p; and 2) a consistent estimator for deep parameters: τ.

2.4 Towards Consistent Estimation of the Deep Parameters of the FPMO Model with an Endogenous Treatment

As is the case in any full information estimation context, in order to implement the relevant maximum likelihood estimator, the explicit specification of the pmf characterizing the data generating process (DGP) for $[Y \;\; X_p \;\; X_o \;\; X_u]$ is needed.

Without imposing certain conditions, the DGP does not directly follow from the PO model in (2). Terza (2018) discusses the requisite conditions and I argue that these conditions are satisfied in the present context. To summarize, these conditions are sufficient to ensure that, given X_o and X_u, Y and X_p be conditionally independent, in which case the true data generating process (DGP) for $[Y \;\; X_p \;\; X_o \;\; X_u]$ is such that[10]

$$\text{pmf}(Y \mid X_p, X_o, X_u)$$

$$= f_{(Y_{X_p^*} | X_o, X_u)}(Y, X_p, X_o, X_u; \tau)$$

[9] See Appendix A detailing the asymptotic standard errors for the ATE estimators in (7).
[10] See Appendix B

$$= \prod_{j=1}^{J} \pi_j(X_p, X_o, X_u; \tau)^{Y_j} \quad (8)$$

Expression (8) amounts to substitution of Y for $Y_{X_p^*}$, and X_p for X_p^*, in (2).

Equation (8) is important because it is the key for the estimation of the deep parameters to be discussed in the next chapter. Equation (8) bridges the gap between the potential outcome framework and the true DGP so that the casually interpretable ATEs and deep parameters in the potential outcome framework can be estimated using data.

Chapter 3: General and Specific Versions of the FPMO Model with an Endogenous Treatment

Chapter 3 begins by detailing the random utility maximization (RUM) behavioral model underlying structure of the MO models to be considered in this dissertation. It is shown how, in this RUM context, alternative structural specifications for the PO (and by implication the corresponding DGP) follow from the pmf formulation detailed at the end of Chapter 2 and alternative specifications for the random components in the RUM model. The underlying motivation for this discussion is assessment of the ICF afforded by such alternative specifications weighed against difficulty of implementation (Stata coding and computational efficiency). A fully generic version of the RUM-based MO is developed in the PO framework. This generic model is one that imposes no ICF restrictions. For simplicity of exposition, and without substantive loss of generality, the remainder of the dissertation focuses on the trinomial outcome case. A "standardized" version of this model is derived. In order to better understand the implications of this standardized model, its fully restricted version is detailed (zero covariance among the MO categories). Based on this standardized model, a method for assessing the inter-category covariance restrictiveness of a particular MO PO modeling specification is developed. This assessment method allows direct comparison of the ICF afforded by chosen model specifications with the fully restrictive version of the generic standardized model. At this point in the chapter, I note that three MO PO specifications will be considered in this dissertation – multinomial logit (MNL), multinomial probit (MNP) and nested multinomial logit (NMNL). MNL is computationally easy to implement but affords no ICF (zero covariance among the MO categories). MNP lies at the other extreme of the spectrum. It

offers maximum ICF but is quite computationally challenging. NMNL resides somewhere between MNL and MNP in that it allows some degree of ICF but is computationally practical.

At this point, I suspend more detailed discussions of the MNL, MNP and NMNL modeling approaches until later in the chapter, turning instead to a discussion of the resolution of items (a) and (b) mentioned at the end of Chapter 2 for a given MO PO specification – viz., (a) how to handle the unobservable controls; and b) how to consistently estimate the deep parameters of the model. Following the general approach suggested by Terza (2009), I propose that item (a) be resolved by assuming a conventional parametric binary response specification for the causal treatment variable whose random component is a scalar comprising the unobservable controls needed for invoking conditional independence between the causal treatment variable and the MO PO. Implicit is such a specification is an assumed distribution for the said random component. For example, one can assume a probit model for the causal binary treatment model (as is done in the simulation and real data analyses conducted in chapters 4 and 5 – discussed later). Implicit in this model is the assumption that the random component, which by assumption encompasses the requisite unobservable regression controls in the MO PO model) is standard normally distributed. This resolves item (a).

Following the general modeling approach suggested by Terza (2009), using the assumed binary response model for the causal treatment variable (e.g. binary probit), and a given MO PO specification (e.g., MNL, MNP or NMNL), Chapter 3 then continues with a derivation of the pmf of the joint density of the MO PO and the causal treatment variable conditional on the observable regression controls and identifying instrumental variables

(IV). (The IV are needed to identify: i) the deep parameters of the MO model that pertain to the causal treatment variable; and ii) by implication, the ATE of the causal treatment variable.) Given the conditional mean independence of the MO PO and the causal treatment variable (i.e., conditional on the observable and unobservable controls), I derive the relevant joint pmf of the observed causal treatment variable and the observed MO conditional on the observed regression controls and the IV. From this pmf follows the relevant log-likelihood function which can be maximized to obtain consistent estimates of the deep parameters of the model. This resolved item (b).

As promised earlier, the remainder of Chapter 3 details the MNL, MNP and NMNL specifications for the MO PO conditional on the observable and unobservable confounders (regression controls). In this part of the discussion, special attention is paid to the extent of ICF afforded by each of the alternative specifications. In the main, ICF is characterized in terms of the independence from irrelevant alternatives (IIA) condition. Models that are subject to the IIA condition are more likely to produce biased estimates of the targeted causal effects and yield counterintuitive predictions. In the case of NMNL, in addition to discussing how this specification can avoid IIA, I demonstrate how the standardized measure (discussed earlier in this chapter) can be applied in the simplest version of the model.

3.1 The General Formulation of the PO and DGP for the FPMO Model with an Endogenous Treatment

The PO specification in expression (2) [and by implication the DGP in (8)] requires an explicit form for $\pi_j(X_p^*, X_o, X_u; \tau)$. There are a number of alternative specifications

for these category probabilities; many of which are derived from a primitive behavioral model called the *random utility model (RUM).*[11] In the RUM it is assumed that

$$Y_{X_p^*} \equiv [0 \ldots \underset{\substack{\uparrow \\ \text{jth element}}}{1} \ldots 0] \qquad \text{iff.} \qquad Y^o_{X^*_{pj}} = \max\{Y^o_{X^*_{pr}};\ r = 1, \ldots, J\} \tag{9}$$

where

$$Y^o_{X^*_{pr}} = X^*_p \beta^o_{pr} + X_o \beta^o_{or} + X_u \beta^o_{ur} + \varepsilon_r \tag{10}$$

$X_o \equiv$ a vector of observation-specific characteristics that do not vary across alternatives)

$X_u \equiv$ the scalar representing the unobservable confounders

$\beta^{o\prime}_r = [\beta^o_{pr} \quad \beta^{o\prime}_{or} \quad \beta^o_{ur}]$ is the vector of coefficient parameters for the rth utility index[12]

and $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \ldots \varepsilon_J]$ is a vector of random error terms having a constant mean vector and covariance matrix

[11] It is not necessary that the multinomial category probabilities be based on an underlying RUM structure. As will be seen later in the dissertation, one can formulate a coherent multinomial probability structure that is not based on the RUM.

[12] Note here that there other types of relevant observable confounders that might, if available in the data, be included in the specification of (10); e.g., a vector of alternative-specific characteristics that do no vary across individuals, a vector of observable characteristics that vary across both individuals and alternatives and/or a vector of observable characteristics of the alternatives that are constant across the alternatives in each of a specified set of subgroups (nests) but vary across subgroups (nests). I include only observation-specific confounders that do not vary across alternatives for simplicity of exposition but without loss of substantive generality.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1J} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1J} & \sigma_{2J} & \cdots & \sigma_J^2 \end{pmatrix}. \qquad (11)$$

Note that β_{ur}^o and the elements of β_r^o vary across alternatives.

For simplicity of exposition, and without substantive loss of generality, let us focus on the trinomial case (J = 3). From the general version of the model given in (10) and (11) there are

3 coefficient parameter vectors for the utility indexes, viz.,

$$[\beta_{p1}^o \quad \beta_{o1}^{o\prime} \quad \beta_{u1}^o], \; [\beta_{p2}^o \quad \beta_{o2}^{o\prime} \quad \beta_{u2}^o] \text{ and } [\beta_{p3}^o \quad \beta_{o3}^{o\prime} \quad \beta_{u3}^o].$$

3 variance parameters σ_j (j = 1, 2, 3)

and

3 covariance parameters $\sigma_{j\ell}$ (j, ℓ = 1, 2, 3; j$\neq\ell$).

In this framework, if the distribution of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$ is known, the multinomial outcome probability for the first category, as defined in equation (2), can then be written as[13]

$$\pi_j(X_p^*, C; \tau) = \Pr(Y_{X_{pj}^*} = 1 \mid C)$$
$$\pi_1(X_p^*, X_o, X_u; \tau) = \Pr(Y_{X_{p1}^*} = 1 \mid X_o, X_u)$$

[13] See Appendix C detailing similar derivations for the general multinomial model (arbitrary J)

$$= \Pr(\xi_2 \le -V_2^*, \xi_3 \le -V_3^*)$$

$$= G_1(-V_2^*, -V_3^*) \tag{12}$$

and for categories $j = 2$ or 3

$$\pi_2(X_p^*, X_o, X_u; \tau) = \Pr(Y_{X_p^* 2} = 1 \mid X_o, X_u)$$

$$= \Pr(-\xi_2 \le V_2^*, \xi_3 - \xi_2 \le V_2^* - V_3^*)$$

$$= G_2(V_2^*, V_2^* - V_3^*) \tag{13}$$

$$\pi_3(X_p^*, X_o, X_u; \tau) = \Pr(Y_{X_p^* 3} = 1 \mid X_o, X_u)$$

$$= \Pr(\xi_2 - \xi_3 \le V_3^* - V_2^*, -\xi_3 \le V_3^*)$$

$$= G_3(V_3^* - V_2^*, V_3^*) \tag{14}$$

where

$$V_j^* = X_p^* \beta_{pj}^* + X_o \beta_{oj}^* + X_u \beta_{uj}^* \tag{15}$$

$$\beta_j^* = \beta_j^o - \beta_1^o$$

$G_1(\bullet, \bullet)$ denotes the bivariate cdf of ξ_2 and ξ_3

$G_2(\bullet, \bullet)$ denotes the bivariate cdf of $-\xi_2$ and $\xi_3 - \xi_2$

$G_3(\bullet, \bullet)$ denotes the bivariate cdf of $\xi_2 - \xi_3$ and $-\xi_3$

and

$$\xi_j = \varepsilon_j - \varepsilon_1 . \tag{16}$$

The exact formulations of $G_1(\bullet, \bullet)$, $G_2(\bullet, \bullet)$ and $G_3(\bullet, \bullet)$ follow from the assumed distribution of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$. Note that an admissible reduction of the model is evident in (12), (13) and (14) because it allows the relevant probabilities to be expressed in terms of the $J - 1$ normalized coefficient parameter vectors of the form $[\beta_1^* \quad \beta_j^*]$.[14]

I now turn to the derivation of the standardized versions of (12), (13) and (14). I do this in order to gain insight regarding parametric identification and as a means of standardizing the characterization of covariance flexibility for the various FPMO models arising from alternative specifications for the distribution of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$. In the trinomial case there are 2 unique utility differences ξ_2 and ξ_3. This means that there are at most 2 identified variance parameters

$$\text{var}(\xi_j) = \omega_j^* = \sigma_j^2 + \sigma_1^2 - 2\sigma_{1j} \tag{17}$$

for j = 2 or 3, and a one covariance parameter

$$\text{cov}(\xi_2, \xi_3) = \omega_{23}^* = \sigma_{23} - \sigma_{12} - \sigma_{13} + \sigma_1^2. \tag{18}$$

After standardizing the relevant random variables ξ_2, ξ_3 and $(\xi_3 - \xi_2)$ by dividing by their respective standard deviations, the model can be admissibly reduced by

1) dividing V_j^* by $\sqrt{\omega_2^*}$ (19)

and

2) dividing ω_j^* and ω_{23}^* by ω_2^* (20)

[14] For the definition of an *admissible reduction* see Terza (1985).

for j = 1, to obtain the following bivariate expressions

$$\pi_1(X_p^*, X_o, X_u; \tau) = G_1^s\left(-V_2, \frac{-V_3}{\sqrt{\omega_3}}\right) \tag{21}$$

where

$G_1^s(\bullet, \bullet) =$ the bivariate cdf of ξ_2 and $\frac{\xi_3}{\sqrt{\omega_3}}$

$$V_j = \frac{V_j^*}{\sqrt{\omega_2^*}} = X_p^*\beta_{pj} + X_o\beta_{oj} + X_u\beta_{uj} \tag{22}$$

$$\beta_{pj} = \frac{1}{\sqrt{\omega_2^*}}\beta_{pj}^*$$

$$\beta_{oj} = \frac{1}{\sqrt{\omega_2^*}}\beta_{oj}^*$$

$$\beta_{uj} = \frac{1}{\sqrt{\omega_2^*}}\beta_{uj}^*$$

and

$$\omega_3 = \omega_3^* / \omega_2^*.$$

The correlation matrix for $[\xi_2 \quad \xi_3]$ (i.e., the covariance matrix of $[\xi_2 \quad \xi_3/\sqrt{\omega_3}]$) is

$$D_1[\xi_2 \quad \xi_3] = \begin{bmatrix} 1 & \frac{\omega_{23}}{\sqrt{\omega_3}} \\ \frac{\omega_{23}}{\sqrt{\omega_3}} & 1 \end{bmatrix} \tag{23}$$

where $\omega_{23} = \omega_{23}^* / \omega_2^*$. Similarly, for j = 2 and 3, respectively,

$$\pi_2(X_p^*, X_o, X_u; \tau) = G_2^s\left(V_2, \frac{(V_2 - V_3)}{\sqrt{1+\omega_3 - 2\omega_{23}}} \right) \tag{24}$$

and

$$\pi_3(X_p^*, X_o, X_u; \tau) = G_3^s\left(\frac{(V_3 - V_2)}{\sqrt{1+\omega_3 - 2\omega_{23}}}, \frac{V_3}{\sqrt{\omega_3}} \right) \tag{25}$$

where

$G_2^s(\bullet, \bullet) =$ the bivariate cdf of ξ_2 and $\frac{\xi_3 - \xi_2}{\sqrt{1+\omega_3 - 2\omega_{23}}}$

$G_3^s(\bullet, \bullet) =$ the bivariate cdf of $\frac{\xi_2 - \xi_3}{\sqrt{1+\omega_3 - 2\omega_{23}}}$ and $\frac{\xi_3}{\sqrt{\omega_3}}$.

with respective correlation matrices

$$D_2[\xi_2 \quad (\xi_3 - \xi_2)] = \begin{bmatrix} 1 & \frac{1-\omega_{23}}{\sqrt{1+\omega_3 - 2\omega_{23}}} \\ \frac{1-\omega_{23}}{\sqrt{1+\omega_3 - 2\omega_{23}}} & 1 \end{bmatrix} \tag{26}$$

and

$$D_3[(\xi_2 - \xi_3) \quad \xi_3] = \begin{bmatrix} 1 & \frac{\omega_3 - \omega_{23}}{\sqrt{\omega_3(1+\omega_3 - 2\omega_{23})}} \\ \frac{\omega_3 - \omega_{23}}{\sqrt{\omega_3(1+\omega_3 - 2\omega_{23})}} & 1 \end{bmatrix}. \tag{27}$$

Note that in all of the above expressions, ω_2 is normalized to be equal to 1. The full parameter vector in the standardized model is

$$\tau = [[\beta_{p2} \quad \beta'_{o2} \quad \beta_{u2}] \quad [\beta_{p3} \quad \beta'_{o3} \quad \beta_{u3}] \quad \omega_3 \quad \omega_{23}]. \tag{28}$$

In summary, in this trinomial case, the number of parameters is admissibly reduced from

nine (three random utility coefficient vectors, three variance parameters and three covariance parameters) to four (two identified random utility coefficient vectors one variance parameter and one covariance parameter). The parameterization (28) for this standardized version of the general model, is identifiable in the sense that it is the result of admissible (and, therefore, requisite) reduction. Moreover, as the above discussion makes clear, it accommodates unrestricted correlation amongst the elements of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$ and, therefore, allows concomitant unrestricted correlation flexibility for $[\xi_2 \quad \xi_3]$, $[\xi_2 \quad (\xi_3 - \xi_2)]$ and $[(\xi_2 - \xi_3) \quad \xi_3]$ in that the off-diagonal elements of (23), (26) and (27) all freely range from -1 to 1. As will be discussed later in the dissertation, such covariance flexibility will have a bearing on the generalizability of the estimation results.

Within this framework one can also examine the most covariance restrictive case in which the elements of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$ are assumed to be independently and identically distributed. The relevant version of (11) for the trinomial case is

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}. \tag{29}$$

Correspondingly the following versions of (17) and (18) are

$$\text{var}(\xi_j) = \omega_j^* = 2\sigma^2 \tag{30}$$

and

$$\text{cov}(\xi_2, \xi_3) = \omega_{23}^* = \sigma^2 . \tag{31}$$

After standardizing the relevant random variables ξ_2, ξ_3 and $(\xi_3 - \xi_2)$ by dividing by their respective standard deviations, the model can be admissibly reduced by applying (19) and (20) to obtain the following bivariate expressions for the trinomial category probabilities. For j = 1

$$\pi_1(X_p^*, X_o, X_u; \tau) = G_1^s\left(-V_2, -V_3\right) \tag{32}$$

where

$G_1^s(\bullet, \bullet) =$ the bivariate cdf of the standardized versions of ξ_2 and ξ_3

$$V_j = \frac{V_j^*}{\sqrt{2\sigma^2}} = X_p^*\beta_{pj} + X_o\beta_{oj} + X_u\beta_{uj} \text{ (j = 2 or 3)} \tag{33}$$

$$\beta_{pj} = \frac{1}{\sqrt{2\sigma^2}}\beta_{pj}^*$$

$$\beta_{oj} = \frac{1}{\sqrt{2\sigma^2}}\beta_{oj}^*$$

and

$$\beta_{uj} = \frac{1}{\sqrt{2\sigma^2}}\beta_{uj}^*.$$

The correlation matrix for $[\xi_2 \quad \xi_3]$ in this case is

$$D[\xi_2 \quad \xi_3] = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}. \tag{34}$$

Similarly for j = 2 and 3, respectively,

$$\pi_2(X_p^*, X_o, X_u; \tau) = G_2^s\left(V_2, (V_2 - V_3)\right) \tag{35}$$

and

$$\pi_3(X_p^*, X_o, X_u; \tau) = G_3^s\left((V_3 - V_2), V_3\right) \tag{36}$$

where, respectively

$G_2^s(\bullet, \bullet) =$ the bivariate cdf of the standardized versions of ξ_2 and $\xi_3 - \xi_2$

$G_3^s(\bullet, \bullet) =$ the bivariate cdf of the standardized versions of $\xi_2 - \xi_3$ and ξ_3.

The corresponding correlation matrices are both identical to (34). This model is completely correlation (covariance) restrictive in that the off-diagonal element of (34) is fixed at $\frac{1}{2}$.

Writing such FPMO models in standard form, as above, provides a "standardized" way of comparing alternative model specifications with respect to the degree of covariance restrictiveness. The two models discussed above represent the extremes and are characterized in this regard by the specifications of (23), (26) and (27) for the least covariance restrictive case and by (34) for the most restrictive case. Models that lie in the middle ground between these two extremes are usually specified via a parametric reduction of the unrestricted model comprising (10) and (11) – say $\eta = \eta(\tau)$, where τ is defined in (28) and the dimension of η is smaller than that of τ. Under such a reduction, using my standardization approach, the covariance (correlation) flexibility of the model can be characterized using the following version of (23)

$$D_{1\eta}[\xi_2 \quad \xi_3] = \begin{bmatrix} 1 & \frac{\omega_{23}(\eta)}{\sqrt{\omega_3(\eta)}} \\ \frac{\omega_{23}(\eta)}{\sqrt{\omega_3(\eta)}} & 1 \end{bmatrix} \tag{37}$$

and similarly rewritten versions of (26) and (27) $\{D_{2\eta}[-\xi_2 \quad (\xi_3 - \xi_2)]$ and $D_{3\eta}[(\xi_2 - \xi_3) \quad -\xi_3]\}$ wherein the functional forms of $\omega_3(\eta)$ and $\omega_{23}(\eta)$ are known. Given the known distribution of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$ and the known formulations of $\omega_3(\eta)$ and $\omega_{23}(\eta)$, the ranges of the off-diagonal elements of $D_{1\eta}$, $D_{2\eta}$ and $D_{3\eta}$ over the domain of η can be traced. This will yield a standardized relative measure of covariance flexibility vis-a-vis the two extreme cases discussed above.

As I have noted above, covariance flexibility will depend on the assumed distribution for $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \ldots \varepsilon_J]$. In this dissertation, I discuss alternative specifications for this distribution and characterize them with regard to covariance flexibility using the standardization approach described above. The most computationally tractable of the FPMO models is multinomial logit (MNL). MNL, while computationally simple, resides at the most covariance restrictive extreme of the spectrum because it is based on the iid assumption embodied in (29) [implying (34)]. Multinomial probit (MNP), on the other hand, affords full covariance flexibility as characterized by (11) [implying (23), (26) and (27)]. This model, however, has proven to be computationally intractable. The third alternative, nested multinomial logit (NMNL), allows some covariance flexibility and is computationally feasible (in a relative sense). For each of these model specifications, respectively, in sections 3.2, 3.3 and 3.4, I discuss their implications regarding an unwanted

consequence of covariance restrictiveness – the so-called independence from irrelevant alternatives (IIA) condition. I now I turn to the analytic development of the generic form of the new modeling/estimation approach proposed in this dissertation.

One can readily see that if X_u were observable, the maximum likelihood estimator (MLE) of τ based on (8) would be consistent. By the same token, if X_u is a confounder for Y and X_p, a MLE based on (8) that ignores X_u (e.g., fixes its coefficient at 0) will be inconsistent. This is an instance of classical endogeneity bias. If X_u is a confounder for Y and X_p, failure to account for its role in (8), will result in spurious attribution of its influence on Y to X_p. In order to account for such potential endogeneity, I formalize the confounding relationship between X_u and X_p as follows:

$$X_p = I(W\alpha + X_u > 0) \tag{38}$$

where W is a vector of observable regressors, α is vector of parameters conformable with W, the distribution of $(X_u \mid W)$ is known and I(C) is the indicator function whose value is 1 if condition C holds and 0 if not. Assume that $W = [X_o \quad W^+]$, where W^+ is a vector of identifying instrumental variables (IV). To qualify as a vector of IV, W must satisfy the following conditions

(1) $E[X_u \mid W] = 0$

(2) Exclusion restriction:

$$f_{(Y|X_p,X_o,X_u)}(Y, X_p, X_o, X_u; \tau)$$

$$= \text{pmf}(Y \mid X_p, X_o, X_u)$$

$$= \text{pmf}(Y \mid X_p, W, X_u)$$

$$= f_{(Y|X_p,W,X_u)}(Y, X_p, W, X_u; \tau)$$

which implies

$$\pi_j(X_p, X_o, X_u; \tau) = E[Y_j \mid X_p, X_o, X_u]$$

$$= E[Y_j \mid X_p, W, X_u] = \pi_j(X_p, W, X_u; \tau)$$

and

(3) $\text{COV}(X_p, W)$ is sufficiently different from zero. (39)

Condition (1) implies that W is not correlated with X_u. Condition (2) is called the exclusion restriction because it articulates the fact that the IV in W^+ play no role in [are excluded from] (8). Condition (3) is often paraphrased as "the IV must be strong", where the metric for "strength" is the level of correlation between X_p and W.

One cannot base an MLE for τ on (8) because X_u is unobservable. Following Terza (2009), combining (8) and (38) I can, however, obtain the joint pdf of Y and X_p conditional on W as

$$f_{(Y,X_p|W)}(Y, X_p, W; \tau, \alpha) = \int_{-\infty}^{\infty} f_{(Y,X_p,X_u|W)}(Y, X_p, X_u, W; \tau, \alpha) dX_u$$

$$= \int_{-\infty}^{\infty} f_{(Y|X_p,W,X_u)}(Y, X_p, W, X_u; \tau, \alpha) g_{(X_p,X_u|W)}(X_p, X_u, W; \tau, \alpha) dX_u \quad (40)$$

From (8) and the exclusion restriction in (39) it follows that

$$f_{(Y|X_p,W,X_u)}(Y, X_p, W, X_u; \tau) = \prod_{j=1}^{J} \pi_j(X_p, X_o, X_u; \tau)^{Y_j} . \quad (41)$$

Moreover, the joint pdf of X_p, X_u given W is

$$g_{(X_p,X_u|W)}(X_p, X_u, W; \tau, \alpha) = (g(X_u)_{[X_u > -W\alpha]})^{X_p} (g(X_u)_{[X_u \le -W\alpha]})^{1-X_p} \quad (42)$$

where g( ) denotes the known pdf of $(X_u \mid W)$ and $g(A)_{[r]}$ represents g( ) subject to support restriction "r". Combining (40) through (42) yields

$$\begin{aligned} & f_{(Y,X_p|W)}(Y, X_p, W; \tau, \alpha) \\ & = \int_{-\infty}^{\infty} \prod_{j=1}^{J} \pi_j(X_p, X_o, X_u; \tau)^{Y_j} (g(X_u)_{[X_u > -W\alpha]})^{X_p} (g(X_u)_{[X_u \le -W\alpha]})^{1-X_p} dX_u \\ & = \left[\int_{-W\alpha}^{\infty} \prod_{j=1}^{J} \pi_j(X_p, X_o, X_u; \tau)^{Y_j} g(X_u)\, dX_u \right]^{X_p} \\ & \quad \times \left[\int_{-\infty}^{-W\alpha} \prod_{j=1}^{J} \pi_j(X_p, X_o, X_u; \tau)^{Y_j} g(X_u)\, dX_u \right]^{1-X_p} \end{aligned} \quad (43)$$

Using (43) the following sample log-likelihood function can be constructed

$$q(\hat{\tau}, \hat{\alpha}, Z_i) = \sum_{i=1}^{n} \ln f_{(Y,X_p|W)}(Y_i, X_{pi}, W_i; \hat{\tau}, \hat{\alpha}) \quad (44)$$

where $Z_i = [Y_i \quad X_{pi} \quad W_i]$ is the data vector.

Consistent estimates of τ and α can be obtained by maximizing (44). Note that as there are non-closed-form integrals in the log-likelihood function (44), Gauss-Legendre quadrature will be used to numerically approximate them, which adds more computational complexity into the implementation of the estimator. It is important to acknowledge this

because one criteria of assessing different models is their computational efficiency. Inference using these estimates can be conducted via standard asymptotic theory. In particular, correct asymptotic standard errors for the MLE parameters $\hat{\tau}$ and $\hat{\alpha}$ can be obtained as the square roots of the elements of the diagonal of the estimated information matrix.

In the remainder of this chapter, I will detail the MNL, MNP and NMNL specifications for the MO PO conditional on the observable and unobservable confounders.

3.2 Multinomial Logit and Independence from Irrelevant Alternatives

Suppose that $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$ is i.i.d. log-Weibull distributed (i.e., the pdf of ε_{ij} is $g(\varepsilon_j) = \exp[-\varepsilon_j - \exp\{-\varepsilon_j\}]$ and is independent across j). Under this assumption $E[\varepsilon_j] = \gamma$, where $\gamma \approx .5772$ is the Euler-Mascheroni constant and $\sigma^2 = \text{var}(\varepsilon_j) = \frac{\Pi^2}{6}$.[15] As have been seen, the i.i.d. assumption implies that the covariance matrix of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$ is (29). Given the results from McFadden (1973) for this FPMO modeling specification, and the discussion supporting equations (29) through (36), the standardized version of this model yields the following relevant versions of (32), (35) and (36)

$$\pi_1(X_p^*, X_o, X_u; \tau) = G_1^s\left(-V_2, -V_3\right)$$

$$= \frac{1}{1 + \sum_{r=2}^{3} \exp\left(X_p^* \beta_{pr} + X_o \beta_{or} + X_u \beta_{ur}\right)} \qquad (45)$$

[15] For simplicity of exposition and without loss of substantive generality, I continue to focus on the trinomial case.

$$\pi_2(X_p^*, X_o, X_u; \tau) = G_2^s\left(V_2, (V_2 - V_3)\right)$$

$$= \frac{\exp\left(X_p^*\beta_{p2} + X_o\beta_{o2} + X_u\beta_{u2}\right)}{1 + \sum_{r=2}^{3} \exp\left(X_p^*\beta_{pr} + X_o\beta_{or} + X_u\beta_{ur}\right)} \tag{46}$$

$$\pi_3(X_p^*, X_o, X_u; \tau) = G_2^s\left(V_2, (V_2 - V_3)\right)$$

$$= \frac{\exp\left(X_p^*\beta_{p3} + X_o\beta_{o3} + X_u\beta_{u3}\right)}{1 + \sum_{r=2}^{3} \exp\left(X_p^*\beta_{pr} + X_o\beta_{or} + X_u\beta_{ur}\right)} \ . \tag{47}$$

and V_j is defined as in (33). The correlation matrices for this standardized MNL model are all identical to (34). Therefore, this FPMO specification lies at the most restrictive extreme of the covariance flexibility spectrum.

As a result of such extreme covariance restrictiveness, the model can produce counterintuitive predictions. This fact is well illustrated by the classic Red Bus - Blue Bus problem. Suppose that initially commuters are faced with the three options (walk, car, red bus) with corresponding choice likelihoods as defined in (45) through (47). Consider the addition of a fourth alternative -- blue bus. Suppose that the blue bus alternative is identical to the red bus alternative in *every* respect except color -- an attribute that is assumed to have no influence on mode choice. Let P_j^b denote the likelihood of the j^{th} alternative *before* the introduction of the blue bus, where J = 3 with j =1 if walk, 2 if car and 3 if red bus. Likewise and let P_j^a denote the likelihood *after* the blue bus is introduced, where J = 4 and j =4 denotes the blue bus alternative. The expanded version of the model in this scenario comprises (45) through (47) with the summation in the denominators running from 2 to 4

and the addition of a blue bus (j = 4) category likelihood which is identical to (47) with the aforementioned denominator adjustment. Logic dictates that $P_3^a = P_4^a = (1/2)P_3^b$, $P_1^a = P_1^b$, and $P_2^a = P_2^b$. It is, however, easy to show by example that this logical outcome is contradicted by the MNL model. For example, let $P_1^b = P_2^b = P_3^b = 1/3$. It is easy to see that in the MNL model for all j and ℓ the ratio P_j / P_ℓ does not depend on the characteristics of alternatives other than j and ℓ. In our example, this implies that $P_j / P_\ell = 1$ for all pairs of alternatives both before and after the addition of the blue bus; in other words $P_1^a = P_2^a = P_3^a$. In addition, as mentioned earlier, according to logic $P_3^a = P_4^a$ would be expect. These conditions taken together, however, imply that the category probabilities after the introduction of the blue bus alternative must be such that $P_1^a = P_2^a = P_3^a = P_4^a = 1/4$. But this contradicts the logical result that the probabilities of walking and driving should both have remained equal to 1/3. This example highlights the restriction imposed by the MNL specification that is commonly termed the *Independence from Irrelevant Alternatives* (IIA) condition. Formally, a RUM is said to be subject to the IIA condition if for all j and ℓ the ratio P_j / P_ℓ does not depend on the characteristics of alternatives other than j and ℓ.

Because the MNL model is plagued by the IIA condition, alternative models that are more covariance flexible and, therefore, not subject to the IIA restriction have been developed. In the following section I consider one such model, viz. the multinomial probit (MNP) model. The MNP model lies at the least restrictive extreme of the covariance flexibility spectrum. Unfortunately, the MNP has proven difficult to implement in empirical research due to the substantial computational burden it imposes and convergence problems related to practical identification issues.

3.3 Multinomial Probit and IIA

In the MNL model, it is assumed that the random components of the utility indexes are independently distributed. It is this independence assumption that lies at the heart of the IIA problem plaguing the MNL model. In the multinomial probit (MNP) model the IIA condition can be circumvented by specifying a covariance-flexible random utility error structure. The MNP model is defined as in (9) and (10) with the additional assumption that $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \ldots \varepsilon_J]$ is multivariate normal distributed with mean 0 and unrestricted covariance matrix Σ [as defined in (11)]. When J = 3, I can show that standardized version of the model comprises[16]

$$\pi_1(X_p^*, X_o, X_u; \tau) = \Phi_2\left[-V_2, \frac{-V_3}{\sqrt{\omega_3}}; \frac{\omega_{23}}{\sqrt{\omega_3}}\right] \tag{48}$$

$$\pi_2(X_p^*, X_o, X_u; \tau) = \Phi_2\left[V_2, \frac{V_2 - V_3}{\sqrt{1+\omega_3-2\omega_{23}}}; \frac{1-\omega_{23}}{\sqrt{1+\omega_3-2\omega_{23}}}\right] \tag{49}$$

$$\pi_3(X_p^*, X_o, X_u; \tau) = \Phi_2\left[\frac{V_3 - V_2}{\sqrt{1+\omega_3-2\omega_{23}}}, \frac{V_3}{\sqrt{\omega_3}}; \frac{\omega_3-\omega_{23}}{\sqrt{\omega_3(1+\omega_3-2\omega_{23})}}\right]. \tag{50}$$

where V_j is defined in (22), ω_3 and ω_{23} are defined as in (23) and $\Phi_s[a_1, \ldots, a_s; R]$ denotes the s-variate standard normal cdf with correlation matrix R evaluated at

[16] In this case V_2 and V_3 are normally distributed because they are linear combinations of normal. This, of course, implies that $V_2 - V_3$ is also normal and any pair of these three variates is bivariate normal; hence the appearance of $\Phi_2[\bullet, \bullet; \bullet]$ in (48) through (50). The remaining details from those expressions follow from (21) through (27).

$(a_1, ..., a_s)$.[17] It is clear that all of the relevant correlation matrices in this model are unrestricted – i.e., the correlation coefficients range from -1 to 1. Therefore, this FPMO specification lies at the least restrictive extreme of the covariance flexibility spectrum.

Let's now examine how the MNP model circumvents the IIA condition. I begin by discussing the red bus/blue bus example introduced above. As I noted earlier, intuitively I expect that after the introduction of the blue bus alternative $P_1^a = P_1^b$, $P_2^a = P_2^b$, and $P_3^a = \frac{1}{2}P_3^b$. I can show that with the addition of the blue bus alternative[18]

$$P_1^a = \Phi_3\left[-V_2, \frac{-V_3}{\sqrt{\omega_3}}, \frac{-V_3}{\sqrt{\omega_3}}, D_1\right] \tag{51}$$

where

$$D_1 = \begin{bmatrix} 1 & \frac{\omega_{23}}{\sqrt{\omega_3}} & \frac{\omega_{23}}{\sqrt{\omega_3}} \\ & 1 & 1 \\ & & 1 \end{bmatrix}. \tag{52}$$

But because the second and third arguments of (51) are equal and the 23th element of D_1 is equal to 1, the third integral is redundant so by (48)

$$P_1^a = \Phi_2\left[-V_2, \frac{-V_3}{\sqrt{\omega_3}}; \frac{\omega_{23}}{\sqrt{\omega_3}}\right] = P_1^b. \tag{53}$$

[17] To simplify exposition I have, for the bivariate expressions (48) – (50), replaced the correlation matrices (R) with the relevant correlation coefficient.
[18] See Appendix D detailing the MNP model for the J=4 case.

It can likewise be shown that $P_2^a = P_2^b$. By (50)

$$P_3^b = \Phi_2\left[\frac{V_3 - V_2}{\sqrt{1+\omega_3-2\omega_{23}}}, \frac{V_3}{\sqrt{\omega_3}}; \frac{\omega_3-\omega_{23}}{\sqrt{\omega_3(1+\omega_3-2\omega_{23})}}\right] \tag{54}$$

and I can show that

$$P_3^a = \Phi_3\left[\frac{(V_3 - V_2)}{\sqrt{1+\omega_3-2\omega_{23}}}, \frac{V_3}{\sqrt{\omega_3}}, 0; D_3\right] \tag{55}$$

where

$$D_3 = \begin{bmatrix} 1 & \frac{\omega_3-\omega_{23}}{\sqrt{\omega_3(1+\omega_3-2\omega_{23})}} & 0 \\ & 1 & 0 \\ & & 1 \end{bmatrix}. \tag{56}$$

Therefore

$$P_3^a = \Phi_2\left[\frac{(V_3 - V_2)}{\sqrt{1+\omega_3-2\omega_{23}}}, \frac{V_3}{\sqrt{\omega_3}}; \frac{(\omega_3-\omega_{23})}{\sqrt{\omega_3(1+\omega_3-2\omega_{23})}}\right] \bullet \Phi(0) = \frac{1}{2}P_3^b \tag{57}$$

where $\Phi(\bullet)$ denotes the univariate standard normal cdf. Therefore, in the red-bus blue-bus scenario the MNP model is consistent with logic. This example can easily be generalized to prove that the MNP model is not subject to IIA.

Although the MNP offers maximum covariance flexibility and, therefore, fully circumvents the IIA, it has not been widely implemented using the unfettered specification detailed above. The main problematic issue encountered in implementing the MNP is the

fact that the formulations of the category likelihoods involve multi-dimensional integrals that do not have closed forms. In trinomial models like the one described here, packaged and fairly accurate bivariate standard normal cdf numerical approximation algorithms are available. For models with more than four categories, however, reliable and efficient numerical integral approximation algorithms are not generally available. Moreover, even in the trinomial case, convergence problems are invariably encountered because of near non-concavities in the likelihood function (problems with so-called practical identification).

3.4 Nested Multinomial Logit and IIA

I now consider a model that lies in the middle ground between the extremely covariance restrictive but computationally attractive MNL model and the unrestrictive but computationally burdensome MNP model -- the nested logit (NMNL) model. The NMNL model is defined as in (9) and (10) with the additional assumption that $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \ldots \varepsilon_J]$ is generalized extreme value (GEV) distributed. For the purpose of exposition and without loss of generality with regard to the main conceptual points to be made I focus on the trinomial case in which the cumulative distribution function (cdf) of ε is the following member of the GEV class of distributions

$$F(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \exp\left(-\left[e^{-\varepsilon_1} + \left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{1-\eta} \right] \right) \tag{58}$$

where scalar η is the only parameter of the distribution and, in the most general case, can be any value in the nonnegative half of the real line aside from 1.[19,20] I refer to η as *inter-category covariance-related parameter*. First note that (58) can be factored in the following way

$$F(\varepsilon_1, \varepsilon_2, \varepsilon_3) = F_{\varepsilon_1}(\varepsilon_1) F_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3) \tag{59}$$

where

$$F_{\varepsilon_1}(\varepsilon_1) = \exp(-e^{-\varepsilon_1}) \tag{60}$$

is the log-Weibull marginal cdf of ε_1, and

$$F_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3) = \exp\left(-\left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{1-\eta} \right) \tag{61}$$

is the joint marginal cdf of ε_2 and ε_3. The marginal distributions of ε_2 and ε_3 are also log-Weibull.[21] Equation (59) embodies the dependence between ε_2 and ε_3 and independence of ε_1 from both of these variates. In general, such GEV specifications for the errors in the random utility indices are appropriate for cases in which the multinomial outcome categories can be partitioned such that there is possible covariance among the categories within a partition but independence from categories in other partitions. Such partitions are called *nests*.

[19] See equation 3.27 of Maddala (1983).
[20] If $0 \leq \eta < 1$, the corresponding model specification is consistent with the RUM.
[21] I prove this in Appendix E.

In the present trinomial example there are two nests – one that includes category 1 only and the other comprising categories 2 and 3. From this the relevant version of the generic error covariance matrix in (11) can be written as

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & \sigma_{23}(\eta) \\ 0 & \sigma_{23}(\eta) & \sigma^2 \end{pmatrix} \tag{62}$$

where $\sigma^2 = \frac{\Pi^2}{6}$, $\sigma_{23}(\eta) = \text{cov}(\varepsilon_2, \varepsilon_3)$ and Π denotes the well-known trigonometric constant. The covariance parameter $\sigma_{23}(\eta)$ is a non-closed-form function of η. As I noted in the discussion surrounding equation (37) above, a standardized measure of the covariance restrictiveness of this model by tracing the ranges of the off-diagonal elements can be obtained (which are correlations in the standardized model) of the following version of (37)[22]

$$D_{1\eta}[\xi_2 \quad \xi_3] = \begin{pmatrix} 1 & \frac{3\sigma_{23}(\eta)}{\Pi^2} + \frac{1}{2} \\ \frac{3\sigma_{23}(\eta)}{\Pi^2} + \frac{1}{2} & 1 \end{pmatrix} \tag{63}$$

and the following relevant versions of $D_{2\eta}[-\xi_2 \quad (\xi_3 - \xi_2)]$ and $D_{3\eta}[(\xi_2 - \xi_3) \quad -\xi_3]$

$$D_2[-\xi_2 \quad (\xi_3 - \xi_2)] = D_3[(\xi_2 - \xi_3) \quad -\xi_3]$$

[22] See Appendix F detailing the derivations of (63) and (64).

$$= \begin{pmatrix} 1 & \frac{1}{2}\left(1 - \frac{6\sigma_{23}(\eta)}{\Pi^2}\right)^{\frac{1}{2}} \\ \frac{1}{2}\left(1 - \frac{6\sigma_{23}(\eta)}{\Pi^2}\right)^{\frac{1}{2}} & 1 \end{pmatrix}. \qquad (64)$$

If I had a closed-form expression for the function $\sigma_{23}(\eta)$, I could evaluate the covariance restrictiveness of this NMNL model by varying η and thereby tracing the range of the relevant correlation functions in (63) and (64). This would reveal the "practical" correlation range for this NMNL model vis-a-vis the unrestricted "ideal" model in which the relevant correlations can take on any values in the interval $(-1, 1)$. Of particular interest in such an assessment would be evaluating the extent to which these relevant correlations can deviate from $\frac{1}{2}$ -- the correlation values for extremely covariance restricted models like MNL. By the same token, I could evaluate how nearly this model represents extremely covariance flexible models like MNP.

Although I do not have a closed form expression for $\sigma_{23}(\eta)$, I can plot its values numerically. I have that

$$\sigma_{23}(\eta) = \int_0^\infty \int_0^\infty \varepsilon_2 \varepsilon_3 \, f_{\varepsilon_2 \varepsilon_3}(\varepsilon_2, \varepsilon_3) \, d\varepsilon_2 \, d\varepsilon_3 - \gamma^2 \qquad (65)$$

where $f_{\varepsilon_2 \varepsilon_3}(\varepsilon_2, \varepsilon_3)$ denotes the joint marginal pdf of ε_2 and ε_3. I have shown that[23]

$$f_{\varepsilon_2 \varepsilon_3}(\varepsilon_2, \varepsilon_3)$$

[23] See Appendix G

$$= \exp\left(-\left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{1-\eta} - \frac{\varepsilon_2 + \varepsilon_3}{1-\eta} \right) \left(\left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{-2\eta} + \frac{\eta}{1-\eta} \left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{-\eta-1} \right) \tag{66}$$

Moreover, for a given value of η, (65) can be evaluated using bivariate Gauss-Legendre quadrature.

It is easy to see that the above NMNL model relaxes but does not entirely eliminate the IIA restriction. If $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$ is GEV distributed as in (58) then the category likelihoods are represented by the following versions of (12), (13) and (14)[24]

$$\pi_1(X_p^*, X_o, X_u; \tau) = G_1(-V_2^*, -V_3^*)$$

$$= \frac{1}{1 + \exp((1-\eta)I)} \tag{67}$$

$$\pi_2(X_p^*, X_o, X_u; \tau) = G_2\left(V_2^*, (V_2^* - V_3^*)\right)$$

$$= \frac{\exp\left(X_p^* \frac{\beta_{p2}^*}{(1-\eta)} + X_o \frac{\beta_{o2}^*}{(1-\eta)} + X_u \frac{\beta_{u2}^*}{(1-\eta)} - \eta I \right)}{1 + \exp((1-\eta)I)} \tag{68}$$

and

$$\pi_3(X_p^*, X_o, X_u; \tau) = G_3\left((V_3^* - V_3^*), V_3^*\right)$$

$$= \frac{\exp\left(X_p^* \frac{\beta_{p3}^*}{(1-\eta)} + X_o \frac{\beta_{o3}^*}{(1-\eta)} + X_u \frac{\beta_{u3}^*}{(1-\eta)} - \eta I \right)}{1 + \exp((1-\eta)I)} \tag{69}$$

where

[24] See Maddala (1983) Sections 3.7 and 3.8.

$$I = \ln\left(\exp\left(X_p^* \frac{\beta_{p2}^*}{(1-\eta)} + X_o \frac{\beta_{o2}^*}{(1-\eta)} + X_u \frac{\beta_{u2}^*}{(1-\eta)} - \eta I \right) + \exp\left(X_p^* \frac{\beta_{p3}^*}{(1-\eta)} + X_o \frac{\beta_{o3}^*}{(1-\eta)} + X_u \frac{\beta_{u3}^*}{(1-\eta)} - \eta I \right) \right) \tag{70}$$

The component "I" defined in expression (70) is called the *inclusive value*.

It is easy to see that the above NMNL model relaxes but does not entirely eliminate the IIA restriction. Consider the following three possible ratios that can be formed using the category probabilities in (67) through (69)

$$\frac{\pi_1(X_p, X_o, X_u; \tau)}{\pi_2(X_p, X_o, X_u; \tau)} = \frac{1}{\exp\left(X_p^* \frac{\beta_{p2}^*}{(1-\eta)} + X_o \frac{\beta_{o2}^*}{(1-\eta)} + X_u \frac{\beta_{u2}^*}{(1-\eta)} - \eta I \right)} \tag{71}$$

$$\frac{\pi_1(X_p, X_o, X_u; \tau)}{\pi_3(X_p, X_o, X_u; \tau)} = \frac{1}{\exp\left(X_p^* \frac{\beta_{p3}^*}{(1-\eta)} + X_o \frac{\beta_{o3}^*}{(1-\eta)} + X_u \frac{\beta_{u3}^*}{(1-\eta)} - \eta I \right)} \tag{72}$$

$$\frac{\pi_2(X_p, X_o, X_u; \tau)}{\pi_3(X_p, X_o, X_u; \tau)} = \frac{\exp\left(X_p^* \frac{\beta_{p2}^*}{(1-\eta)} + X_o \frac{\beta_{o2}^*}{(1-\eta)} + X_u \frac{\beta_{u2}^*}{(1-\eta)} - \eta I \right)}{\exp\left(X_p^* \frac{\beta_{p3}^*}{(1-\eta)} + X_o \frac{\beta_{o3}^*}{(1-\eta)} + X_u \frac{\beta_{u3}^*}{(1-\eta)} - \eta I \right)}. \tag{73}$$

It is clear that, IIA holds for pairs of categories within nest but not for pairs in which one alternative is within nest and the other is not in the nest.

The MLE defined by (43) and (44) is designed to account for endogeneity in the generic FPMO -- i.e., regardless of the particular MO specification [$\pi_j(X_p, X_o, X_u; \tau)$] that is implemented. In the remainder of the dissertation, I focus on the NMNL version of

the generic model [(43) and (44)] in which $\pi_j(X_p, X_o, X_u; \tau)$ is specified as in (67), (68) and (69) and $(X_u \mid W)$ in (38) is assumed to be standard normal distributed. As discussed earlier, this model/estimator, unlike MNL and MNP, promises to be both covariance flexible and computationally feasible. To investigate how well the NMNL (with endogenous treatment) estimator delivers on these margins, I conduct a simulation study in Chapter 4. A real data illustration of the implementation of the estimator is discussed in Chapter 5.

Chapter 4: Nested Logit with an Endogenous Treatment: Computational Issues and a Simulation Study

The focus in this chapter is on NMNL because it lies in the middle ground between the extremely covariance restrictive but computationally attractive MNL model and the unrestrictive but computationally burdensome MNP model. I implement (in Stata's matrix language [Mata]) the trinomial version of the NMNL with an endogenous probit causal treatment variable (NMNL-ET). The main challenge in this implementation is the development and incorporation of Mata code to numerically approximate the non-closed-form integrals required for the calculation of the log-likelihood function. Such integrals appear in the log-likelihood formulation as a means of accounting for the unobservable regression controls. Mata code implementing Gauss-Legendre quadrature was used for this purpose. I then develop and code (in Mata) a novel and relatively simple NMNL data simulator which exploits a unique property of the model. I also write Stata/Mata code for validating the accuracy of the data simulator. The simulator is found to be valid. Using data generated via the validated simulator and the newly developed NMNL-ET Mata software, I conduct a simulation study aimed at: a) assessing the computational feasibility of the NMNL-ET estimator and corresponding Stata/Mata software; b) empirically validating the theoretical consistency of the estimator; and c) investigating potential gains in estimation accuracy associated with the more computationally challenging NMNL-ET estimator vs. the analogous MNL specification (MNL with an endogenous treatment variable – MNL-ET).

For item (a), I first code (in Mata) a conventional optimization algorithm for the log-likelihood function and apply it to datasets generated using the aforementioned

simulator under a variety of sampling designs. For many of the sampling designs, the algorithm fails to converge. It is determined that allowing the inter-category covariance-related parameter to freely vary during the optimization iterations is the cause of the problem. To remedy this, the Stata/Mata code is revised to include a systematic line search (SLS) on the culprit parameter. At each iteration of the SLS, the revised algorithm optimizes the log-likelihood function at a fixed value of the parameter in question. Over the SLS iterations, the algorithm searches for and finds the value of that parameter at which the series of such optima is maximized. When this revised SLS algorithm is implemented, the convergence problems virtually disappear and the NMNL-ET estimator and corresponding code are shown to be feasible. Thus objective (a) is accomplished.

I then choose a particular sampling design and use it to generate a sequence of samples of increasing size. At each sample size, the deep parameters and the ATE are estimated. The latter is compared to the true values of the ATE for the chosen sampling design. As the sample size is increased, the estimated value of the ATE converges to the true value. Thus objective (b) is accomplished.

In establishing objective (b), I show that the NMNL-ET estimator of the ATE is consistent. A reasonable related question to ask is, "Is there substantive difference between this estimator and that obtained via the computationally simpler MNL-ET estimator which imposes severe covariance restrictions?" To shed light on the answer to this question, using a particular sampling design, I generate samples of size n = 100,000 and vary the value of the covariance-related parameter. NMNL-ET and MNL-ET estimates of the ATE are obtained for each sample. I find the differences between these estimates to be potentially quite large. This is in contrast to the results I obtain for ATE estimates obtained

by uncorrected (for endogeneity) NMNL and MNL estimators to samples simulated without endogeneity. Thus objective (c) is accomplished.

4.1 Simulating Data for the Simple Nested Logit Model with an Endogenous Treatment

In this section, I discuss the development of this simple NMNL data simulator and how this data simulator is validated.

In the NMNL model, the category probabilities can be decomposed into two simple binary logit probabilities. In addition, these two binary logit probabilities are effectively stochastically independent. The decomposition will simplify the otherwise difficult simulation task. Therefore, equations (67), (68) and (69) are written in the following alternative way.

Now let

$$B_2 = X_p^* \beta_{p2}^* + X_o \beta_{o2}^* + X_u \beta_{u2}^*$$

and

$$B_3 = X_p^* \beta_{p3}^* + X_o \beta_{o3}^* + X_u \beta_{u3}^*$$

Then

$$\pi_1(X_p^*, X_o, X_u; \tau) = \frac{1}{1 + \exp((1-\eta)I)}$$

$$= \Pr(\text{Alternative 1} \mid \text{Nest 1}) \times \Pr(\text{Nest 1})$$

$$= (1) \times \frac{1}{1 + \exp((1-\eta)I)} \qquad (74)$$

$$\pi_2(X_p^*, X_o, X_u; \tau) = \frac{\exp\left(\frac{B_2}{1-\eta}\right)\exp(-\eta I)}{1 + \exp((1-\eta)I)}$$

$$= \left(\frac{\exp\left(\frac{B_2}{1-\eta}\right)}{\exp(I)}\right)\frac{\exp(I)\exp(-\eta I)}{1 + \exp((1-\eta)I)}$$

$$= \Pr(\text{Alternative 2} \mid \text{Nest 2}) \times \Pr(\text{Nest 2})$$

$$= \left(\frac{\exp\left(\frac{B_2}{1-\eta}\right)}{\exp(I)}\right) \times \left(\frac{\exp((1-\eta)I)}{1 + \exp((1-\eta)I)}\right). \tag{75}$$

$$\pi_3(X_p^*, X_o, X_u; \tau) = \frac{\exp\left(\frac{B_3}{1-\eta}\right)\exp(-\eta I)}{1 + \exp((1-\eta)I)}.$$

$$= \Pr(\text{Alternative 3} \mid \text{Nest 2}) \times \Pr(\text{Nest 2})$$

$$= \left(\frac{\exp\left(\frac{B_3}{1-\eta}\right)}{\exp(I)}\right) \times \left(\frac{\exp((1-\eta)I)}{1 + \exp((1-\eta)I)}\right). \tag{76}$$

Note that the independence property portrayed in (74), (75), and (76) is unique to the NMNL model, and it is this independence property that allows the generation of the data (the outcome variable) in the following two simple steps.

-- **Step 1**

-- Generate the "within Nest 2 alternative" choice outcome

1 if Alternative 2

N2 =

0 if Alternative 3.

where

$$N2 = \mathcal{I}\left(\mathcal{U}_1[0, 1] > \frac{\exp\left(\frac{B_3}{1-\eta}\right)}{\exp(I)} \right) \tag{77}$$

where $\mathcal{I}(\quad)$ is the indicator function and $\mathcal{U}_1[0, 1]$ is a uniform random variable whose support is the unit interval.

-- Step 2

-- Generate the nest choice outcome

N =

1 if nest 2

0 if nest 1

where

$$N = \mathcal{I}\left(\mathcal{U}_2[0, 1] > \frac{1}{1 + \exp((1-\eta)I)} \right) \tag{78}$$

The simulated outcome then is generated as:

If	Then outcome =
N = 0	Alternative 1
N = 1 and N2 = 1	Alternative 2
N= 1 and N2 = 0	Alternative 3.

The variables needed to implement steps 1 and 2 are, X_o, W^+, X_u, and X_p. X_o and W^+ are generated following a uniform distribution with certain values of mean and

variance. In keeping with the assumption that will be made about X_u for the real data analysis, I assume it is standard normal, so it is generated following a standard normal distribution. The endogenous variable X_p is binary, so it is generated using the following indicator function: $X_p = I(W\alpha + X_u > 0)$. The true parameter coefficients are displayed in the second column of Table 3. In this dissertation, the outcome variable Y has three alternatives. In the simplest NMNL case, there are two nests: Nest 1 comprises the first alternative only, and Nest 2 comprises the second and the third alternatives.

To validate the accuracy of this simulation protocol, I need to have a valid data generator. Therefore the validation of the data generator is necessary for a simulation study. The validation includes two components: 1) Calculate the relative frequencies for (X_p, Y) pairs for a sample generated using fixed values of $W\alpha$ and $X_o\beta_o$; and 2) Calculate true probabilities for (X_p, Y) pairs for fixed values of $W\alpha$ and $X_o\beta_o$. If the relative frequencies for (X_p, Y) pairs are consistent with the true probabilities for (X_p, Y) pairs, then the data generator is valid. Table 1 and Table 2 show that the relative frequencies for (X_p, Y) pairs match the true probabilities for (X_p, Y) pairs very well. Therefore, the data generator is valid.

4.2 Implementation of the Simple Nested Logit Model

In this section, I discuss the implementation of the simple NMNL model with an endogenous treatment using the simulated data generated by the protocol detailed in section 4.1. Recall that, in general, the joint pdf of Y and X_p conditional on W is given in (43) and the corresponding relevant log-likelihood function is (44). Combining that generic log-

likelihood specification with the NMNL conditional category probabilities in (67) through (70) under the assumption that $(X_u \mid W)$ is standard normally distributed yields the specific form of the log-likelihood function that I use for the simulation and real data analysis

Consistent estimates of τ and α can be obtained by maximizing this log-likelihood function. However, optimization of the log-likelihood is not easy. I encounter convergence problems when attempting optimization with η (the inter-category covariance-related parameter) directly included one of maximization arguments. On the other hand, when η is held fixed and conventional optimization is conducted on the remaining elements of the parameter vector, those convergence problems disappear for the most part. I therefore decide to take a "systematic line search" (SLS) approach to optimizing the likelihood function.

I introduce the SLS algorithm into the estimator and develop corresponding Mata code for maximizing the log-likelihood function. Most SLS algorithms are designed to find the maximum of a unimodal function of one variable whose domain is a bounded interval of the real line. In the present context, the one-variable function to be maximized is that which represents the profile of optimized values of the log-likelihood function (optimized, that is, with respect to all parameters except η). The SLS algorithm begins with a chosen interval in which η is expected to reside. Denote this interval as [a, c] and at a particular iteration of the algorithm define the "interval of certainty." At the m^{th} iteration, all SLS algorithms are designed to reduce the length of the interval of certainty and terminate if the interval of certainty is smaller in length than some specified tolerance. The typical SLS algorithm begins by choosing a point b within [a, c] thus forming a triple of "bracketing

points" such that q(a)< q(b) > q(c). Note that q( ) is the log-likelihood function discussed in (44). Next the interval of certainty is reduced in length by choosing a fourth point τ and comparing the value of the function at that point to the values q(a), q(b), and q(c). There are four different cases in terms of the locations of the fourth point τ, which leads to new different bracketing triples. To complete the specification of the algorithm, two issues must be settled: 1) whether the new "fourth point" should be placed to the right or to the left of the interior point in the bracketing triple; and 2) the distance from the interior point at which to place the new point. In the present context, q(η) is the profile maximum value (with respect to the remaining parameters), written as a function of η. To summarize, the MLE algorithm that is used comprises two steps: 1) For a given (fixed) value of the inter-category covariance-related parameter, maximize the likelihood function with respect to all other parameters of the model (the αs and βs); and 2) Conduct a SLS across a relevant range of η values to find the η-maximizer of all maxima obtained in the first step. The SLS algorithm successfully resolving the convergence problems that are encountered with conventional optimization in the estimation of the NMNL model with an endogenous treatment variable.

4.3 Simulation Study

There are three objectives in this simulation study: a) assessing the computational feasibility of the NMNL-ET estimator and corresponding Stata/Mata software—I want to show that this NMNL-ET estimator that includes the SLS actually works well; b) empirically validating the theoretical consistency of the estimator—I will show that as the sample size increases, the estimated deep parameters and average treatment effect approach to the true values; and c) investigating potential gains in estimation accuracy associated

with the more computationally challenging NMNL-ET estimator vs. the analogous MNL specification (MNL with an endogenous treatment variable – MNL-ET)—I will demonstrate that it is important to accommodate the endogeneity and inter-category covariance simultaneously, which I will show by measuring the difference between the ATEs for NMNL-ET and MNL-ET estimators.

For the first two objectives, I choose η=.75 and the values of the other parameters are displayed in the second column of Table 3. Table 3 and Table 4 illustrate that as I increase the sample size, both the estimated deep parameters and ATEs get closer to the true values. The results in Tables 3 and 4 serve to empirically validate the consistency of the estimators.

To achieve the third objective, I choose a sampling design with sample size equal to 100,000 and the other parameters displayed in the second column of Table 3. I vary the value of η in each single sample to investigate the potential gains in estimation accuracy. When the covariance-related parameter η falls in the range [0, 1), the model conforms with the random utility model detailed in equations (9) and (10). As η is close to 0, the correlation between the alternatives within the same nest is close to zero and the NMNL model will become the MNL model. As η is close to 1, the correlation between the alternatives within the same nest gets larger. I want to know whether having a correct specification of the model matters as I change the value of η. Other than the range of [0, 1], it is possible that η may fall into other ranges (i.e., η could be greater than 1 or less than 0). When η is outside the unit interval, even though a model might not conform to the random utility model the model is still a legitimate probability model. In addition, in the real data analysis that will be discussed in chapter 5, the estimated η is greater than 1, so

it is important for us to understand when η is outside the unit interval, how the model would perform. I also want to know if endogeneity affects the choice of the model specifications. To evaluate how the change of the value of η will affect the choice of model specification, I divide the values of η into three regions: less than 0, [0, 1), and greater than 1. I thus discuss the simulation results within the three regions.

I first start with model without endogeneity. To compare the NMNL with MNL, I use the data generator discussed in section 4.1 to generate a simulated dataset with sample size equal to 100,000. Note that I remove X_u in the data generating process so that I will have a simulated dataset without endogeneity.

Table 5 illustrates that when η is in the unit interval, the difference between the ATEs for NMNL and MNL are small though it shows a pattern of widening difference as η is close to 1. This pattern is consistent with the theory because as η is close to 1, the correlation between alternatives within the nest is getting larger and, as a result, the use of the NMNL specification may make a difference. However, it is interesting to see that the difference is quite small. This indicates that the NMNL model may not yield results that are very different from those obtained using the MNL model when there is no endogeneity.

As discussed in previous studies, η must be within the unit interval so that the model is consistent with utility-maximizing behavior. Mathematically, however, it is possible that η could be greater than one or less than zero. I therefore also discuss the difference between the ATEs for NMNL and MNL when η is outside the unit interval.

Table 6 shows that when η is greater than 1, the difference between the ATEs for NMNL and MNL gradually decreases as η increases and is the largest when η is close to

1. However, the difference between the ATEs for NMNL and MNL is not very large. Table 7 illustrates that the difference decreases as η decreases when η is less than 0. The difference is also quite small.

Tables 5, 6 and 7 demonstrate that the NMNL model specification may not provide different ATE estimates than the MNL model when there is no endogeneity across the three regions of the domain of η. This means that even though the true model is NMNL, the MNL appears to provide quite accurate ATE estimates.

I now discuss the case when there is endogeneity. I use the data generator discussed in section 4.1 to generate a simulated dataset with endogeneity. The sample size is 100,000.Table 8 illustrates that the difference between the ATEs for NMNL-ET and MNL-ET increases as η increases within the unit interval. This is similar to the pattern in the case when there is no endogeneity but the differences in the present endogeneity case is much bigger. Specifically, when η = .95, the difference between the ATES for NMNL-ET and MNL-ET in three categories are -54.62%, 3.98%, and 25.19%, while the differences are -0.6%, -0.52% and 0.49% in the case when there is no endogeneity. This suggests that it is important to have a correct model specification when there is endogeneity though it may not be important to do so when there is no endogeneity.

Table 9 shows the difference between the ATEs for NMNL-ET and MNL-ET when η is greater than 1. I also find that the difference is the largest when η is close to 1. For instance, when η = 1.5, the difference between the ATEs for NMNL-ET and MNL-ET in the three outcome categories are -79.73%, -615.61% and 13.89%, while the differences are -0.06%, -4.03%, and -1.09% in the case when there is no endogeneity. I also find that as η increases, the difference becomes smaller. For example, when η = 7, the difference

between NMNL-ET and MNL-ET is quite small. It suggests that when η is greater than 1 and close to 1, it is important to have a correct model specification but when η is greater than 1 and gets larger, there may not be much difference between NMNL-ET and MNL-ET.

Table 10 illustrates the difference when η is less than zero. I also find a pattern similar to that which I found in the case when there is no endogeneity. However, the difference is much bigger when there is endogeneity, especially when η is close to zero, the difference is the largest. However, as η decreases, the difference becomes smaller.

Tables 8, 9, and 10 suggest that NMNL-ET may provide different estimates than MNL-ET when there is endogeneity. It is important to have a correct model specification when endogeneity is present. The evidence from this simulation study is also informative for the real data analysis. In real data analysis, if there is no endogeneity, both NMNL and MNL should provide similar results. If there is endogeneity, having a correct model specification is important and it is crucial to use the NMNL-ET instead of the MNL-ET.

Chapter 5: An Illustration: The Effect of Substance Abuse/Dependence on Employment Status

Substance abuse/dependence has been a serious societal and economic problem for decades. It not only causes trouble to individuals and families, but also leads to economic loss for the entire society. Substance abuse/dependence is usually associated with more crime, and thus increases the incarceration costs. A lot of health care resources are needed to treat patients with substance abuse/dependence. Substance abuse/dependence also leads to productivity loss, which constitutes a major part of the economic loss. Substance abuse/dependence decreases an individual's work productivity by affecting his/her work performance, increasing absenteeism, work-related injuries, and risk of being fired or resigning from a job (Blum et al. 1993, Ames et al. 1997, Hoffmann & Larison 1999, Webb et al. 1994). The recent epidemic of opioid use all over the U.S. shows that it is important to understand the impact of substance abuse/dependence.

The economic cost brought by substance abuse/dependence is substantial and productivity loss is the largest. The National Institute on Drug Abuse (NIDA, 2017) estimates that substance abuse costs the country $740 billion a year. The total economic cost of substance abuse/dependence can be further broken down into three categories: tobacco, alcohol, and illicit drugs. This dissertation focuses on the costs of alcohol and illicit drugs. Sacks et al. (2015) estimates excessive drinking costs the U.S. $250 billion. The lost productivity cost was $179 billion. The National Drug Intelligence Center (NDIC, 2011) estimates the cost of illicit drug use was $193 billion and drug abuse costs the U.S. $120 billion per year in lost productivity. The most significant portion of productivity loss caused by illicit drug abuse is reduced labor participation which can be as large as $49

billion. From 2000 to 2015, deaths from opioid overdose increased dramatically, and the opioid overdose has become a serious public health crisis all over the country. Prescription opioid misuse is the driving factor of deaths caused by opioid overdose, bringing a heavy burden to society. Florence et al. (2016) estimates that the total economic burden imposed by prescription opioid abuse is $78.5 billion. The productivity loss attributable to the opioid abuse is about $41.9 billion. Krueger (2017) finds that the labor force participation rate is lower and fell more during the 2000s in areas of this country that have a higher volume of opioid medication prescribed per capita than in other areas.

It is crucial for us to better understand the effect substance abuse/dependence has on society and how large the potential treatment effect would be if substance abuse/dependence were treated or prevented.

Most of previous studies of the effects of substance abuse/dependence on employment status have considered alcohol only.[25] Most studies have focused on how problem drinking affects employment status and the results are mixed. Some studies find that problem drinking increases the probability of unemployment (Mullahy & Sindelar ,1996; Terza , 2002; Booth & Feng ,2002; MacDonald & Shields , 2004; Johansson et al. , 2007) , while others find a positive relationship (Feng et al. ,2001; Asgeirsdottir & McGeary , 2009 ; Balsa & French , 2010). On the contrary, my definition of substance abuse/dependence includes alcohol and a number of illicit drugs (cannabis, cocaine, etc).

I illustrate my novel approach to modeling and estimation by examining the treatment effect of substance abuse/dependence on employment status. This application is well suited for my approach for two reasons: first, it is a multinomial outcome model (out

[25]The exceptions are Buchmueller and Zuvekas (1998) and Bray et al. (2000)

of the labor force, unemployed, or employed) where covariance flexibility among the alternatives is likely to be needed; second, the causal treatment variable (substance abuse) is binary. I argue that individuals who are unemployed or employed are actively involved in the labor market and have unobserved characteristics that are more aligned with each other than those who are out of the labor force. Therefore, it would not seem appropriate to assume independence across these three alternatives. In the relationship between substance abuse/dependence and employment status, endogeneity can be present and thus precludes the use of conventional regression methods to obtain a credible causal estimate of the effect of substance abuse/dependence on employment status. For example, unobserved factors that determine both employment status and substance abuse/dependence status can confound the effect. These include psychiatric problems, chronic health problems, injuries, physical pain, problems with friends and family, and stress (Mullahy & Sindelar, 1996).

To simultaneously account for covariance flexibility (the IIA issue) and endogeneity, I use the NMNL-ET maximum likelihood estimator as discussed in Chapter 4 with a NMNL specification for the category likelihoods (conditional on the observable and unobservable confounders; X_o and X_u, respectively) and a probit specification for the binary endogenous treatment variable. In this NMNL-ET model, therefore there are two nests. The first nest is a degenerate nest and includes only one alternative (out of the labor force). The second nest has two alternatives (unemployed and employed). Figure 1 displays the nest structure of the employment status.

5.1 Model Specification

Under standard normality of $(X_u \mid W)$, (43) can be written as

$$
\begin{aligned}
&f_{(Y,X_p|W)}(Y, X_p, W; \tau, \alpha) \\
&\quad = \left[\int_{-W\alpha}^{\infty} \prod_{j=1}^{J} \pi_j(X_p, X_o, X_u; \tau)^{Y_j} \varphi(X_u)\, dX_u \right]^{X_p} \\
&\quad \times \left[\int_{-\infty}^{-W\alpha} \prod_{j=1}^{J} \pi_j(X_p, X_o, X_u; \tau)^{Y_j} \varphi(X_u)\, dX_u \right]^{1-X_p} \qquad (85)
\end{aligned}
$$

where $\varphi(\ )$ denotes the standard normal pdf and the $\pi_j(\ )$ functions are specified as in (67) through (70). The estimate of the average treatment effect at category j can be written as the following version of (6)

$$
\widehat{ATE_j} = \sum_{i=1}^{n} \frac{1}{n} \left\{ \int_{-\infty}^{\infty} \left[\pi_j(0, X_o, X_u; \hat{\tau}) - \pi_j(1, X_o, X_u; \hat{\tau}) \right] \right\} \varphi(X_u)\, dX_u \qquad (86)
$$

where $\hat{\tau}$ is the MLE of τ obtained by maximizing the version of (43) that implements (85).

5.2 The Data

The data comes from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). I use wave 1 (NESARC I) of the data which was conducted from 2001 to 2002. NESARC is composed of data for adults aged 18 years and older taken from a random sample of U.S. households. NESARC I interviewed 43,093 individuals. I restricted the estimation sample to respondents aged 25 to 59; respondents who are full-time homemakers, permanently disabled, retired, and in school full-time or part-time are all excluded; I also exclude females from the sample. The final sample size is 10,551.

The endogenous variable X_p is the indicator of substance abuse/dependence status. An individual is defined as a substance abuser or dependent if he meets the DSM-IV criteria

for current abuse and/or dependence in the past 12 months.[26] The qualitative employment status variable Y_j has three categories (j = 1, 2, 3) where j =1 denotes the "out of the labor force" category, j = 2 denotes "unemployed," and j = 3 denotes "employed." The vector of observable confounders includes race, age, region of residence, living in an urban setting, education level, the quarter in which the interview took place, the number of problematic health conditions from which an individual currently suffers, and the state level unemployment rate. Following Mullahy and Sindelar (1996) and Terza (2002), I use the following instrumental variables (IV): state beer excise tax rate per gallon, state cigarette excise tax rates per pack, quadratics of the two taxes, and whether the respondent's biological father or mother is an alcoholic or problem drinker. The full list of the variables and their definitions are given in Table 12. Table 13 and Table 14 present the descriptive statistics.

5.3 Estimation and Results

I check the strength of the instrumental variables by using a joint likelihood ratio test. The results of the likelihood ratio test suggest that the IV are strong (see Table 15). I also find, however, no statistical evidence that a model that allows for ICF is needed (the null hypothesis that there is zero inter-category covariance is not rejected) (see Table 16). This means that the correlation between the category "unemployed" and "employed" is not strong. I conduct a likelihood ratio test of the null hypothesis that substance abuse/dependence is exogenous (H_o: $\beta_{u2} = \beta_{u3} = 0$) (see Table 17). Exogeneity is rejected at a 5% significance level. This shows that endogeneity is present in the relationship

[26]Details of the DSM-IV criteria are given in Table 11.

between substance abuse/dependence and employment. In the present modeling context, the above statistical tests indicate that the appropriate model may be MNL-ET (endogeneity corrected MNL).

The deep parameters estimation results can be found in Table 18. The average treatment effect for the three employment categories, as estimated in (84) in NMNL-ET, are reported in column 12 of Table 19. Note that most of the deep parameter coefficients estimated by the NMNL-ET are not significant, but the estimated ATEs are significant.

Table 19 displays ATE estimation results obtained from different methods. The NMNL-ET is the only method that simultaneously accommodates both endogeneity and ICF. The NMNL accommodates ICF but not endogeneity. The MNL-ET tackles the endogeneity issue but ignores ICF. The MNL-ET accommodates neither endogeneity nor ICF. I also provide results obtained from linear models: OLS and linear IV to demonstrate the estimation difference between linear models and nonlinear models. The linear IV addresses endogeneity, while the OLS does not.

The comparisons between OLS and linear IV models, between MNL and MNL-ET models, as well as between NMNL and NMNL-ET models in Table 19, provide strong evidence that it is important to address endogeneity when estimating the treatment effect. The models that do not address endogeneity are downward biased in magnitude. In models that ignore endogeneity, I find the treatment effects in all three categories are quite small compared to other models that address endogeneity. This is consistent with prior studies of substance abuse/dependence and employment where results from using the IV method are usually significantly larger. When compared with linear IV, results obtained from the NMNL-ET are larger in magnitude and more significant. It is interesting to see that the

NMNL-ET yields different results than the MNL-ET, though the likelihood ratio test suggests that the NMNL-ET and MNL-ET is not significantly different. After I group "unemployed" and "employed" into a nest and relax the IIA assumption, I find a significant and positive treatment effect on "employment," a significant and negative treatment effect on "unemployed" and significant and negative treatment effect on "out of the labor force." The MNL-ET, however, does not produce a significant treatment effect on "out of the labor force."

In sum, the statistical tests indicate that the appropriate model may be MNL-ET. The MNL-ET-based ATE estimates imply that substance abuse/dependence treatment can both significantly reduce the probability of being unemployed and significantly increase the probability of being employed. The latter result is consistent with that which was obtained by Terza (2002) using a different dataset. The estimation did, however, reveal one somewhat puzzling result: the NMNL-ET ATE estimates were all statistically significant with intuitively correct signs, despite the fact that neither of the NMNL-ET estimated deep coefficient parameters for the substance abuse variables were significant.

Chapter 6: Summary, Discussion and Conclusions

This dissertation focuses on a commonly encountered and very policy-relevant class of empirical contexts that has heretofore been virtually ignored in the econometric methodological literature – specification and estimation of endogenous binary treatment effects on multinomial outcomes with non-null inter-category covariance. This work contributes to the literature in the following specific ways. First, specification, estimation and inference for such models are placed in a PO framework, thereby making explicit the requisite conditions for causal interpretability of the treatment effect parameters and estimates. Secondly, the general formulation of the log-likelihood function for estimation of the deep parameters of the model is derived and is shown to encompass three important specific versions that cover the inter-category covariance and computational feasibility spectra. As part of this discussion, a standardized measure of ICF is developed. Third, a simulation study of one of the three highlighted models (NMNL-ET) is conducted. This version of the model, unlike the other two, affords both ICF and computational feasibility. The simulation study validates that: a) the model is computationally feasible; b) the estimator is consistent; and c) there are clear potential gains in accuracy with NMNL-ET vs. MNL-ET (the computationally feasible but inter-category restrictive alternative). Finally, for the purpose of illustration, the NMNL-ET model was applied to real data in estimating the potentially endogenous effect of substance abuse/dependence treatment on employment status.

This dissertation should be viewed as a first step toward the development and application of models for the estimation of endogenous causal effects in multinomial outcome contexts. Subsequent topics for future research include: extending the trinomial

NMNL-ET model and code to cases involving four or more categories; extending that model to accommodate multiple nests; exploring possible feasible inter-category covariance flexible MNP specifications; and investigating multinomial modeling strategies beyond the three considered here. There are many other possible extensions of this work.

Chapter 7: Tables and Figure

7.1 Tables

Table 1: True Probabilities

	$X_p = 0$	$X_p = 1$
$Y_1 = 1$	.0027884579	.0359185233
$Y_2 = 1$	.0632400465	.239131992
$Y_3 = 1$	.0007786969	.6581422834

Table 2: Relative Frequencies

	$X_p = 0$	$X_p = 1$
$Y_1 = 1$	.00277565	.03592185
$Y_2 = 1$	.06312685	.2390688
$Y_3 = 1$	.0007804	.65832645

Table 3: Deep Parameters True Values and Estimates

		Sample Size						
Parameters	True Values	1K	5K	10K	25K	50K	100K	500K
α_o	.25	.26	.25	.26	.25	.25	.25	.25
α_{w^+}	.5	.48	.51	.49	.51	.50	.50	.50
α_c	.75	.71	.75	.76	.76	.75	.75	.75
β_{p2}	1	.49	.72	.94	.94	.97	.96	.96
β_{o2}	.25	.17	.28	.22	.32	.28	.26	.27
β_{c2}	1	1.48	1.03	.94	1.22	1.11	1.06	1.06
β_{u2}	-1	-.46	-.97	-1.08	-.83	-.89	-.94	-.97
β_{p3}	.25	-.43	-.16	-.11	.38	.31	.21	.24
β_{o3}	.75	.85	.98	.89	.71	.72	.76	.75
β_{c3}	2	2.72	2.25	2.28	1.99	2.01	2.05	2.02
β_{u3}	.5	1.55	1.11	.88	.33	.45	.56	.48
η	.75	.75	.64	.66	.81	.79	.75	.77

Table 4: True ATEs and Estimated ATEs

Category / n	Category 1			Category 2			Category 3		
Sample Size	True ATE	NMNL-ET ATE	Bias (%)	True ATE	NMNL-ET ATE	Bias (%)	True ATE	NMNL-ET ATE	Bias (%)
1000	.0366408	.0052512	-85.67%	-.1354145	-.1175852	-13.17%	.0987737	.112334	13.73%
5000	.0366408	.0179354	-51.05%	-.1354145	-.1139384	-15.86%	.0987737	.096003	-2.81%
10000	.0366408	.0220933	-39.70%	-.1354145	-.140299	3.61%	.0987737	.1182057	19.67%
25000	.0366408	.0407042	11.09%	-.1354145	-.1335602	-1.37%	.0987737	.092856	-5.99%
50000	.0366408	.0381962	4.24%	-.1354145	-.1341807	-0.91%	.0987737	.0959845	-2.82%
100000	.0366408	.0340311	-7.12%	-.1354145	-.1340003	-1.04%	.0987737	.0999692	1.21%
500000	.0366408	.035323	-3.60%	-.1354145	-.1339534	-1.08%	.0987737	.0986304	-0.15%

Table 5: Proportional Difference between the ATEs for NMNL and MNL

($0 \leq \eta < 1$, No Endogeneity)

	% difference		
Category / η	Category 1	Category 2	Category 3
.05	-0.001%	-0.004%	-0.004%
.25	-0.027%	-0.08%	-0.095%
.5	-0.11%	-0.10%	-0.10%
.75	-0.32%	-0.21%	-0.17%
.95	-0.6%	-0.52%	0.49%

Table 6: Proportional Difference between the ATEs for NMNL and MNL

($\eta > 1$, No Endogeneity)

	% difference		
Category / η	Category 1	Category 2	Category 3
1.5	-0.06%	-4.03%	-1.09%
2.0	-0.005%	-4.31%	-0.50%
4.0	0.026%	-0.22%	0.16%
7.0	-0.003%	-1.42%	1.45%
10.0	-0.003%	-0.33%	0.34%

Table 7: Proportional Difference between the ATEs for NMNL and MNL

($\eta < 0$, No Endogeneity)

	% difference		
Category / η	Category 1	Category 2	Category 3
-1.5	0.015%	0.095%	0.12%
-2.0	0.013%	0.083%	0.10%
-4.0	0.011%	0.041%	0.043%
-7.0	0%	-0.11%	-0.11%
-10.0	0%	0.019%	0.02%

Table 8: Proportional Difference between the ATEs for NMNL-ET and MNL-ET

($0 \leq \eta < 1$, Endogeneity)

	% difference		
Category / η	Category 1	Category 2	Category 3
.05	- 2.99%	0.04%	0.94%
.25	-12.44%	0.2%	4.06%
.5	-35%	0.86%	12.1%
.75	-49.6%	2.06%	19.54%
.95	-54.62%	3.98%	25.19%

Table 9: Proportional Difference between the ATEs for NMNL-ET and MNL-ET

($\eta > 1$, Endogeneity)

	% difference		
Category / η	Category 1	Category 2	Category 3
1.5	-79.73%	-615.61%	13.89%
2.0	-60.95%	-318.55%	7.75%
4.0	-15.67%	-31.36%	-3.33%
7.0	-2.5%	-2.46%	-2.54%
10.0	-1.59%	-3.78%	0.5%

Table 10: Proportional Difference between the ATEs for NMNL-ET and MNL-ET

($\eta < 0$, Endogeneity)

	% difference		
Category / η	Category 1	Category 2	Category 3
-1.5	16.99%	0.41%	-4.11%
-2.0	15.79%	0.78%	-2.44%
-4.0	10.09%	0.65%	-0.31%
-7.0	5.13%	0.12%	0.11%

Table 11: The DSM-IV Criteria

The American Psychiatric Association states that addiction is a maladaptive pattern of substance use, leading to clinically significant impairment or distress, as manifested by three (or more) of the following, occurring at any time in the same 12 month period.
1. Tolerance, as defined by either of the following: A. A need for markedly increased amounts of the substance to achieve intoxication or desired effect. B. Markedly diminished effect with continued use of the same amount of the substance 2. Withdrawal, as manifested by either of the following: A. The characteristic withdrawal syndrome for the substance B. The same (or a closely related) substance is taken to relieve or avoid withdrawal symptoms 3. The substance is often taken in larger amounts or over a longer period than was intended 4. There is a persistent desire or unsuccessful efforts to cut down or control substance use 5. A great deal of time is spent in activities necessary to obtain the substance (e.g., visiting multiple doctors or driving long distances), use the substance (e.g., chain smoking), or recover from its effects 6. Important social, occupational, or recreational activities are given up or reduced because of substance use 7. The substance use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by the substance (e.g., current cocaine use despite recognition of cocaine-induced depression, or continued drinking despite recognition that an ulcer was made worse by alcohol consumption)
The preceding was reprinted form Landry (1997), Exhibit 2.1.

Table 12: Variable Definitions

Dependent Variables

olf: Y_1 = 1 if out of the labor force, 0 otherwise.

unemp: Y_2 = 1 if unemployed, 0 otherwise.

emp: Y_3 = 1 if employed, 0 otherwise

Endogenous Variable

cdrgab : X_p**:** 1 if substance abuse or dependent, 0 otherwise

Variables Included in X_0

numphc: count of the number of health conditions that caused problems in the past year
hhsize: count variable equal to the number of people in the household
married: 1 if married, 0 otherwise
black: 1 if black, 0 otherwise
asian: 1 if asian, 0 otherwise
hispanic: 1 if hispanic, 0 otherwise
other: 1 if other ethnic groups, 0 otherwise
nohschool: 1 if not high school graduate, 0 otherwise
hschool: 1 if a high school graduate only, 0 otherwise
somecol: 1 if some post secondary school education, 0 otherwise
college: 1 if a college graduate or beyond, 0 otherwise
midwest, south, west: 1 if resides in that region, 0 otherwise (northeast excluded)
urban: 1 if living in an urban setting, 0 otherwise
qtrdum2, qtrdum3, qtrdum4: 1 if interview was conducted in that quarter, 0 otherwise (first quarter excluded)
ur: state unemployment rate
age: age in years
agesq: age squared

Instrumental Variables (Included in W^+ **Only)**

bfdum: 1 if biological father was an alcoholic, 0 otherwise
bmdum: 1 if biological mother was an alcoholic, 0 otherwise
alctax: state level alcohol tax
alctaxsq: alctax squared
cigtax: state level cigarette tax
cigtaxsq: cigtax squared

Table 13: Descriptive Statistics

Variable	Mean	SD
olf	0.01	0.10
unemp	0.04	0.21
emp	0.95	0.23
cdrgab	0.14	0.34
numphc	0.39	0.77
hhsize	2.71	1.54
married	0.63	0.48
white	0.58	0.49
black	0.16	0.36
asian	0.04	0.18
hisp	0.21	0.41
other	0.02	0.13
nohschool	0.13	0.33
hschool	0.27	0.44
somecol	0.29	0.46
college	0.31	0.46
northeast	0.19	0.39
midwest	0.21	0.41
south	0.36	0.48
west	0.24	0.43
urban	0.84	0.37
qtrdum1	0.15	0.36
qtrdum2	0.00	0.06
qtrdum3	0.40	0.49
qtrdum4	0.44	0.50
ur	0.06	0.01
age	41.03	9.13
agesq	1,766.74	762.75
bfdum	0.18	0.39
bmdum	0.05	0.22
alctax	0.25	0.22
cigtax	0.75	0.47
alctaxsq	0.11	0.23
cigtaxsq	0.78	0.78
n	10,551	10,551

Table 14: Descriptive Statistics by Substance Abuse Status

Variable	Non-abuser	Abuser
olf	0.01	0.01
	(0.10)	(0.11)
unemp	0.04	0.06
	(0.20)	(0.24)
emp	0.95	0.93
	(0.22)	(0.26)
numphc	0.37	0.49
	(0.75)	(0.89)
hhsize	2.75	2.44
	(1.55)	(1.51)
married	0.65	0.49
	(0.48)	(0.50)
white	0.57	0.63
	(0.50)	(0.48)
black	0.16	0.15
	(0.37)	(0.36)
asian	0.04	0.02
	(0.19)	(0.13)
hisp	0.22	0.18
	(0.41)	(0.38)
other	0.02	0.02
	(0.12)	(0.14)
nohschool	0.13	0.12
	(0.34)	(0.32)
hschool	0.26	0.30
	(0.44)	(0.46)
somecol	0.29	0.32
	(0.45)	(0.47)
college	0.32	0.26
	(0.47)	(0.44)
northeast	0.19	0.18
	(0.40)	(0.39)
midwest	0.20	0.26
	(0.40)	(0.44)
south	0.37	0.31
	(0.48)	(0.46)
west	0.24	0.24
	(0.43)	(0.43)
urban	0.84	0.82
	(0.37)	(0.38)
qtrdum1	0.16	0.13
	(0.36)	(0.34)

qtrdum2	0.00	0.01
	(0.06)	(0.08)
qtrdum3	0.40	0.42
	(0.49)	(0.49)
qtrdum4	0.44	0.44
	(0.50)	(0.50)
ur	0.06	0.06
	(0.01)	(0.01)
age	41.40	38.69
	(9.10)	(8.95)
agesq	1,796.97	1,577.15
	(764.50)	(723.71)
bfdum	0.17	0.27
	(0.37)	(0.44)
bmdum	0.04	0.10
	(0.20)	(0.29)
alctax	0.26	0.23
	(0.22)	(0.20)
cigtax	0.75	0.76
	(0.47)	(0.45)
alctaxsq	0.12	0.09
	(0.24)	(0.20)
cigtaxsq	0.78	0.77
	(0.78)	(0.75)
n	9,100	1,451

Note:The numbers in parentheses are standard deviations

Table 15: Joint Likelihood Ratio Test of Instrumental Variables

LR chi2(6)	84.57
p-value	<0.001

Note: H0: The first stage probit regression without IVs is nested in probit regression with IVs

Table 16: Test of Inter-category Covariance-related Coefficient

LR chi2(2)	0.61
p-value	0.736

Note: H0: $\eta = 0$

Table 17: Endogeneity Test in NMNL-ET Model

LR chi2(2)	8.36
p-value	0.015

Note: H0: $\beta_{u2} = \beta_{u3} = 0$

Table 18: NMNL-ET Deep Parameters Estimation Results

Variable	Unemployed		Employed	
	Estimate	t-statistics	Estimate	t-statistics
cdrgab	-17.84	-.99	-2.29	-1.30
numphc	-.71	-1.35	-.19	-1.48
hhsize	-.37	-.99	-.00	-.04
married	4.22	1.15	.05	.07
black	-2.81	-.97	.81	.60
asian	-3.09	-.90	.72	.65
hisp	-.81	-.63	.33	.61
other	-.53	-.22	1.28	.85
hschool	2.89	1.24	.56	1.46
somecol	3.96	1.25	.87	1.70
college	5.95	1.28	1.28	2.01**
midwest	1.51	.85	-.18	-.31
south	1.15	.94	-.06	-.11
west	.99	.79	-.05	-.11
urban	-.18	-.17	.71	1.38
qtrdum2	-4.34	-1.10	-.42	-.19
qtrdum3	.57	.56	-.41	-.95
qtrdum4	1.04	.88	-.18	-.43
ur	-73.15	-1.03	-14.14	-.80
age	.15	.54	.01	-.06
agesq	-.00	-.70	.00	.02
constant	24.47	1.01	6.13	1.94*
Xu	7.92	1.00	.84	1.08
η	6.22	1.13	6.22	1.13

Note: *** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

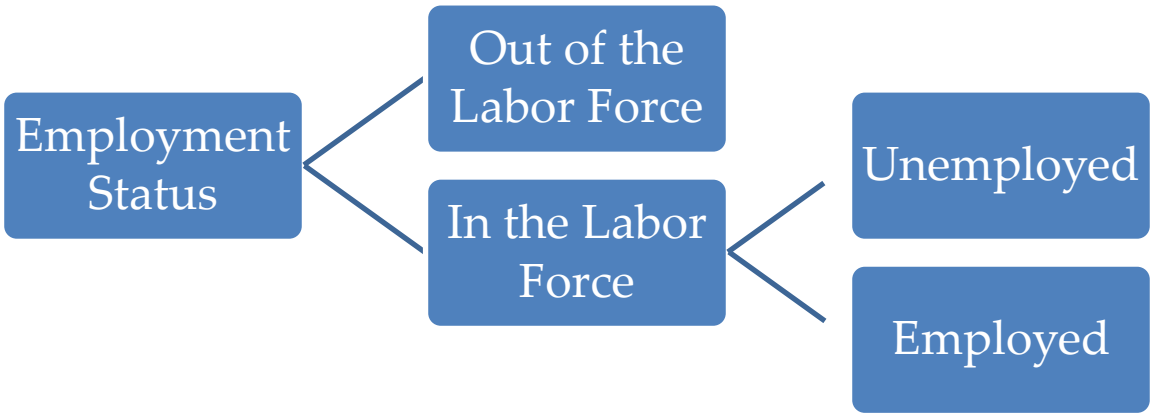
Table 19: Average Treatment Effect of Substance Abuse on the Employment Status

	OLS		Linear IV		MNL		MNL-ET		NMNL		NMNL-ET	
	ATE	t-statistics	ATE	t-statistics	ATE	t-statistics	ATE	t-statistics	ATE	t-statistics	ATE	t-statistics
Out of Labor Force	-0.001	-0.41	-0.178*	-1.75	-0.001	-0.40	-0.032	-0.59	-0.001	1.96	-0.264*	-1.67
Unemployed	-0.013*	-1.90	0.032	0.46	-0.012*	-1.92	-0.209*	-1.78	-0.011	-1.44	-0.107***	-2.92
Employed	0.014*	1.92	0.146*	1.96	0.013*	1.93	0.241**	2.26	0.012	1.12	0.371***	2.54
Sample size	10,551	10,551	10,551	10,551	10,551	10,551	10,551	10,551	10,551	10,551	10,551	10,551

Note: *** Significant at 1% level, ** Significant at 5% level, * Significant at 10% level

7.2 Figure

Figure 1: The Nest Structure of the Employment Status



Appendix A

Terza (2016a, b) explains the way to obtain the asymptotic standard error for the commonly encountered two-stage optimization estimator (2SOE). In this dissertation, we can view the average treatment effect (ATE) as a 2SOE. 2SOE is the sample mean of a parametric data transformation of the following general form

$$\hat{\gamma} = \sum_{i=1}^{n} \frac{g(\hat{\tau}, X_i)}{n} \tag{A-1}$$

which is, under general conditions, consistent for

$$\gamma = E[g(\tau, X)] \tag{A-2}$$

where g( ) is a specified (known) function; $\hat{\tau}$ is a estimate of τ, a vector of "deep" parameters and X_i denotes a vector of observed data on X, a vector of observable variates, for the ith member of a sample of size n (i = 1, ..., n). Specifically, $\hat{\gamma}$ represents $\widehat{ATE}_j$, g( ) in (A-1) represents $E_{X_u}\left[\pi_j(0, X_o, X_u; \hat{\tau}) - \pi_j(1, X_o, X_u; \hat{\tau})\right]$ in equation (7). X_i represents X_p, X_o, X_u in equation (7).

The asymptotic property of $\hat{\gamma}$ (in this dissertation it is the average treatment effect) is as the following

$$\sqrt{\frac{n}{\widehat{a\,var}(\hat{\gamma})}}\left(\hat{\gamma} - \gamma\right) \xrightarrow{d} n(0,1) \tag{A-3}$$

(A-1) is indeed a 2SOE has been established by Terza (2016 a and b). It is also shown therein that because (A-1) is a 2SOE, under general conditions its asymptotic variance is

$$\text{a var}(\hat{\gamma}) = \mathrm{E}\left[\nabla_{\tau} g(\tau, X)\right] \mathrm{AVAR}(\hat{\tau}) \mathrm{E}\left[\nabla_{\tau} g(\tau, X)\right]' + \mathrm{E}\left[(g(\tau, X) - \gamma)^2\right] \tag{A-4}$$

which can be consistently estimated as

$$\widehat{\text{a var}}(\hat{\gamma}) = A + B \tag{A-5}$$

where

$$A = C\, \widehat{\mathrm{AVAR}}(\hat{\tau}) C'$$

$$B = \frac{\sum_{i=1}^{n} \left(g(\hat{\tau}, X_i) - \hat{\gamma}\right)^2}{n}$$

$$C = \left(\frac{\sum_{i=1}^{n} \nabla_{\tau} g(\hat{\tau}, X_i)}{n} \right)$$

where

$$A = \left(\frac{\sum_{i=1}^{n} \nabla_{\tau} g(\hat{\tau}, X_i)}{n} \right) \widehat{AVAR}(\hat{\tau}) \left(\frac{\sum_{i=1}^{n} \nabla_{\tau} g(\hat{\tau}, X_i)}{n} \right)'$$

$$B = \frac{\sum_{i=1}^{n} \left(g(\hat{\tau}, X_i) - \hat{\gamma} \right)^2}{n}$$

$\nabla_{\tau} g(\tau, X) \equiv$ the gradient of $g(\beta, X)$ (a row vector)

$\nabla_{\tau} g(\hat{\tau}, X_i) \equiv \nabla_{\tau} g(\tau, X)$ evaluated at X_i and $\hat{\tau}$

and

$AVAR(\hat{\tau}) \equiv$ the asymptotic covariance matrix of $\hat{\tau}$

with $\widehat{AVAR}(\hat{\tau})$ being a consistent estimator of $AVAR(\hat{\tau})$

Appendix B

By equations (3) and (5) we have

$$E[Y_{X_p^* j} \mid C] = \Pr(Y_{X_p^* j} = 1 \mid C) = \pi_j(X_p^*, C; \tau). \qquad (j = 1, ..., J) \qquad \text{(B-1)}$$

where

$$C = [X_o \quad X_u] \qquad \text{(B-2)}$$

Given the definitions of X_o and X_u, it follows that

$$E[Y_{X_p^* j} \mid X_p, C] = E[Y_{X_p^* j} \mid C]. \qquad \text{(B-3)}$$

Therefore, by Theorem 11 of Terza (2018), under some general conditions, (B-3) implies that

$$E[Y_j \mid X_p, X_o, X_u] = \Pr(Y_j = 1 \mid X_p, X_o, X_u) = \pi_j(X_p, X_o, X_u; \tau) \quad (j = 1, ..., J) \qquad \text{(B-4)}$$

which, in turn, implies that

$$\begin{aligned} &\text{pmf}(Y \mid X_p, X_o, X_u) \\ &\quad = f_{(Y_{X_p^*} \mid X_o, X_u)}(Y, X_p, X_o, X_u; \tau) \\ &\quad = \prod_{j=1}^{J} \pi_j(X_p, X_o, X_u; \tau)^{Y_j}. \end{aligned} \qquad \text{(B-5)}$$

Appendix C

Generalizing to the case in which $j = 1, ..., J$, the J relevant pmf values can then be written

$$\pi_1(X_p^*, X_o, X_u; \tau) = \Pr(Y_{X_p^* 1} = 1 \mid X_o, X_u)$$

$$= \Pr(\xi_2 \le -V_2^*, \xi_3 \le -V_3^* ... \xi_J \le -V_J^*)$$

$$= G_1(-V_2^*, -V_3^* ..., -V_J^*) \qquad \text{(C-1)}$$

and for $j = 2, ..., J$

$$\pi_j(X_p^*, X_o, X_u; \tau) = \Pr(Y_{X_p^* j} = 1 \mid X_o, X_u)$$

$$= \Pr(\xi_2 - \xi_j \le V_j^* - V_2^*, ..., -\xi_j \le V_j^* - V_j^*, ..., \xi_J - \xi_j \le V_j^* - V_J^*)$$

$$= G_j(V_j^* - V_2^*, ..., V_j^* - V_j^*, ..., V_j^* - V_J^*) \qquad \text{(C-2)}$$

where

$G_1(\bullet, \ldots, \bullet)$ denotes the multivariate cdf of ξ_2 , ξ_3 . . . ξ_J

and

$G_j(\bullet, \ldots, \bullet)$ denotes the multivariate cdf of $\xi_2 - \xi_j$, ... $-\xi_j$ ... $\xi_J - \xi_j$.

The exact formulations of $G_1(\bullet, \ldots, \bullet)$ and $G_j(\bullet, \ldots, \bullet)$ follow from the assumed distribution of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \ ... \ \varepsilon_J]$. The general form of the model detailed in (C-1) and (C-2) has the following parameters

J coefficient parameter vectors for the utility indexes, viz.,

$$[\beta^*_{p1} \quad \beta^{*\prime}_{o1} \quad \beta^*_{u1}], \ldots, [\beta^*_{pJ} \quad \beta^{*\prime}_{oJ} \quad \beta^*_{uJ}].$$

σ^*_j -- J variance parameters (j = 1, …, J)

and

$\sigma^*_{j\ell}$ -- J(J-1)/2 covariance parameters (j, ℓ = 1, …, J).

First note that an admissible reduction is evident in (C-1) and (C-2) because it allows the relevant probabilities to be expressed in terms of the $J-1$ normalized coefficient parameter vectors of the form[27]

$$[(\beta^o_{pj} - \beta^o_{p1}) \quad (\beta^{o\prime}_{oj} - \beta^{o\prime}_{o1}) \quad (\beta^o_{uj} - \beta^o_{u1})].$$

Moreover, as we will see later when we discuss the "standardized" version of the model, it is typically possible to further admissibly reduce the model by normalizing on one of the $J-1$ variance parameters of $[\xi_2 \ldots \xi_J]$.

I now turn to the derivation of the standardized versions of (C-1) and (C-2). I do this in order to gain insight regarding parametric identification and as a means of characterizing the covariance flexibility of the various FPMO models arising from alternative specifications for the distribution of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \ldots \varepsilon_J]$. I begin with the trinomial model in (12), (13) and (14). As can be seen in (12), there are only two unique difference random variables, ξ_2 and ξ_3, so there can be at most three identified variance-covariance parameters, namely $\omega^*_2 = \text{var}(\xi_2)$, $\omega^*_3 = \text{var}(\xi_3)$ and $\omega^*_{23} = \text{cov}(\xi_2, \xi_3)$ We can

[27] For the definition of an *admissible reduction* see Terza (1985).

write the standardized versions of (12), (13) and (14), respectively, as

$$\pi_1(X_p^*, X_o, X_u; \tau)$$

$$= \Pr\left(\frac{\xi_2}{\sqrt{\omega_2^*}} \le \frac{-V_2^*}{\sqrt{\omega_2^*}} \text{ and } \frac{\xi_3}{\sqrt{\omega_3^*}} \le \frac{-V_3^*}{\sqrt{\omega_3^*}} \right). \tag{C-3}$$

and the correlation between ξ_2 and ξ_3 is

$$d_1 = \text{corr}(\xi_2, \xi_2) = \frac{\omega_{23}^*}{\sqrt{\omega_2^* \omega_3^*}}. \tag{C-4}$$

Similarly, for the remaining two trinomial categories j = 2 and 3, respectively, we have

$$\pi_2(X_p^*, X_o, X_u; \tau)$$

$$= \Pr\left(\frac{-\xi_2}{\sqrt{\omega_2^*}} \le \frac{V_2^*}{\sqrt{\omega_2^*}} \text{ and } \frac{\xi_3 - \xi_2}{\sqrt{\omega_2^* + \omega_3^* - 2\omega_{23}^*}} \le \frac{V_2^* - V_3^*}{\sqrt{\omega_2^* + \omega_3^* - 2\omega_{23}^*}} \right) \tag{C-5}$$

with relevant correlation

$$d_2 = \text{corr}(-\xi_2, \xi_3 - \xi_2) = \frac{\omega_2^* - \omega_{23}^*}{\sqrt{\omega_2^{*2} + \omega_2^* \omega_3^* - 2\omega_2^* \omega_{23}^*}} \tag{C-6}$$

and

$$\pi_3(X_p^*, X_o, X_u; \tau)$$

$$= \Pr\left(\frac{\xi_2 - \xi_3}{\sqrt{\omega_2^* + \omega_3^* - 2\omega_{23}^*}} \le \frac{V_3^* - V_2^*}{\sqrt{\omega_2^* + \omega_3^* - 2\omega_{23}^*}} \text{ and} \frac{-\xi_3}{\sqrt{\omega_3^*}} \le \frac{V_3^*}{\sqrt{\omega_3^*}} \right)$$

with relevant correlation

$$d_3 = \text{corr}(\xi_2 - \xi_3, -\xi_3) = \frac{\omega_3^* - \omega_{23}^*}{\sqrt{\omega_3^{*2} + \omega_2^*\omega_3^* - 2\omega_2^*\omega_{23}^*}}. \tag{C-7}$$

It is clear from (C-3) through (C-7) that the model can be further admissibly reduced (and, therefore must be admissibly reduced if the parameter vector is to be identified) by

1) dividing V_2^* and V_3^* by $\sqrt{\omega_2^*}$

and

2) dividing ω_2^*, ω_3^* and ω_{23}^* by ω_2^*.

Under these additional normalizations we get

$$\pi_1(X_p^*, X_o, X_u; \tau) = G_1^s\left(-V_2, \frac{-V_3}{\sqrt{\omega_3}}\right) \tag{C-8}$$

$$d_1 = \text{corr}(\xi_2, \xi_2) = \frac{\omega_{23}}{\sqrt{\omega_3}} \tag{C-9}$$

$$\pi_2(X_p^*, X_o, X_u; \tau) = G_2^s\left(V_2, \frac{V_2 - V_3}{\sqrt{1 + \omega_3 - 2\omega_{23}}}\right) \tag{C-10}$$

$$d_2 = \text{corr}(-\xi_2, \xi_3 - \xi_2) = \frac{1 - \omega_{23}}{\sqrt{1 + \omega_3 - 2\omega_{23}}} \tag{C-11}$$

$$\pi_3(X_p^*, X_o, X_u; \tau) = G_3^*\left(\frac{V_3 - V_2}{\sqrt{1 + \omega_3 - 2\omega_{23}}}, \frac{V_3}{\sqrt{\omega_3}}\right) \tag{C-12}$$

$$d_3 = \text{corr}(\xi_2 - \xi_3, -\xi_3) = \frac{\omega_3 - \omega_{23}}{\sqrt{\omega_3(1 + \omega_3 - 2\omega_{23})}} \tag{C-13}$$

where for j =1

$G_1^s(\bullet, \bullet) =$ the bivariate cdf of ξ_2 and $\frac{\xi_3}{\sqrt{\omega_3}}$ (the cdf of the standardized version of $[\xi_2 \quad \xi_3]$)

and for j = 2 or 3

$G_j^*(\bullet, \bullet) =$ the bivariate cdf of $-\xi_j$ and $\frac{\xi_r - \xi_j}{\sqrt{\text{var}(\xi_r - \xi_j)/\omega_2^*}}$ such that $r \neq j$ (the cdfs of the standardized versions of $[-\xi_2 \quad (\xi_3 - \xi_2)]$ and $[(\xi_2 - \xi_3) \quad -\xi_3]$)

with

$$V_2 = \frac{V_2^*}{\sqrt{\omega_2^*}} = X_p^* \beta_{p2} + X_o \beta_{o2} + X_u \beta_{u2}$$

$$V_3 = \frac{V_3^*}{\sqrt{\omega_2^*}} = X_p^* \beta_{p3} + X_o \beta_{o3} + X_u \beta_{u3}$$

$$\omega_3 = \omega_3^* / \omega_2^*$$

$$\omega_{23} = \omega_{23}^* / \omega_2^*$$

$$\beta_{pj} = \frac{1}{\sqrt{\omega_2^*}} \beta_{pj}^*$$

$$\beta_{oj} = \frac{1}{\sqrt{\omega_2^*}} \beta_{oj}^*$$

$$\beta_{uj} = \frac{1}{\sqrt{\omega_2^*}} \beta_{uj}^*$$

$$\tau = [[\beta_{p2} \quad \beta'_{o2} \quad \beta_{u2}] \quad [\beta_{p3} \quad \beta'_{o3} \quad \beta_{u3}] \quad \omega_3 \quad \omega_{23}].$$

The exact formulations of $G_1^s(\bullet, \bullet)$, $G_2^s(\bullet, \bullet)$ and $G_3^s(\bullet, \bullet)$ follow from the assumed distribution of $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$. In summary, in the J = 3 case, the number of parameters is admissibly reduced from 9 (3 random utility coefficient vectors and 6 variance-covariance parameters) to 4 (2 identified random utility coefficient vectors and 2 identified variance-covariance parameters).

In the general case there are $J - 1$ unique utility differences $\xi_2, \xi_3, \ldots, \xi_J$. This means that there are at most $J - 1$ identified variance parameters

$$\text{var}(\xi_j) = \omega_j^* \tag{C-14}$$

and $\frac{(J-1)(J-2)}{2}$ covariance parameters

$$\text{cov}(\xi_j, \xi_\ell) = \omega_{j\ell}^* \tag{C-15}$$

$j, \ell = 2, \ldots, J; j \neq \ell$. I can write a standardized version of this multinomial model analogous to that of the trinomial model in (C-3) through (C-7). This multinomial model can be admissibly reduced by

1) dividing V_j^* by $\sqrt{\omega_2^*}$

and

2) dividing ω_j^* and $\omega_{j\ell}^*$ by ω_2^*

to obtain the following $(J - 1)$ – variate analogues to (C-8) through (C-12)

$$P_1^* = \pi_1(X_p^*, X_o, X_u; \tau) = G_1^s\left(-V_2, \frac{-V_3}{\sqrt{\omega_3}}, ..., \frac{-V_J}{\sqrt{\omega_J}}\right) \quad \text{(C-16)}$$

where

$G_1^s(\bullet, \bullet, \ldots, \bullet) =$ the multivariate cdf of ξ_2 , $\frac{\xi_3}{\sqrt{\omega_3}}$ ... $\frac{\xi_J}{\sqrt{\omega_J}}$

$$V_j = \frac{V_j^*}{\sqrt{\omega_2^*}} = X_p^*\beta_{pj} + X_o\beta_{oj} + X_u\beta_{uj}$$

$$\beta_{pj} = \frac{1}{\sqrt{\omega_2^*}}\beta_{pj}^*$$

$$\beta_{oj} = \frac{1}{\sqrt{\omega_2^*}}\beta_{oj}^*$$

$$\beta_{uj} = \frac{1}{\sqrt{\omega_2^*}}\beta_{uj}^*$$

and

$$\omega_j = \omega_j^* / \omega_2^* .$$

The associated correlation matrix for $[\xi_2 \quad \xi_3 \ldots \xi_J]$ is

$$D_1[\xi_2 \quad \xi_3 \ldots \xi_J]$$

$$= \begin{bmatrix} 1 & \frac{\omega_{23}}{\sqrt{\omega_3}} & \cdots & & & \frac{\omega_{2J}}{\sqrt{\omega_J}} \\ & 1 & & & & \frac{\omega_{3J}}{\sqrt{\omega_3\omega_J}} \\ & & 1 & \cdots & & \vdots \\ & & & \ddots & & \vdots \\ & & & & 1 & \frac{\omega_{J-2,J-1}}{\sqrt{\omega_{J-2}\omega_{J-1}}} \quad \frac{\omega_{J-2,J}}{\sqrt{\omega_{J-2}\omega_J}} \\ & & & & & 1 \quad \frac{\omega_{J-1,J}}{\sqrt{\omega_{J-1}\omega_J}} \\ & & & & & 1 \end{bmatrix} \tag{C-17}$$

where $\omega_{\ell j} = \omega^*_{\ell j} / \omega^*_2$ for $\ell \neq j$. Similarly, for the remaining three multinomial categories j = 2, . . ., J, respectively

$$P^*_j = \pi_j(X^*_p, X_o, X_u; \tau) = G^s_j\left(\frac{(V_j - V_2)}{\sqrt{1+\omega_j - 2\omega_{2j}}}, ..., \frac{V_j}{\sqrt{\omega_j}}, ..., \frac{(V_j - V_J)}{\sqrt{\omega_j + \omega_J - 2\omega_{jJ}}}\right) \tag{C-18}$$

where

$G^s_j(\bullet, \bullet, \ldots, \bullet) =$ the multivariate cdf of $\frac{\xi_2 - \xi_j}{\sqrt{1+\omega_j - 2\omega_{2j}}}$, ... $\frac{-\xi_j}{\sqrt{\omega_j}}$ ... $\frac{\xi_J - \xi_j}{\sqrt{\omega_j + \omega_J - 2\omega_{J4}}}$.

The associated correlation matrix

$$D_j[(\xi_2 - \xi_j) \ldots - \xi_j \ldots (\xi_J - \xi_j)] \tag{C-19}$$

like $D_1[\xi_2 \quad \xi_3 \ldots \xi_J]$, has 1's along the main diagonal. To define the remaining elements of D_j I need only specify its upper triangle (excluding the diagonal, of course). The elements of the upper triangle corresponding to correlations involving $\xi_j = \varepsilon_j - \varepsilon_1$ have the following general form

$$\frac{(\omega_j - \omega_{j\ell})}{\sqrt{\omega_j(\omega_j + \omega_\ell - 2\omega_{j\ell})}}.$$

The remaining elements of the upper triangle of D_j are of the form

$$\frac{(\omega_j - \omega_{jm} - \omega_{\ell j} - \omega_{\ell m})}{\sqrt{(\omega_j + \omega_\ell - 2\omega_{j\ell})(\omega_j + \omega_m - 2\omega_{jm})}}$$

where j, ℓ, m = 2, ... , J; ℓ, m are not equal to j; m > ℓ. Note that in all of the above expressions ω_2 is normalized to be equal to 1. The full parameter vector is

$$\tau = [[\beta_{p2} \quad \beta'_{o2} \quad \beta_{u2}{}'] \ldots [\beta_{pJ} \quad \beta'_{oJ} \quad \beta_{uJ}{}'] \quad \omega_3 \ldots \omega_j \quad \omega_{23} \quad \omega_{24} \ldots \omega_{(J-1)J}].$$

In summary, in the general case, the number of parameters is admissibly reduced from $\frac{J(J+3)}{2}$ (J random utility coefficient vectors, J variance parameters and $\frac{J(J-1)}{2}$ covariance parameters) to $\frac{J^2 + J - 4}{2}$ (J - 1 identified random utility coefficient vectors, $J - 2$ identified variance parameters [recall the normalization on ω_2^*] and $\frac{(J-1)(J-2)}{2}$ identified covariance parameters).

Appendix D

When J = 4, there are only three unique utility differences, ξ_2, ξ_3, and ξ_4 so although there are 10 variance-covariance parameters in the primitive model, there can be at most 6 identified variance-covariance parameters, viz.,

$$\text{var}(\xi_2) = \omega_2^*,\ \text{var}(\xi_3) = \omega_3^*,\ \text{var}(\xi_4) = \omega_4^*$$

$$\text{cov}(\xi_2, \xi_3) = \omega_{23}^*,\ \text{cov}(\xi_2, \xi_4) = \omega_{24}^*, \text{ and } \text{cov}(\xi_3, \xi_4) = \omega_{34}^*.$$

We can write a standardized version of this quadrinomial model that is analogous to that of the trinomial model in (C-3) through (C-7). This quadrinomial model can be admissibly reduced by

1) dividing $V_2^*,\ V_3^*$, and V_4^* by $\sqrt{\omega_2^*}$

and

2) dividing $\omega_2^*,\ \omega_3^*,\ \omega_4^*,\ \omega_{23}^*,\ \omega_{24}^*$ and ω_{34}^* by ω_2^*

to obtain the following trivariate analogues to (C-8), (C-9) and (C-13)

$$P_1^* = \pi_1(X_p^*, X_o, X_u; \tau) = G_1^s\left(-V_2, \frac{-V_3}{\sqrt{\omega_3}}, \frac{-V_4}{\sqrt{\omega_4}}\right) \qquad \text{(D-1)}$$

$$D_1[\xi_2 \quad \xi_3 \quad \xi_4] = \begin{bmatrix} 1 & \frac{\omega_{23}}{\sqrt{\omega_3}} & \frac{\omega_{24}}{\sqrt{\omega_4}} \\ & 1 & \frac{\omega_{34}}{\sqrt{\omega_3\omega_4}} \\ & & 1 \end{bmatrix} \equiv \text{correlation matrix for } [\xi_2 \quad \xi_3 \quad \xi_4]$$

(D-2)

$$P_2^* = \pi_2(X_p^*, X_o, X_u; \tau) = G_2^s\left(V_2, \frac{V_2 - V_3}{\sqrt{1+\omega_3-2\omega_{23}}}, \frac{V_2 - V_4}{\sqrt{1+\omega_4-2\omega_{24}}}\right)$$

(D-3)

$$D_2[-\xi_2 \quad (\xi_3-\xi_2) \quad (\xi_4-\xi_2)]$$

$$= \begin{bmatrix} 1 & \frac{1-\omega_{23}}{\sqrt{1+\omega_2+2\omega_{23}}} & \frac{1-\omega_{24}}{\sqrt{1+\omega_4-2\omega_{24}}} \\ & 1 & \frac{1+\omega_{34}-\omega_{23}-\omega_{24}}{\sqrt{(1+\omega_3-2\omega_{23})(1+\omega_4-2\tau\omega_{24})}} \\ & & 1 \end{bmatrix}$$

$$\equiv \text{correlation matrix for } [-\xi_2 \quad (\xi_3-\xi_2) \quad (\xi_4-\xi_2)] \qquad \text{(D-4)}$$

$$P_3^* = \pi_3(X_p^*, X_o, X_u; \tau) = G_3^s\left(\frac{V_3 - V_2}{\sqrt{1+\omega_3-2\omega_{23}}}, \frac{V_3}{\sqrt{\omega_3}}, \frac{V_3 - V_4}{\sqrt{\omega_3+\omega_4-2\omega_{34}}}; D_3\right)$$

(D-5)

$$D_3[(\xi_2-\xi_3) \quad -\xi_3 \quad (\xi_4-\xi_3)]$$

$$= \begin{bmatrix} 1 & \dfrac{\omega_3 - \omega_{23}}{\sqrt{\omega_3 + \omega_3^2 - 2\omega_3\tau\omega_{23}}} & \dfrac{\omega_3 + \omega_{24} - \omega_{23} - \omega_{34}}{\sqrt{(1+\omega_3 - 2\omega_{23})(\omega_3 + \omega_4 - 2\tau\omega_{34})}} \\ & 1 & \dfrac{\omega_3 - \omega_{34}}{\sqrt{\omega_3^2 + \omega_3\omega_4 - 2\omega_3\omega_{34}}} \\ & & 1 \end{bmatrix}$$

$\equiv$ correlation matrix for $[(\xi_2 - \xi_3) \quad -\xi_3 \quad (\xi_4 - \xi_3)]$ (D-6)

$$P_4^* = \pi_4(X_p^*, X_o, X_u; \tau) = G_1^s\left(\frac{V_4 - V_2}{\sqrt{1+\omega_4 - 2\omega_{24}}}, \frac{V_4 - V_3}{\sqrt{\omega_3 + \omega_4 - 2\omega_{34}}}, \frac{V_4}{\sqrt{\omega_4}}\right)$$

(D-7)

$$D_4[(\xi_2 - \xi_4) \quad (\xi_3 - \xi_4) \quad -\xi_4] \tag{D-8}$$

$$= \begin{bmatrix} 1 & \dfrac{\omega_4 + \omega_{23} - \omega_{24} - \omega_{34}}{\sqrt{(1+\omega_4 - 2\omega_{24})(\omega_3 + \omega_4 - 2\omega_{34})}} & \dfrac{\omega_4 - \omega_{24}}{\sqrt{\omega_4 + \omega_{42} - 2\omega_4\omega_{24}}} \\ & 1 & \dfrac{\omega_4 - \omega_{34}}{\sqrt{\omega_4^2 + \omega_4\omega_{43} - 2\omega_4\omega_{34}}} \\ & & 1 \end{bmatrix}$$

$\equiv$ correlation matrix for $[(\xi_2 - \xi_4) \quad (\xi_3 - \xi_4) \quad -\xi_4]$ (D-9)

where for j =1

$G_1^s(\bullet, \bullet, \bullet)$ = the trivariate cdf of ξ_2 , $\dfrac{\xi_3}{\sqrt{\omega_3}}$ and $\dfrac{\xi_4}{\sqrt{\omega_4}}$ (the cdf of the standardized version of $[\xi_2 \quad \xi_3 \quad \xi_4]$)

and for j = 2, 3 or 4

$G_j^s(\bullet, \bullet, \bullet)$ = the trivariate cdf of $\frac{\xi_2 - \xi_j}{\sqrt{1+\omega_j - 2\omega_{2j}}}$, $\frac{-\xi_j}{\sqrt{\omega_j}}$ and $\frac{\xi_4 - \xi_j}{\sqrt{\omega_j + \omega_4 - 2\omega_{j4}}}$ (the cdf of the standardized version of $[\xi_2 - \xi_j \quad -\xi_j \quad \xi_4 - \xi_j]$)

with

$$V_2 = \frac{V_2^*}{\sqrt{\omega_2^*}} = X_p^* \beta_{p2} + X_o \beta_{o2} + X_u \beta_{u2}$$

$$V_3 = \frac{V_3^*}{\sqrt{\omega_2^*}} = X_p^* \beta_{p3} + X_o \beta_{o3} + X_u \beta_{u3}$$

$$V_4 = \frac{V_4^*}{\sqrt{\omega_2^*}} = X_p^* \beta_{p4} + X_o \beta_{o4} + X_u \beta_{u4}$$

$$\omega_3 = \omega_3^* / \omega_2^*$$

$$\omega_4 = \omega_4^* / \omega_2^*$$

$$\omega_{23} = \omega_{23}^* / \omega_2^*$$

$$\omega_{24} = \omega_{24}^* / \omega_2^*$$

$$\omega_{34} = \omega_{34}^* / \omega_2^*$$

$$\beta_{pj} = \frac{1}{\sqrt{\omega_2^*}} \beta_{pj}^*$$

$$\beta_{oj} = \frac{1}{\sqrt{\omega_2^*}} \beta_{oj}^*$$

$$\beta_{uj} = \frac{1}{\sqrt{\omega_2^*}} \beta_{uj}^*$$

and

$$\tau = [[\beta_{p2} \quad \beta'_{o2} \quad \beta_{u2}] \quad [\beta_{p3} \quad \beta'_{o3} \quad \beta_{u3}] \quad [\beta_{p4} \quad \beta'_{o4} \quad \beta_{u4}] \quad \omega_3 \quad \omega_4 \quad \omega_{23} \quad \omega_{24} \quad \omega_{34}].$$

In summary, in the J = 4 case, the number of parameters is admissibly reduced from 14 (4 random utility coefficient vectors and 10 variance-covariance parameters) to 8 (3 identified random utility coefficient vectors and 5 identified variance-covariance parameters).

Note that the MNP model will be a special case for the J = 4 case if the error terms are assumed to be multivariate normal distributed,

Appendix E

To get the marginal pdfs of ε_2 and ε_3 and the joint marginal pdf of ε_2 and ε_3 we need only focus on the joint marginal cdf of ε_2 and ε_3 as defined in equation (51) in the main text

$$F_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3) = \exp\left(-\left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{1-\eta} \right). \tag{E-1}$$

To be specific, these marginal pdfs can be derived as

$$f_{\varepsilon_2}(\varepsilon_2) = \left. \frac{\partial F_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3)}{\partial \varepsilon_2} \right|_{\varepsilon_3=\infty}$$

$$f_{\varepsilon_3}(\varepsilon_3) = \left. \frac{\partial F_{\varepsilon_2\varepsilon_3}(\varepsilon_1, \varepsilon_3)}{\partial \varepsilon_3} \right|_{\varepsilon_2=\infty}$$

and

$$f_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3) = \frac{\partial^2 F_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3)}{\partial \varepsilon_2 \partial \varepsilon_3}.$$

We have

$$\frac{\partial F_{\varepsilon_2\varepsilon_3}(\varepsilon_1, \varepsilon_3)}{\partial \varepsilon_2} = F_{\varepsilon_2\varepsilon_3}(\varepsilon_1, \varepsilon_3) \times C(\varepsilon_1, \varepsilon_3)$$

where

$$C(\varepsilon_2, \varepsilon_3) = \left(e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right)^{-\eta} \times \left(e^{-\frac{\varepsilon_2}{1-\eta}} \right).$$

Now

$$F_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3)\Big|_{\varepsilon_3=\infty} = \exp\left(-\left(e^{-\frac{\varepsilon_2}{1-\eta}} \right)^{1-\eta} \right)$$

$$= \exp\left(-e^{-\varepsilon_2} \right)$$

and

$$C(\varepsilon_2, \varepsilon_3)\Big|_{\varepsilon_3=\infty} = e^{\left(-\frac{\varepsilon_2}{1-\eta}\right)(-\eta)} \left(e^{\left(-\frac{\varepsilon_2}{1-\eta}\right)} \right)$$

$$= e^{\left(\frac{\varepsilon_2}{1-\eta}\right)(\eta)} \left(e^{\left(-\frac{\varepsilon_2}{1-\eta}\right)} \right)$$

$$= e^{\left(\frac{\varepsilon_2\eta}{1-\eta} - \frac{\varepsilon_2}{1-\eta}\right)}$$

$$= e^{\left(\frac{-\varepsilon_2(1-\eta)}{1-\eta}\right)}$$

$$= e^{-\varepsilon_2}.$$

Combining these results yields

$$f_{\varepsilon_2}(\varepsilon_2) = \exp(-e^{-\varepsilon_2})\exp(-\varepsilon_2)$$

$$= \exp(-\varepsilon_2 - \exp(-\varepsilon_2))$$

so the marginal distribution of ε_2 is log-Weibull.

Appendix F

From Appendix E and the discussion surrounding equations (58) and (59) in the text, we have that if $\varepsilon = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]$ is GEV distributed as in (58), then its marginal distributions are log-Weibull. From this we get

$$E[\varepsilon_1] = E[\varepsilon_2] = E[\varepsilon_3] = \gamma$$

and

$$\text{var}(\varepsilon_1) = \text{var}(\varepsilon_2) = \text{var}(\varepsilon_3) = \frac{\Pi^2}{6}.$$

Also, as we have defined in the text

$$\text{cov}(\varepsilon_1, \varepsilon_2) = \text{cov}(\varepsilon_1, \varepsilon_3) = 0$$

and

$$\text{cov}(\varepsilon_2, \varepsilon_3) = \sigma_{23}(\eta).$$

We have then that

$$E[\varepsilon_2\varepsilon_1] = E[\varepsilon_2]E[\varepsilon_1] = \gamma^2$$

$$E[\varepsilon_3\varepsilon_1] = E[\varepsilon_3]E[\varepsilon_1] = \gamma^2$$

$$E[\varepsilon_2\varepsilon_3] = \sigma_{23}(\eta) + \gamma^2$$

and

$$E[\varepsilon_1^2] = E[\varepsilon_2^2] = E[\varepsilon_3^2] = \frac{\Pi^2}{6} + \gamma^2.$$

For the two relevant "difference" random variables, $\xi_2 = \varepsilon_2 - \varepsilon_1$ and $\xi_3 = \varepsilon_3 - \varepsilon_1$ we have

$$E[\xi_2] = E[\xi_3] = 0$$

and for $\omega_2^* = \text{var}(\xi_2) = E[\xi_2^2]$ and $\omega_3^* = \text{var}(\xi_3) = E[\xi_3^2]$ and $\omega_{23}^* = \text{cov}(\xi_2, \xi_3) = E[\xi_2 \xi_3]$ we get

$$\omega_2^* = E[\xi_2^2] = E[\varepsilon_2^2] - 2E[\varepsilon_2 \varepsilon_1] + E[\varepsilon_1^2]$$

$$= \frac{\Pi^2}{6} + \gamma^2 - 2\gamma^2 + \frac{\Pi^2}{6} + \gamma^2$$

$$= \frac{\Pi^2}{3}$$

and, likewise

$$\omega_3^* = \frac{\Pi^2}{3}$$

moreover

$$\omega_{23}^* = E[\xi_2 \xi_3] = E[\varepsilon_2 \varepsilon_3] - E[\varepsilon_2 \varepsilon_1] - E[\varepsilon_3 \varepsilon_1] + E[\varepsilon_1^2]$$

$$= \sigma_{23}(\eta) + \gamma^2 - \gamma^2 - \gamma^2 + \frac{\Pi^2}{6} + \gamma^2$$

$$= \sigma_{23}(\eta) + \frac{\Pi^2}{6}. \qquad \text{(F-1)}$$

Applying the admissible and requisite reduction of the model we get

$$\omega_3 = \omega_3^* / \omega_2^* = 1$$

and

$$\omega_{23} = \omega_{23}^* / \omega_2^* = \frac{\sigma_{23}(\eta) + \frac{\Pi^2}{6}}{\frac{\Pi^2}{3}} = \frac{3\sigma_{23}(\eta)}{\Pi^2} + \frac{1}{2}.$$

So we can write the relevant version of (23) in this case as

$$D_1[\xi_2 \quad \xi_3] = \begin{pmatrix} 1 & \frac{\omega_{23}^*}{\sqrt{\omega_2^* \omega_3^*}} \\ \frac{\omega_{23}^*}{\sqrt{\omega_2^* \omega_3^*}} & 1 \end{pmatrix} = \begin{pmatrix} 1 & \omega_{23} \\ \omega_{23} & 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{3\sigma_{23}(\eta)}{\Pi^2} + \frac{1}{2} \\ \frac{3\sigma_{23}(\eta)}{\Pi^2} + \frac{1}{2} & 1 \end{pmatrix} \tag{F-2}$$

Note also that from the above we get

$$\text{var}(\xi_3 - \xi_2) = \text{var}(\xi_2 - \xi_3) = \omega_2^* + \omega_3^* - 2\omega_{23}^*$$

Applying the admissible and requisite reduction of the model we get the identified versions of these variances as

$$2(1-\omega_{23}) = 2\left(1-\left[\frac{3\sigma_{23}(\eta)}{\Pi^2}+\frac{1}{2}\right]\right) = 1-\frac{6\sigma_{23}(\eta)}{\Pi^2}$$

So we can write the relevant versions of (26) and (27) in this case as

$$D_2[-\xi_2 \quad (\xi_3-\xi_2)] = \begin{pmatrix} 1 & \frac{\omega_2^*-\omega_{23}^*}{\sqrt{\omega_2^*+\omega_3^*-2\omega_{23}^*}} \\ \frac{\omega_2^*-\omega_{23}^*}{\sqrt{\omega_2^*+\omega_3^*-2\omega_{23}^*}} & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{\omega_2^*-\omega_{23}^*}{\sqrt{\omega_2^*+\omega_3^*-2\omega_{23}^*}} \\ \frac{\omega_2^*-\omega_{23}^*}{\sqrt{\omega_2^*+\omega_3^*-2\omega_{23}^*}} & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{1-\omega_{23}}{\sqrt{2(1-\omega_{23})}} \\ \frac{1-\omega_{23}}{\sqrt{2(1-\omega_{23})}} & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{\frac{1}{2}\left(1-\frac{6\sigma_{23}(\eta)}{\Pi^2}\right)}{\sqrt{1-\frac{6\sigma_{23}(\eta)}{\Pi^2}}} \\ \frac{\frac{1}{2}\left(1-\frac{6\sigma_{23}(\eta)}{\Pi^2}\right)}{\sqrt{1-\frac{6\sigma_{23}(\eta)}{\Pi^2}}} & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{1}{2}\left(1 - \frac{6\sigma_{23}(\eta)}{\Pi^2}\right)^{\frac{1}{2}} \\ \frac{1}{2}\left(1 - \frac{6\sigma_{23}(\eta)}{\Pi^2}\right)^{\frac{1}{2}} & 1 \end{pmatrix}$$

Appendix G

For the joint marginal pdf of ε_2 and ε_3 we have

$$f_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3) = \frac{\partial^2 F_{\varepsilon_2\varepsilon_3}(\varepsilon_2, \varepsilon_3)}{\partial\varepsilon_2\partial\varepsilon_3}$$

$$= \exp\left(-\left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{1-\eta} - \frac{\varepsilon_2 + \varepsilon_3}{1-\eta} \right) \left(\left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{-2\eta} + \frac{\eta}{1-\eta} \left\{ e^{-\frac{\varepsilon_2}{1-\eta}} + e^{-\frac{\varepsilon_3}{1-\eta}} \right\}^{-\eta-1} \right)$$

.

References

Ames, G. M., Grube, J. W., & Moore, R. S. (1997). The relationship of drinking and hangovers to workplace problems: an empirical study. *Journal of Studies on Alcohol*, *58*(1), 37–47.

Asgeirsdottir, T. L., & McGeary, K. A. (2009). Alcohol and labor supply: the case of Iceland. *European Journal of Health Economics*, *10*(4), 455–465.

Balsa, A. I., & French, M. T. (2010). Alcohol use and the labor market in Uruguay. *Health Economics*, *19*(7), 833–854.

Blum, T. C., Roman, P. M., & Martin, J. K. (1993). Alcohol consumption and work performance. *Journal of Studies on Alcohol*, *54*(1), 61–70.

Booth, B. M., & Feng, W. (2002). The impact of drinking and drinking consequences on short-term employment outcomes in at-risk drinkers in six southern states. *The Journal of Behavioral Health Services & Research*, *29*(2), 157–166.

Feng, W., Zhou, W, Butler, J.S., Booth, B.M., & French, M. (2001). The impact of problem drinking on employment. *Health Economics*, *10*(6), 509–521.

Florence, C. S., Zhou, C., Luo, F., & Xu, L. (2016). The economic burden of prescription opioid overdose, abuse, and dependence in the United States, 2013. *Medical Care*, 54(10), 901–906.

Heckman, J. J., Tobias, J. L., and Vytlacil, E. J. (2003). Simple estimators for treatment parameters in a latent-variable framework. *The Review of Economics and Statistics*, 85(3), 748–755.

Heckman, J. J. and Vytlacil, E. J. (2005). Strucutral equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3), 669–738.

Hoffmann, J., & Larison, C. (1999). Drug use, workplace accidents and employee turnover. *Journal of Drug Issues*, *29*(2), 341–364.

Landry, M. (1997). Overview of addiction treatment effectiveness. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Johansson, E., Alho, H., Kiiskinen, U., & Poikolainen, K. (2007). The association of alcohol dependency with employment probability: evidence from the population survey “health 2000 in Finland”. *Health Economics*, *16*(7), 739–754.

Krueger, A. (2017) Where have all the workers gone? An inquiry into the decline of the U.S.. labor force participation rate *Brookings Papers on Economic Activity Conference*

MacDonald, Z., & Shields, M. A. (2004). Does problem drinking affect employment? evidence from england. *Health Economics*, *13*(2), 139–155.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge university press.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometric*, Cambridge university press, New York.

Mullahy, J., & Sindelar, J. (1996). Employment, unemployment, and problem drinking. *Journal of Health Economics*, *15*(4), 409–434.

National Drug Intelligence Center (NDIC) (2011). National drug threat assessment, 2011.https://www.justice.gov/archive/ndic/pubs44/44849/44849p.pdf.

National Institute on Drug Abuse (NIDA) (2017). Trends & statistics. https://www.drugabuse.gov/related-topics/trends-statistics.

Sacks, J. J., Gonzales, K. R., Bouchery, E. E., Tomedi, L. E., & Brewer, R. D. (2015). 2010 national and state costs of excessive alcohol consumption. *American Journal of Preventive Medicine*, *49*(5), 73–79.

Terza, J. (1985). Reduced-form trinomial probit: a quantal response model without a priori restrictions. *The Journal of Business and Economic Statistics*, 3, 54-59.

Terza, J. V. (2002). Alcohol abuse and employment: a second look. *Journal of Applied Econometrics*, *17*(4), 393–404.

Terza, J.V. (2009). Parametric nonlinear regression with endogenous switching. *Econometric Reviews*, 28, 555-580.

Terza, J.V. (2016a). Inference using sample means of parametric nonlinear data transformations. *Health Services Research*, 51, 1109-1113.

Terza, J.V. (2016b). Supplementary appendix to “inference using sample means of parametric nonlinear data transformations”. *Health Services Research*

Terza, J.V. (2018). Regression-based causal analysis from the potential outcomes perspective. Unpublished manuscript, Department of Economics, Indiana University Purdue University Indianapolis.

Webb, G. R., Redman, S., Hennrikus, D. J., Kelman, G. R., Gibberd, R. W., & Sanson-Fisher, R. W. (1994). The relationships between high-risk and problem drinking and the occurrence of work injuries and related absences. *Journal of Studies on Alcohol*, *55*(4), 434–446.

Curriculum Vitae

Shichao Tang

Education

- Ph.D. in Economics, Indiana University, Indianapolis, 09/2018
- M.P.A., Indiana University, Bloomington, 05/2013
- M.A. in Economics, Renmin University of China, 06/2011
- B.A. in Economics, Renmin University of China, 06/2009

Fields of Specialization

- Applied Econometrics, Health Economics

Working Papers

- "Specification, Estimation and Testing of Treatment Effects in Multinomial Outcome Models: Dealing with Endogeneity and Independence from Irrelevant Alternatives," with Joseph V. Terza
- "Causal Effect Estimation in Models with Endogenous Treatments and Sample Selection: Implementing Gauss-Legendre Quadrature in Mata," with Joseph V. Terza
- "Binary and Nonnegative Outcomes with Endogenous Treatments: The Extended Olsen Model without Numerical Integration," with Joseph V. Terza

Professional Experience

- Statistician Intern, Eli Lilly and Company, 08/2016-03/2017
- Health Services Research Analyst, Medical Decision Modeling Inc., 05/2015-07/2015
- Statistician, IU Lilly Family School of Philanthropy, 06/2014-05/2015

- Summer Research Analyst, China Asset Management Co. Ltd., 05/2012-08/2012
- Part-time research Analyst, Bain & Company, 09/2010-12/2010

Teaching Experience

- Instructor, Introduction to Statistical Theory for Economics and Business, IUPUI, Spring 2017, Fall 2017
- Instructor, Introduction to Microeconomics, IUPUI, Spring 2018
- Teaching Assistant, Statistics Analysis for Effective Decision Making (MPA core course), IU Bloomington, Fall 2012
- Teaching Assistant, Data Analysis and Modeling (MPA core course), IU Bloomington, Spring 2013

Awards and Honors

- ICPSR Summer Program Travel Award , University of Michigan , 2015
- Graduate Assistantship, IUPUI, 2013-2018
- University Fellowship for Graduate Student, Renmin University of China , 2009-2011
- Graduated with Honor, Renmin University of China , 2009
- Scholarship for Distinguished Academic Performance, Renmin University of China, 2006-2008
- University Outstanding Student Award, Renmin University of China, 2008
- Scholarship for Distinguished Volunteer Service, Renmin University of China, 2006

Qualifications/Skills

- Computer Skills: Stata, R, SAS, Python, SQL, MS Office and LaTeX

- Language: English (fluent), Chinese (native)