UNIVERSITY OF CALIFORNIA, SAN DIEGO

Semiparametric Methods for Choice Models in Panel Data with Persistence

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Kelly C Paulson

Committee in charge:

Professor Ivana Komunjer, Chair
Professor Graham Elliott
Professor Karsten Hansen
Professor Vincent Nijs
Professor Andres Santos

2013

The dissertation of Kelly C Paulson is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2013

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Many thanks to Ivana Komunjer for her support and guidance while I developed my research agenda. All chapters of this dissertation have benefited from the insights of Andres Santos. Karsten Hansen's guidance on the empirical portion of this paper has been indispensable. An early version of this work was much improved by conversations with Hal White. Thanks also to Vincent Nijs and Graham Elliott. I am grateful for many engaging conversations with David Kaplan and Dalia Ghanem.

VITA

2005	B.A. in Economics, minor in Mathematics, Stanford University
2008	M.Sc. in Economics, London School of Economics
2013	Ph.D. in Economics, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Semiparametric Methods for Choice Models in Panel Data with Persistence

by

Kelly C Paulson

Doctor of Philosophy in Economics

University of California, San Diego, 2013

Professor Ivana Komunjer, Chair

This research explores the intersection of econometric theory and consumer choice applications. Consumer choice panel data often exhibit persistence in choices which could be explained by unobservable heterogeneity across consumers, like brand preferences, or structural state dependence, like habit formation and brand loyalty. A semiparametric method for identifying and estimating structural parameters in a binary choice model with structural state dependence in the form of a lagged choice variable is presented. The method requires the availability of auxiliary data that satisfy a conditional exogeneity assumption the additional data must adequately explain any systematic relationship between observable and unobservable components of the model. However, it is not necessary to specify the

functional form of the relationship. The distribution of the error term is also left unspecified, and certain types of serial correlation of the errors are accommodated. A constructive two-step estimation procedure is proposed.

The method is applied to consumer choice data using the IRI Academic Dataset. For a variety of datasets that may be available to marketing researchers, data that may satisfy the assumptions required for the new method is suggested. This discussion highlights specific applications where using the method described above can be helpful in disentangling structural state dependence from unobservable heterogeneity. Simulations show that the semiparametric method estimates structural state dependence better than the usual techniques. A brand choice application using data from the milk product category indicates that standard techniques may overestimate structural state dependence.

Chapter 1

Identification

Semiparametric identification of the dynamic binary choice model is a difficult problem since static model identification strategies cannot easily be extended to accommodate lagged dependent variables. Existing identification results for dynamic models require restrictive assumptions. In this paper, a conditional exogeneity assumption allows for identification in a dynamic binary choice model with a linear structure and additively separable heterogeneity. By appropriately choosing the conditioning variables, we allow for accommodation of endogenous explanatory variables and persistence though an individual fixed effect, a lagged dependent variable and serially correlated error terms.

1.1 Introduction

This paper is concerned with the application of binary choice models to panel datasets that exhibit persistence, possibly though structural state dependence. Binary choice models have been widely-used to describe individual choice behavior.

There is considerable persistence in binary outcome variables in applications in labor economics, health economics, consumer choice, industrial organization and corporate finance. Distinguishing between the different avenues of persistence is important in these applications since the different mechanisms have different policy and strategy implications.

In labor economics, structural state dependence in unemployment and labor market participation has been studied by Hyslop (1999) in the US, Arulampalam, Booth, and Taylor (2000) in the UK , Croda and Kyriazidou (2002) in Germany, Knight, Harris, and Loundes (2002) in Australia and Lee and Tae (2005) in South Korea. Policymakers are interested to know if persistence in unemployment and nonparticipation is due to individuals' unobservable characteristics or a response of labor markets to individuals' past statuses. In these models, structural state dependence can be interpreted as the "scarring" effect of unemployment - past unemployment causing an individual to be less likely to leave unemployment. Empirical results indicating significant structural state dependence have motivated governments to subsidize firms that hire new employees from long-term unemployment.

In consumer choice applications, data sometimes exhibits persistence in brand choice. Browning and Carro (2009) have studied product choice using binary choice models, while Dubé, Hitsch, and Rossi (2010) have used multinomial models to examine brand choice. In these applications, differentiating between unobservable heterogeneity and structural state dependence has important implications for pricing strategy. Persistence from structural state dependence could be due to consumer habit formation or switching costs. It may also be possible to observe variety-seeking behavior in which a consumer's purchase of a good in an earlier period causes them to be less likely to purchase the good in the current period. Structural state dependence induces firms to set prices dynamically. These pricing implications, discussed in Dubé, Hitsch, Rossi, and Vitorino (2008) and Dubé, Hitsch, and Rossi (2009), may have a sizable effect on firm profits, market structure and the level of competition within the market.

For $i = 1, ..., N$ and $t = 1, ...T$, the dynamic binary choice model with unobservable heterogeneity has the form:

$$Y_{it} = 1\{X'_{it}\beta + \gamma Y_{it-1} + A_i + U_{it} \geq 0\} \tag{1.1}$$

Note that $Y_{it} \in \{1, 0\}$. A latent variable $X'_{it}\beta + \gamma Y_{it-1} + A_i + U_{it}$ determines the value of Y_{it} in a threshold-crossing fashion. There are k observable explanatory variables contained in the vector X_{it}. The lagged dependent variable Y_{it-1} is also observable. The individual-specific scalar effect A_i and the individual-time-specific scalar shock U_{it} are not observable.

In this model, persistence in behavior is allowed through three mechanisms:

1. Unobservable heterogeneity A_i
2. Direct effect of lagged dependent variable Y_{it-1}
3. Possible serial correlation in U_{it}

The question crucial to identification is how persistence observed in the data is allocated between these possible sources. The literature on static binary choice models attributes all persistence to the individual effect. While this may be plausible in some circumstances, the direct effect of the lagged dependent variable and serial correlation of the error terms may be important in many applications. A strategy that allows for the identification and estimation of the relative magnitudes of these avenues of persistence is superior to strategies that necessitate *ex ante* assumptions about the importance of each mechanism.

Existing identification results Heckman (1981) addressed the nature of state dependence in binary choice models. Heckman differentiates between "spurious" state dependence (persistence due to unobservable heterogeneity or lingering shocks) and "structural" state dependence (persistence due to the direct effect of the lagged dependent variable). Several recent developments in the semiparameteric binary choice model literature are related to the method presented here.

For static binary choice models with exogenous explanatory variables and unobservable heterogeneity in the form of additive unobservable heterogeneity, Magnac (2004) extends the conditional logit approach developed by Rasch (1960) and Andersen (1973), and formalized by Chamberlain (1984). The identification

strategy relies on the sum of the dependent variable for each individual being a "sufficient statistic" for the individual fixed effect. This method requires assuming that state dependent occurs *only* through the unobservable heterogeneity. State dependence through the serial correlation in the error term or a lagged dependent variable are explicitly ruled out. Although Magnac describes general conditions under which the sufficient statistic identifies the fixed effect, it is not possible to extend this approach further to allow for a lagged dependent variable. To extend the model in this way, it would be necessary to specify how the persistence is split between the individual fixed effects and the lagged dependent variable. Making an assumption of this sort is undesirable as this is what we are trying to estimate from the data.

For static binary choice models with endogenous variables, a recent topic of interest has been semiparametric identification and estimation using a control function approach (Blundell and Powell (2004), Hoderlein (2008)). Although the lagged dependent variable is an endogenous variable, the methods proposed in this literature are not well-suited to the lagged dependent variable problem because they require a complete vector of classical instruments to explain the variation in the the lagged dependent variable. Generally there is not a plausible instrument available for the lagged dependent variable.

Hoderlein and White (2010a) consider identification of marginal effects in nonseparable models with fixed effects using differencing intuition, and illustrate their technique using the static binary choice model. They consider observations for which variables that are not of immediate interest are not changing, and calculate the marginal effect of variables of interest using differencing intuition. Although this is a clever extension of differencing intuition from the linear case to a nonlinear situation, it does not allow for a lagged dependent variable. It is not possible to include Y_{it-1} as an explanatory variable since it violates Hoderlein and White's definition of exogeneity. Treating Y_{it-1} as a conditioning variable is not possible either since it violates necessary stationarity conditions.

Several recent contributions can accommodate a lagged dependent variable. Arellano and Carrasco (2003) present a random effects identification strategy for binary choice panel models with predetermined variables. The unobservable individual-specific time-varying component U_{it} is assumed to have a specific distribution, conditional on lags of X_{it} and Y_{it-1}. Also, Wooldridge (2005) presents a parametric identification strategy for average partial effects in the dynamic binary choice model. It involves specifying the relationship between the A_i, the initial condition y_{i0} and x_i, as well as the distribution of U_{it}. Although a clever solution to the initial conditions problem for short panels, the approach relies on correct specification of the distribution. Although it is possible to choose a flexible distribution, it is still desirable to show nonparametric identification.

Honorè and Kyriazidou (2000) propose a semiparametric identification strategy that uses a differencing intuition to accommodate an individual fixed effect and identify β and γ in Equation (1.1). By considering only observations that meet certain conditions, they are able to invoke conditional maximum score results from Manski (1987) to identify β and γ. However, the observations that meet these conditions are not necessarily representative of the sample or population. For instance, they consider trends in the dependent variable {1,0,1,1} and {0,1,0,0}, only when the exogenous variables between the second and third observations are equal. In many applications, observations with this sort of variation may not be representative of the true variation in the population.

Honorè and Lewbel (2002) employ a "special regressor" approach to identifying the dynamic binary choice model, based on Lewbel (2000). They require a variable V_{it} with unbounded support that affects the dependent variable. Also V_{it} should to be independent of the sum of unobservables, conditional on X_{it}. The unbounded support assumption ensures that $-V_{it}$ can take all values that $\beta X_{it} + \gamma Y_{it} + A_i + U_{it}$ can take, so that the probability that $Y_{it} = 1$ can be arbitrarily close to 0 or 1. The approach presented here also relies on a relevant explanatory

variable with large support, but extends the special regressor approach in the dynamic binary choice setting by requiring a different and possibly less stringent exogeneity condition

.

Honorè and Tamer (2006) approach the identification problem from another direction. Instead of considering assumptions that, if satisfied, would lead to point identification of model parameters, they bound the range of possible parameters using linear programming methods, making only mild assumptions. In simulations, they find that when the true value of γ is small, bounds will be very tight around true values of β and γ. However, for structurally persistent processes, bounds are very large and possibly uninformative. This emphasizes the restrictiveness of the assumptions used to point-identify β and γ in this literature.

Chernozhukov, Fernàndez-Val, Hahn, and Newey (2010) investigate the possibility of using a linear model to estimate the marginal effect in a binary choice framework. For Equation (1.1), they show that generally linear fixed effects estimators for parameters are not identified, but it is possible to identify bounds around the marginal effect of interest. Further, they show that the bounds can become small (and hence informative) when T gets large or when more restrictive assumptions are made (for instance, specifying the distribution of Y_{it} conditional on X_{it} and A_i). However, their approach has several drawbacks that prevent it from being useful in an applied context. Their results hold only for discrete variables, and the discretization of continuous variables makes the bounds somewhat arbitrary. Also, the result that the bounds shrink as T grows relies on the assumption that the support of X_{it} is also the marginal support of X_{it}. This is a strong assumption in applications like labor supply, where it would be unlikely that some explanatory variables, education level for instance, vary on the entire support of X_{it}.

Although the recent work on bounding relevant parameters in dynamic binary choice models is promising, it is not useful from a practical perspective at

this point because it is not clear how to do inference with bounds. Chernozhukov, Fernàndez-Val, Hahn, and Newey (2010) present some inference results, but note that it is not straightforward. For this reason, the development of point identification results for dynamic binary choice models that are more widely applicable than existing literature is a research priority.

Approach This paper proposes identification of β and γ using a two-step method - first identifying the coefficients of the continuous variables, then using them to identify the coefficients on the binary variables.

Utilizing conditioning covariates that allow for conditional independence assumptions to be made, it is possible to equate conditional probabilities observed in the data with a function G characterized by the distributions of the error term and the unobservable heterogeneity. Then, by exploiting the linear structure of Equation (1.1), it is possible to show that the derivative of the joint distribution with respect to every x_ℓ^c is equal to β_ℓ multiplied by the derivative of the joint distribution that does not depend on ℓ. Then, the ratio of β_{ℓ_1} and β_{ℓ_2} is identified for any ℓ_1 and ℓ_2. The normalization $\beta_1^c = 1$ provides point identification.

To identify the binary variable coefficients, it is first necessary to identify the function G. Using the subset of data for which all binary variables are equal to zero, it is possible to trace out G by varying $x^c\beta^c$ along its entire support. This is possible because of the large-support continuous regressor. When G has been identified, it is possible to identify a binary variable coefficient by considering the subset of data for which all binary variables are equal to zero except for the coefficient of interest. It will be possible to equate the observed conditional probability with G as a function of known components $x^c\beta^c$ (since β^c was identified in the first step) and conditioning covariates q, and the binary coefficient of interest. As long as G is invertible, which occurs when, for instance, the distribution of the error term has positive density at any point along the real line, then it will be possible to solve for the binary coefficient of interest in terms of known quantities. This process can

be repeated for each binary variable, including the lagged dependent variable.

Contributions The identification strategy described here contributes to the dynamic binary choice model literature by providing an identification strategy that has has several advantages relative to existing methods. First, it can accommodate endogeneity when classical instruments are not available and without the parametric assumptions required by some control function methods. Second, both continuous and binary endogenous variables can be treated with the same identification strategy. Also, this identification strategy is unique in the dynamic binary choice literature in accommodating serial correlation in U_{it} when suitable conditioning covariates are available.

Equation (1.1) can be interpreted as a transformation model in which the explanatory variables enter linearly and are transformed by the indicator function. The method of identifying β^c through the ratio of derivatives of CDFs is similar to methods that have been used in the transformation model literature (Ridder (1990), Horowitz (1996), Chiappori, Komunjer, and Kristensen (2011)). Typically in this literature, the transformation is assumed to be invertible. However, this paper shows that similar methods can be used to identify a model in which the transformation is not smooth, but rather an indicator function. Integrating over the distribution of unobservable heterogeneity provides a sufficiently smooth function to apply the derivative ratio method. In this way, the results presented here extend the transformation model literature to accommodate transformations that are not invertible

Also, this paper extends the class of models for which structural parameters can be identified using conditioning covariates (Altonji and Matzkin (2005), Hoderlein and Mammen (2007) and Imbens and Newey (2009)) to include the dynamic binary choice model. Hoderlein and White (2010b) detail identification results for nonseparable panel data models with a lagged dependent variable when the dependent variable is a smooth function of the independent variables, but it is

not possible apply these results when the indicator function is present. Hoderlein and White (2010a) present identification results utilizing conditioning covariates for static binary choice models, but their assumptions preclude specifications including a lagged dependent variable.

Outline The main identification result is contained in Section 1.2. Section 1.2.1 discusses the assumptions sufficient for identification of β and γ in Equation (1.1). Section 1.2.2 details the main identification result. Section 1.2.3 shows that under additional assumptions, it is possible to incorporate more information into the identification strategy. Section 1.3 will compare the assumptions required here with key assumptions utilized in other identification strategies and summarize the advantages of the method presented here.

1.2 Identification Result

1.2.1 Assumptions

This section outlines the necessary assumptions for identification of β and γ in Equation (1.1). Identifying these objects is of interest to researchers because it allows for comparison of the relative magnitudes of different explanatory variables. In particular, it is possible to compare the effect of the lagged dependent variable with the other variables and do inference on the estimated coefficients.

We assume the following in order to identify β and γ in Equation (1.1).

A1 Data generating process

(i) Equation (1.1) holds.

(ii) There are at least two continuous variables in X_{it}. Also, X_{it} may contain binary variables. Denote continuous components with superscript c and binary components with superscript b: $X_{it} = [X_{it}^c \; X_{it}^b]$, with coefficients $\beta = [\beta^c \; \beta^b]$. X_{it} is a $[k \times 1]$ vector.

(iii) A vector of conditioning instruments Q_{it} exists. These instruments are related to the determinants of the dependent variable in a way that will be specified below.

(iv) Y_{it}, Y_{it-1}, X_{it}, and Q_{it} are observable. The individual-specific scalar effect A_i and the individual-time-specific scalar shock U_{it} are not observable.

Correct specification as described in **A1 (i)** is assumed in all of the papers discussed above. Identification of parameters when Equation (1.1) is misspecified is an open research topic. However, as will be discussed in Section 1.2.2, this method produces a variety of overidentifying restrictions that can be used to test the specification of Equation (1.1). Of particular interest is testing whether β and γ are common across individuals and if the linear structure is appropriate.

Note that **A1** accommodates both continuous and binary variables. Existing literature on endogeneity in static binary choice models allows for only continuous endogenous regressors (Blundell and Powell (2004), Hoderlein and White (2010a)) or only discrete endogenous regressors (Lewbel (2000)). Hoderlein (2008)'s method can accommodate both discrete and continuous endogenous regressors, but his results identify average structural effects rather than parameters.

A2 Large support variable

For at least one element in X_{it}^c, $X_{it}^{c^*}$, the support of $X_{it}^{c^*}$ is the entire real line.

As per Chamberlain (2010)'s impossibility result, it is necessary to have at least one variable with large support for identification in this framework. The approach here is similar to the special regressor approach suggested by Honorè and Lewbel (2002). However, the large support regressor may be endogenous, unlike Honorè and Lewbel's requirement.

A3 Normalization

$$\beta_1^c = 1$$

A scale normalization is required, in line with identification in binary choice models generally. Without loss of generality, the coefficient on the first component of X_{it}^c is normalized to 1.

A4 Conditional Exogeneity

$$U_{it} \perp X_{it}, Y_{it-1}, A_i \mid Q_{it} \tag{1.2}$$

$$A_i \perp X_{it}, Y_{it-1} \mid Q_{it} \tag{1.3}$$

A conditional exogeneity assumption accommodates endogeneity and allows for identification of the effect of the lagged dependent variable. One benefit of this identification strategy is the flexibility that comes from the researcher's ability to select components of the conditioning matrix as suggested by economic theory. A detailed discussion about when appropriate conditioning instruments are likely to be available and how to select them can be found in Chalak and White (2006).

Note that it is possible to accommodate serial correlation in U_{it} when an appropriate Q_{it} is available. This seems plausible when, for instance, we observe a sufficiently good proxy variable for a process causing serial correlation. The methods discusses earlier are unable to accommodate serial correlation, with the exception of Honorè and Kyriazidou (2000). Since their method relies on Manski (1987) which allows for some types of serial correlation, it may be possible to extend Honorè and Kyriazidou (2000)'s result to incorporate mild forms of serial correlation.

Here, it is sufficient to make two assumptions about the distribution of unobservables.

A5 Characteristics of U_{it}

The pdf of U_{it} has positive density everywhere on the real line.

A6 Differentiability

The distribution of $-U_{it}$ conditional on Q_{it} is unknown and denoted by $G(u, q)$. The distribution of A_i conditional on X_{it}, Y_{it-1} and Q_{it} is unknown and denoted by $F(a|x, y, q)$. It is necessary that $\tilde{G}(x + a, q) \equiv \int G(x + a, q)dF(a|x, y, q)$ be differentiable with respect to the first argument.

Beyond this, no assumption about the functional form of A_i and U_{it} are required. This feature alleviates the need to make functional form assumptions as in Arellano and Carrasco (2003), Wooldridge (2005) and the parametric literature. Robustness to functional form assumptions is very important since economic theory rarely gives researchers as to what sort of functional form to anticipate. Moreover, functional form assumptions sometimes induce arbitrary features. For instance, Arellano and Carrasco (2003) assume that for some period t^*, the difference between U_{it} and the conditional expectation of A_i has a normal distribution, implying that for other periods, the difference does not have a normal distribution. This leads to the uncomfortable implication that the researcher must arbitrarily select one period to have a normal distribution in a setting where the impact of misspecification are unknown.

1.2.2 Main result

The method presented here shows that it is possible to use conditional probabilities observed in the data to identify parameters of interest.

Theorem 1

Under A1 - A6, β and γ are identified.

The proof of this result requires three steps - first identification of β^c, then identification of an intermediate function, which allows for identification of the coefficients on the binary variables, β^b and γ.

Identification of β^c Suppose that unobservable fixed effect A_i were observable. Then it would be possible to observe the conditional probability $Pr(Y_{it} = 1|X_{it} = x, Y_{it-1} = y, Q_{it} = q, A_i = a)$.

$$\begin{aligned} Pr(Y_{it} = 1|X_{it} = x, Y_{it-1} = y, Q_{it} = q, A_i = a) &= Pr(x\beta + \gamma y + a + U_{it} \geq 0) \\ &= Pr(x\beta + \gamma y + a \geq -U_{it}) \end{aligned}$$

$G(\cdot, \cdot)$, a function of $-U_{it}$ and Q_{it}, can be evaluated at $X_{it} = x$, $Y_{it-1} = y$, $A_i = a$ and $Q_{it} = q$. X_{it}, Y_{it-1} and A_i enter in a linear fashion as described in **A1** so that the first argument is the latent value at which the marginal distribution is evaluated $(x\beta + \gamma y + a)$. The second argument is q, which determines the shape of the distribution.

As long as $U_{it} \perp X_{it}, Y_{it-1}, A_i \mid Q_{it}$, it is possible to equate

$$Pr(Y_{it} = 1 \mid X_{it} = x, Y_{it-1} = y, A_i = a, Q_{it} = q) = G(x\beta + \gamma y + a, q) \qquad (1.4)$$

Since A_i is not actually observable, it is not possible to observe $Pr(Y_{it} = 1|X_{it} = x, Y_{it-1} = y, A_i = a, Q_{it} = q)$ in the data. Instead, $Pr(Y_{it} = 1|X_{it} = x, Y_{it-1} = y, Q_{it} = q)$ is observable, as are less specific quantities like $Pr(Y_{it} = 1|X_{it} = x, Q_{it} = q)$.

The identification method proposed here has the advantage of identifying β and γ for any combination of (x, y, q).

It is desirable to use as much information from the data as possible in identification and estimation. The method proposed here also has the advantage that averaging can be done across binary variables so that β^c can be identified for every (x^c, q) rather than for every (x^c, x^b, y, q) using the less specific conditional probabilities mentioned above. This extension will be discussed in Section 1.2.3.

By integrating A_i out of Equation (1.4), it is possible to use observable

conditional probabilities to identify β^c.

$$\begin{aligned} Pr(Y_{it}=1|X_{it}=x, Y_{it-1}=y, Q_{it}=q) &= \int G(x\beta+\gamma y+a, q)dF(a|x,y,q) \\ &= \int G(x\beta+\gamma y+a, q)dF(a|q) \\ &\equiv \tilde{G}(x\beta+\gamma y, q) \end{aligned} \tag{1.5}$$

The second equality is a consequence of $A_i \perp X_{it}, Y_{it-1} \mid Q_{it}$. Denote this quantity by $\tilde{G}(\cdot,\cdot)$.

$\tilde{G}(\cdot,\cdot)$ is a function characterized by the joint distribution of A_i and U_{it}. A topic for further investigation is how deconvolution methods can be applied to $\tilde{G}(\cdot,\cdot)$. It would be interesting to know what additional assumptions are needed in order to recover $G(x\beta+\gamma y, q)$ and $F(x\beta+\gamma y|q)$ from $\tilde{G}(x\beta+\gamma y, q)$.

From here, it is possible to identify β^c as long as there are two continuous variables. For continuous variables x_1^c and x_2^c,

$$\frac{\partial}{\partial x_1^c} Pr(Y_{it}=1|X_{it}^c=x^c, X_{it}^b=x^b, Y_{it-1}=y, Q_{it}=q) = \beta_1^c \tilde{G}'(x^c\beta^c + x^b\beta^b + \gamma y, q) \tag{1.6}$$

$$\frac{\partial}{\partial x_2^c} Pr(Y_{it}=1|X_{it}^c=x^c, X_{it}^b=x^b, Y_{it-1}=y, Q_{it}=q) = \beta_2^c \tilde{G}'(x^c\beta^c + x^b\beta^b + \gamma y, q) \tag{1.7}$$

where $\tilde{G}'(x^c\beta^c+x^b\beta^b+\gamma y, q)$ is the derivative of $\tilde{G}(x^c\beta^c+x^b\beta^b+\gamma y, q)$ with respect to the first argument. Recall that **A6** guarantees that this quantity exists.

Then, note that the ratio of Equation (1.7) and Equation (1.6) identifies β_2^c

relative to β_1^c.

$$\frac{\frac{\partial}{\partial x_2^c} Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = x^b, Y_{it-1} = y, Q_{it} = q)}{\frac{\partial}{\partial x_1^c} Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = x^b, Y_{it-1} = y, Q_{it} = q)} = \frac{\beta_2^c \tilde{G}'(x^c\beta^c + x^b\beta^b + \gamma y, q)}{\beta_1^c \tilde{G}'(x^c\beta^c + x^b\beta^b + \gamma y, q)}$$
$$= \frac{\beta_2^c}{\beta_1^c}$$

It is necessary to use a normalization in order to identify β^b and γ. Justification for the normalization will be given in Section 1.2.2. With β_1^c normalized to 1 under **A3**,

$$\frac{\frac{\partial}{\partial x_2^c} Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = x^b, Y_{it-1} = y, Q_{it} = q)}{\frac{\partial}{\partial x_1^c} Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = x^b, Y_{it-1} = y, Q_{it} = q)} = \beta_2^c \tag{1.8}$$

Using the ratio of quantities in which certain components cancel in order to identify effects of interest can also be found in Chiappori, Komunjer, and Kristensen (2011) for transformation models and Hoderlein (2008) for identification of average structural effects in static binary choice models.

This method can be extended to any number of continuous variables. Any β_ℓ^c can be identified from

$$\frac{\frac{\partial}{\partial x_\ell^c} Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = x^b, Y_{it-1} = y, Q_{it} = q)}{\frac{\partial}{\partial x_1^c} Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = x^b, Y_{it-1} = y, Q_{it} = q)} = \beta_\ell^c \tag{1.9}$$

Identification of $\tilde{G}$ Prior to showing the identification of γ and β^b, it is necessary to identify $\tilde{G}$. To do this, consider observations for which $Y_{it-1} = 0$ and $X_{it}^b = 0$. Then,

$$Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = 0, Y_{it-1} = 0, Q_{it} = q) = \tilde{G}(x^c\beta^c, q) \tag{1.10}$$

Since we observe x^c and β^c was shown to be identified above, it is possible to identify $\tilde{G}(\cdot, q)$ by varying X_{it}^{c*}, the large support variable, over its entire support and tracing out the function as the first argument changes, holding the

second argument constant.

Recall **A3**, the normalization $\beta_1^c = 1$. Without this normalization, we could identify the ratios of any $\beta_{\ell_1}^c$ to any $\beta_{\ell_2}^c$, but it would not be possible to identify $\tilde{G}(\cdot,\cdot)$, γ or β^b.

To see this, suppose $\frac{\beta_2^c}{\beta_1^c} = 2$. It is the case that $\{\beta_1^c, \beta_2^c\} = \{1, 2\}$ is consistent with this ratio, as well as $\{\beta_1^c, \beta_2^c\} = \{2, 4\}$. When $\{\beta_1^c, \beta_2^c\} = \{1, 2\}$, it is possible to trace out a function $\tilde{G}(\cdot, q)$ by varying the large support variable in $\tilde{G}(1x_{it1}^c + 2x_{it2}^c, q)$. Suppose here the large support variable is x_{it1}^c. However, it is also possible to trace out a function $\tilde{G}(\cdot, q)$ by varying x_{it1}^c in $\tilde{G}(2x_{it1}^c + 4x_{it2}^c, q)$. The $\tilde{G}(\cdot, q)$ function traced out in these cases is generally not the same. In other words, without the normalization, there are many possible $\tilde{G}(\cdot, q)$ functions that are observationally equivalent, and an econometrician observing the ratio $\frac{\beta_2^c}{\beta_1^c}$ cannot distinguish between many possibilities for $\tilde{G}(\cdot, q)$.

It will be shown that identification of β^b and γ depends on $\tilde{G}(\cdot, q)$. When $\tilde{G}(\cdot, q)$ is not uniquely identified, it will not be possible to identify β^b and γ. Normalizing $\beta_1^c = 1$ pins down $\tilde{G}(\cdot, q)$, allowing β^b and γ to be identified. Normalizations of this sort are not thought to be restrictive in this literature.

Identification of β^b and γ With β^c and $\tilde{G}$ identified, it is possible to identify the coefficients on the binary variables.

First, consider identification of γ. For observations with $Y_{it-1} = 1$ and $X_{it}^b = 0$,

$$Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = 0, Y_{it-1} = 1, Q_{it} = q) = \tilde{G}(x^c \beta^c + \gamma, q) \tag{1.11}$$

Since β^c and $\tilde{G}(\cdot, q)$ are identified, and x^c and $Pr(Y_{it} = 1 | X_{it}^c = x^c, X_{it}^b = 0, Y_{it-1} = 1)$ are observable, it is possible to identify γ if $\tilde{G}$ is invertible.

To show invertibility, it is sufficient to show that:

$$\frac{\partial}{\partial t}\tilde{G}(t,q) > 0$$

A6 ensures that $\tilde{G}$ is differentiable. It remains to show that $\tilde{G}$ is strictly a increasing function over its entire domain. Two observations enable us to conclude this. First, $\tilde{G}$ is increasing because it is the integral of the product of two always positive functions, as defined in Equation (1.5). Then, **A5** ensures that $\tilde{G}$ is strictly increasing since U_{it} having positive density everywhere necessitates that the value of the integral will be strictly increasing across the domain.

Then, γ is identified by:

$$\tilde{G}^{-1}(Pr(Y_{it}=1|X_{it}^c=x^c, X_{it}^b=0, Y_{it-1}=1, Q_{it}=q), q) - x^c\beta^c = \gamma \qquad (1.12)$$

Identification of β^b follows in the same way. Consider observation with $Y_{it-1}=0$, $X_{it\ell}^b=1$ and $X_{it(-\ell)}^b=0$, where $X_{it\ell}^b$ denotes the ℓ^{th} element in X_{it}^b, and $X_{it(-\ell)}^b$ describes the other variables in X_{it}^b. Then,

$$Pr(Y_{it}=1|X_{it}^c=x^c, x_{it\ell}^b=1, x_{it(-\ell)}^b=0, Y_{it-1}=0, Q_{it}=q) = \tilde{G}(x^c\beta^c+\beta_\ell^b, q)$$

and it is possible to identify β_ℓ^b.

$$\tilde{G}^{-1}(Pr(Y_{it}=1|X_{it}^c=x^c, x_{it\ell}^b=1, x_{it(-\ell)}^b=0, Y_{it-1}=0, Q_{it}=q), q) - x^c\beta^c = \beta_\ell^b$$

This technique can be used to identify the other binary variable coefficients.

Sequential Identification of Binary Variables Note that it is also possible to identify the binary variables sequentially, in addition to the method described above. Suppose γ has been identified in the way described above. It is then possible to identify β^b using data for which $Y_{it-1}=1$, in addition to the method using

$Y_{it-1} = 0$ described above. The conditional probability observed in the data is,

$$Pr(Y_{it} = 1|X_{it}^c = x^c, x_{it\ell}^b = 1, x_{it(-\ell)}^b = 0, Y_{it-1} = 1, Q_{it} = q) = \tilde{G}(x^c\beta^c + \gamma + \beta_\ell^b, q)$$

Since $\tilde{G}(\cdot, q)$ is invertible, it is possible to identify β_ℓ^b with this quantity.

$$\tilde{G}^{-1}(Pr(Y_{it} = 1|X_{it}^c = x^c, x_{it\ell}^b = 1, x_{it(-\ell)}^b = 0, Y_{it-1} = 1, Q_{it} = q), q) - x^c\beta^c - \gamma = \beta_\ell^b \tag{1.13}$$

The binary variable coefficients can be identified in any order. These properties will be used in later work to propose a specification test based on overidentifying restrictions.

1.2.3 Extensions

When additional assumptions hold, it is possible to show identification using a strategy that incorporates more data than the method described in Theorem 1. Recall that Theorem 1 shows identification of β and γ for each (x, y, q) combination. Detailed in the Appendix, Theorem 2 describes the additional conditions under which it is possible to incorporate both values of a binary dependent variable into the identification strategy for β^c rather than just one particular value of y and x^b. β^c, identified with these averages, can be used to identify γ and β^b, as before.

1.3 Discussion

This paper provides a method of identification for dynamic binary choice models that does not require functional form assumptions to be made about the distributions of unobservables. It relaxes assumptions made in Arellano and Carrasco (2003) and Wooldridge (2005). Using a two-step identification method, it is possible to identify parameters while accommodating both continuous and binary variables. Serially correlated errors can also be incorporated. The method is useful when suitable conditioning covariates are available, and can be used when

classical instruments are not available. Broadening the horizon of special regressor literature, this method can be used when large-support variables that do not meet the requirements of Honorè and Lewbel (2002) are available.

Parameter identification results for any (x, y, q) provide two advantages. First, this strategy makes it possible to provide identification using data from all (or any) individuals. Compared to Honorè and Kyriazidou (2000), which uses data from only individuals who exhibit frequent switches in Y_{it}. Also, the method produces a variety of overidentifying restrictions for parameters identified with various (x, y, q) combinations. This allows for specification testing. In light of recent results by Browning and Carro (2009) suggesting that the specification described in Equation (1.1) does not allow for enough unobservable heterogeneity for common applications, this property may be very useful.

Chapter 2

Estimation

This chapter proposes a nonparametric technique to estimate a binary choice model with unobservable heterogeneity and a lagged dependent variable. The estimation technique, a two-step method, employs a ratio of weighted average derivative estimators and a quantile estimator in order to estimate structural parameters. These estimators are shown to be asymptotically normal and converge at the parametric rate, despite the nonparametric set-up. A model specification test for the stability of structural parameters over time is suggested.

2.1 Introduction

Section 2.2 explains notation and recalls relevant identification results from Chapter 1. Section 2.3 proposes estimators for β and γ. Section 2.4 describes the asymptotic behavior of the proposed estimators.

2.2 Notation and identified objects

2.2.1 Notation

Y_{it}, X_{it} and Q_{it} are random variables. Realizations of random variables are denoted by lower case letters. Note that y denotes the realization of the lagged dependent variable (Y_{it-1} instead of Y_{it}).

X_{it} contains both continuous and binary variables. Since continuous and binary variables are treated differently, it will sometimes be clearer to denote $X_{it} = \begin{pmatrix} X_{it}^c \\ X_{it}^b \end{pmatrix}$, vectors with d^{x^c} and d^{x^b} elements, respectively. Q_{it} contains both continuous and discrete variables, sometimes denoted $Q_{it} = \begin{pmatrix} Q_{it}^c \\ Q_{it}^d \end{pmatrix}$, vectors with d^{q^c} and d^{q^d} elements, respectively. Note that X_{its} refers to the s^{th} element of the X_{it} vector, and x_s refers to the s^{th} element of the realization vector x.

2.2.2 Identification of β^c

Each component of β^c can be identified from a ratio of the derivative of the conditional probability with respect to x_s to the derivative of the conditional probability with respect to x_1 (since β_1^c is normalized to 1).

$$\frac{\frac{\partial}{\partial x_s^c} Pr(Y_{it} = 1 | X_{it} = x, Y_{it-1} = y, Q_{it} = q)}{\frac{\partial}{\partial x_1^c} Pr(Y_{it} = 1 | X_{it} = x, Y_{it-1} = y, Q_{it} = q)} = \beta_s^c \tag{2.1}$$

β_s^c is identified for each (x, y, q) combination.

2.2.3 Identification of β^b and γ

Once β^c has been identified, it is possible to use it to identify β^b and γ.

$$\tilde{G}^{-1}(Pr(Y_{it} = 1 | X_{it}^C = x^c, X_{it}^B = 0, Y_{it-1} = 1, Q_{it} = q) | q) - x^c \beta^c = \gamma \tag{2.2}$$

$$\tilde{G}^{-1}(Pr(Y_{it} = 1 | X_{it}^C = x^c, X_{sit}^B = 1, X_{(-s)it}^B = 0, Y_{it-1} = 0, Q_{it} = q) | q) - x^c \beta^c = \beta_s^b \tag{2.3}$$

Recall that

$$\tilde{G}(x\beta + \gamma y | q) = \int g(x\beta + \gamma y + a | q) dF(a | q) \tag{2.4}$$

where $g(\cdot | q)$ is a conditional pdf. $G(x\beta + \gamma y | q)$ is a conditional CDF. Assumptions

guarantee that the $G(x\beta + \gamma y|q)$ is invertible with respect to the first argument.

Note that γ and β^b are identified for each (x, y, q).

2.3 Proposed estimator

These objects can be estimated with either kernel or series estimators. Using kernels, it is possible to use a local constant estimator, local linear estimator or a higher-order local polynomial estimator. Here, a kernel estimator is proposed. This was chosen over a series estimator to be similar to estimation methods in the transformation model literature. Asymptotically, the kernel and series estimators are equivalent.

2.3.1 Local linear estimator

In particular, a local linear kernel estimator is used. This estimator has the same asymptotic variance as the standard local constant estimator, but has two advantages. It provides an estimator of both the conditional probability and the derivative of the conditional probability with respect each x. Computationally this is more straightforward that estimating the derivative of the conditional probability analytically or using perturbation methods. Moreover, since estimation of β^c requires estimating derivatives of conditional probabilities and estimation of β^b and γ requires estimation of conditional probabilities, using the local linear estimator reduces the complexity of the bandwidth selection problem. Instead of finding the optimal exponent for two types of estimators (conditional density and derivative of conditional density), it is possible to find it only for one. Also, the local linear estimator may have lower bias than the local constant estimator.

This section will present a procedure for estimating β and γ for a particular realization of (x, y, q). Methods of averaging across different realizations of (x, y, q)

will be discussed in the next section.

The local linear estimator solves the minimization problem:

$$\min_{a,b,c,d} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - a - (X_{it} - x)'b - (Y_{it-1} - y)'c - (Q_{it} - q)'d - \right)^2 \mathbf{K_h}(X_{it}, Y_{it-1}, Q_{it}, x, y, q)$$

where the generalized kernel K is the product kernel

$$\mathbf{K}_h(X_{it}, Y_{it-1}, Q_{it}, x, y, q) = \prod_{m=1}^{d^{x^c}} k\left(\frac{X_{itm}^c - x_m^c}{h}\right) 1\{X_{it}^b = x^b\} 1\{Y_{it-1} = y\} \prod_{n=1}^{d^{q^c}} k\left(\frac{Q_{itn}^c - q_n^c}{h}\right) 1\{Q_{it} = q^d\} \tag{2.5}$$

for a univariate kernel k. Properties of k will be discussed in the next section.

This minimization problem can be rewritten as a GLS problem and estimated by

$$\begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \\ \hat{d} \end{pmatrix} = \left[\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{K}_h(X_{it}, Y_{it-1}, Q_{it}, x, y, q) \begin{pmatrix} 1 \\ X_{it} - x \\ Y_{it-1} - y \\ Q_{it} - q \end{pmatrix} \begin{pmatrix} 1 & (X_{it} - x)' & Y_{it-1} - y & (Q_{it} - q)' \end{pmatrix} \right]^{-1}$$
$$\times \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{K}_h(X_{it}, Y_{it-1}, Q_{it}, x, y, q) \begin{pmatrix} 1 \\ X_{it} - x \\ Y_{it-1} - 1 \\ Q_{it} - q \end{pmatrix} Y_{it}$$

Bandwidth selection using a cross-validation method is recommended for this type of estimator.

Note that $\hat{a}(x, y, q)$ estimates $Pr(Y_{it} = 1 | X_{it} = x, Y_{it-1} = y, Q_{it} = q)$. This quantity is used to estimate γ and β^b.

As discussed in Fan and Gijbels (1996), $\hat{b}(x, y, q)$ estimates the vector of derivatives corresponding to the derivative of $Pr(Y_{it} = 1 | X_{it} = x, Y_{it-1} = y, Q_{it} = q)$ with respect to each component of X_{it}^c. These quantities are used to estimate β^c.

2.3.2 Estimator for β^c

Let $\hat{b}_s(x, y, q)$ denote the derivative of the conditional probability with respect to x_s^c.

Then, it is possible to estimate β_s^c:

$$\hat{\beta}_s = \frac{\hat{b}_s(x, y, q)}{\hat{b}_1(x, y, q)} \tag{2.6}$$

2.3.3 Estimator for β^b and γ

Prior to the estimation of β^b and γ, it is necessary to estimate $\tilde{G}$.

To do this, consider the subset of data for which $X_{it}^b = 0$ and $Y_{it-1} = 0$. The conditional probability $Pr(Y_{it} = 1|X_{it}^c = x^c, X_{it}^b = 0, Y_{it-1} = 0, Q_{it} = q)$ can be estimated with $\hat{a}(x^c, x^b = 0, y = 0, q)$, as described in Section 2.3.1.

Because

$$Pr(Y_{it} = 1|X_{it}^c = x^c, X_{it}^b = 0, Y_{it-1} = 0, Q_{it} = q) = \tilde{G}(x^c\beta^c|q)$$

and one component of X_{it}^c has a large support, it is possible to trace out the conditional CDF $\tilde{G}(\cdot|q)$ using the conditional probabilities by varying x^{c^*} along across the entire support. Then, it is possible to use the estimate of $\tilde{G}(\cdot|q)$ to estimate β^b and γ.

First consider estimation of γ. It is possible to estimate $Pr(Y_{it} = 1|X_{it}^C = x^c, X_{it}^B = 0, Y_{it-1} = 1, Q_{it} = q)$ by $\hat{\alpha}_\gamma \equiv \hat{a}(x^c, x^b = 0, y = 1, q)$.

Quantile estimation methods can be applied in order to estimate $\tilde{G}^{-1}$, evaluated at the α_γ^{th} quantile.

The estimator takes the form

$$\hat{\hat{G}}^{-1}(\hat{\alpha}_\gamma|q) = \arg\min_k|\hat{\alpha}_\gamma - \hat{\hat{G}}(k|q)|$$

From here, it is possible to estimate γ by

$$\hat{\gamma} = \hat{\hat{G}}^{-1}(\hat{\alpha}_\gamma|q) - x^c\hat{\beta}^c \tag{2.7}$$

Similarly, β_s^b can be estimated using the subset of data for which $X_{its}^b = 1$ and all other binary variables are equal to zero. It is possible to estimate $Pr(Y_{it} = 1|X_{it}^c = x^c, X_{sit}^b = 0, X_{(-s)it}^b = 0, Y_{it-1} = 0, Q_{it} = q)$ by $\hat{\alpha}_s \equiv \hat{a}(x^c, x_s^b = 1, x_{-s}^b = 0, y = 0, q)$.

Then as before, the estimator takes the form

$$\hat{\hat{G}}^{-1}(\hat{\alpha}_s|q) = \arg\min_k|\hat{\alpha}_s - \hat{\hat{G}}(k|q)|$$

and, for each binary element of X_{it}, it is possible to estimate β_s^b by

$$\hat{\beta}_s^b = \hat{\hat{G}}^{-1}(\hat{\alpha}_s|q) - x^c\hat{\beta}^c \tag{2.8}$$

2.4 Asymptotic behavior

In the previous section, Equation (2.6), Equation (2.8) and Equation (2.7) provide point estimators for β^c, β^b and γ (point estimator meaning an estimator for each (x, y, q), not an average). However, these point estimators do not converge at the $\sqrt{N}$ rate because of the derivative associated with estimation of β^c. Since β^c, β^b and γ are identified for each (x, y, q), it is possible to average the point estimates in order to achieve convergence at the $\sqrt{N}$ rate (see Powell, Stock, and Stoker (1989) for a detailed discussion of convergence rates of derivatives and averaged derivatives).

By averaging, it is possible to control for another potential problem related to the ratio form of $\hat{\beta}^c$. When estimating ratios of densities, it is possible that the denominator can be very close to zero, yielding erratic behavior in the estimator. In the literature, two methods have been used to control this problem - trimming and weighting. The trimming method excludes point estimates from inclusion in the averaged estimator when the denominator is smaller than some value (see Hardle and Stoker (1989) for an example in the average derivative context). The weighting method excludes point estimates with small denominators by giving them very little weight (see, for example, Horowitz (1996) and Chiappori, Komunjer, and Kristensen (2011)). As suggested in Horowitz (1996), it is not possible to achieve the $\sqrt{N}$ convergence rate for an averaged estimator using the trimming method, but it is possible to do so using the weighting method. For this reason, the weighting method will be implemented here.

Another difficulty in considering the asymptotic behavior of the estimator proposed in the previous section is that it requires one explanatory variable to have a large support. Khan and Tamer (2009) discuss the limitations of "irregularly identified" estimators. These models are characterized as models that

> attain identification by requiring that covariate variables take support in regions with arbitrary small probability mass. These identification strategies sometimes lead to estimators that are weighted by a density, a conditional probability or weights that take arbitrarily small values on these regions of small mass.

Khan and Tamer (2009) discuss the difficulties of estimating models with this type of estimation strategy, emphasizing that generally it is not possible to attain the parametric rate. Using Lewbel (1997) paper on binary choice models as an example, Khan and Tamer show that the rate of convergence depends on the relative thicknesses of the tails of the distributions of the large support variable and the error term. Generally, when the large support variable has fat tails (for instance, if it has a Cauchy distribution) and the error term has thinner tails (like logit or probit), then it is possible to achieve the $\sqrt{N}$ rate. However, for general or unspecified distributions, it is not possible to achieve the parametric rate

of convergence. However, typically it is still possible to show asymptotic normality.

One approach to dealing with asymptotics in this situation is to assume that the large support variable has fatter tails than the error term, and use this assumption to show the parametric rate can be achieved.

Another approach would be to acknowledge that we generally do not achieve the parametric rate with this sort of estimator, and use the method illustrated in Khan and Tamer (2009). They show that although the parameter itself does not converge at the parametric rate, a studentized version of the parameter does. This may be a good approach if Khan and Tamer (2009) provide a feasible estimation strategy. However, since the parameter estimate is used for estimation of the second step, it will be necessary to figure out how the results for the studentized estimator can be incorporated into the second-stage asymptotics.

The approach taken here exploits an aspect of this identification strategy that is not present in Lewbel's setup and not discussed by Khan and Tamer. Since it is possible to calculate β for every value of (x, y, q), even though each estimate of $\beta(x, y, q)$ may not converge at the parametric rate because it is a derivative point estimate, it may be possible to average across different values of (x, y, q) in order to achieve the $\sqrt{N}$ rate.

2.4.1 Asymptotic distribution of β^c

Denote the average across $\beta^c(x, y, q)$ as

$$\beta^c \equiv \int_{\mathcal{X}} \int_{\mathcal{Y}} \int_{\mathcal{Q}} w(x, y, q) \beta^c(x, y, q) dx dy dq \tag{2.9}$$

where $w(x, y, q)$ is the joint distribution of (x, y, q), and

$$\hat{\beta}^c = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\beta}^c(x_{it}, y_{it-1}, q_{it}) \tag{2.10}$$

where $(x_{it}, y_{it-1}, q_{it})$ denotes the realization of $(X_{it}, Y_{it-1}, Q_{it})$ that individual i has at time t.

Assume the following assumptions hold, in addition to assumptions required for identification:

Theorem 1 shows that this estimator has an asymptotically normal distribution and converges at the $\sqrt{N}$ rate.

Theorem 1 *Asymptotic distribution of average continuous coefficients*

$$\sqrt{n}(\hat{\beta}^c - \beta^c) \rightarrow^d N(0, \Omega)$$

The proof is in the Appendix.

2.4.2 Estimation of binary variable coefficients

Asymptotic results for β^b and γ are straightforward applications of arguments made in the quantile estimation literature.

Theorem 2 *Asymptotic distribution of average binary coefficients*

For each $s = 1, ..., d^{x^c}$

$$\sqrt{n}(\hat{\beta}^b_s - \beta^b_s) \rightarrow^d N(0, \Omega)$$
$$\sqrt{n}(\hat{\gamma} - \gamma) \rightarrow^d N(0, \Omega)$$

The proof is in the Appendix.

2.5 Discussion

Estimators for the dynamic binary choice model estimation strategy proposed in Chapter 1 are proposed and discussed. It is possible to use standard kernel

estimation techniques to estimate each $\beta(x, y, q)$ and $\gamma(x, y, q)$. These estimators, averaged across the empirical distribution of (x, y, q) are shown to converge at the $\sqrt{n}$ rate. A data-driven bandwidth selection strategy specific to the unique form of this estimator is proposed.

Chapter 3

Application

Distinguishing between structural state dependence and other sources of persistence is an important question in consumer choice data since different types of persistence suggest different optimal pricing policies. Typical strategies involve estimating a model with a flexible specification of unobservable heterogeneity and including information about lagged choice as an explanatory variable. It is common practice to assume that variation in prices generates sufficient variation in past choices to identify structural state dependence. For a simple model with structural state dependence and unobserved heterogeneity, this paper presents evidence suggesting that this is not the case. This evidence motivates the need for a method with a clear strategy for disentangling structural state dependence from unobservable heterogeneity. A novel semiparametric technique for identifying structural state dependence in binary choice models, developed by the author in a previous paper, is introduced in the consumer choice context. The method incorporates other types of marketing data to provide a better estimate of structural state dependence. Simulations show that this method can provide better estimates of structural state dependence than commonly-used techniques. An application of the new method to a brand choice model with structural state dependence uses IRI data from the milk product category. Results suggest that commonly-used methods overestimate the amount of structural state dependence.

3.1 Introduction

Persistence in consumer choice data has been observed often. For many consumer packaged goods product categories, consumers are more likely to purchase products they have purchased in the past. A simple explanation for this inertia is that consumers tend to purchase the product that they prefer to the alternatives. The product was purchased in the past because the consumer prefers it to the alternatives, and it is repurchased later for the same reason. Inertia in choices that is caused by persistence in unobservable preferences has been characterized as "spurious" state dependence by Heckman (1981). Another possible explanation for inertia in choices is structural state dependence. This occurs when a consumer's past decisions directly affect later choices through a causal mechanism. A variety of structural state dependence mechanisms have been considered in the literature. Dubé, Hitsch, and Rossi (2010) look for evidence of consumer search and learning processes as mechanisms for structural state dependence. As an alternative, they posit that experience with a product can induce a form of loyalty to the product that may be manifested in future purchases.

Estimating the amount of structural state dependence in consumer choice data is of interest because it has important implications for pricing. When state dependence is structural, there are important dynamic pricing incentives (Dubé, Hitsch, and Rossi (2009), Dubé, Hitsch, Rossi, and Vitorino (2008)). When state dependence is spurious, however, a dynamic pricing strategy would be suboptimal.

Researchers typically estimate choice models that allow for persistence in consumer behavior through both unobservable heterogeneity across households and structural state dependence. Models allow for unobservable heterogeneity by allowing each household to respond to marketing mix variables differently, and structural state dependence is incorporated by inclusion of lagged choice information. Keane (1997) estimates a model of ketchup brand choice using first a model with persistence only through structural state dependence, specifically a term of expo-

nentially weighted lagged choices. Then, he estimates specifications that allow for complex forms of unobservable heterogeneity too. Dubé, Hitsch, and Rossi (2010) estimate brand choice models for orange juice and margarine using a model with a flexible distribution of unobservable heterogeneity and state dependence through a lagged choice indicator. These papers find that both unobservable heterogeneity and structural state dependence are responsible for persistence in the data. Dubé, Hitsch, and Rossi demonstrate that their flexible estimators fit the data better than estimators with simpler forms of heterogeneity.

It is thought that when unobservable heterogeneity has been sufficiently captured, structural state dependence can be identified by price variation that induces variation in lagged choices. This paper argues that price variation is generally insufficient to identify structural state dependence using standard methods. Lagged choice variables confound unobservable heterogeneity and structural state dependence, even when flexible forms of unobservable heterogeneity are used. This is because a consumer's previous choices are influenced by both unobservable preferences and the unobservable structural state dependence parameter. The lagged choice variable is endogenous in the sense that it is correlated with the unobservable parameters. Methods that do not take this dependence into account can provide misleading evidence about the nature of persistence in the data.

Intuitively, the variation in observed choices due to price promotions may not provide enough information to explain how much of the persistence in the data comes from structural state dependence. Suppose a consumer with fairly persistent choices is observed. Because lagged choices depend on two unobservable factors, unobservable heterogeneity in preferences and structural state dependence, it is difficult to estimate how much persistence comes from each source. When the consumer does not often respond to other items' price promotions, it is not clear if this is because the consumer has a strong preference for the item or because of structural state dependence, based on variation in price data alone. It may be the case that there are multiple combinations of unobservable parameter values that

are consistent with the observed choice and price data.

The approach of this paper is to demonstrate, for a binary choice model, that this is a problem of practical importance, and then to propose a solution. A semiparametric method for identifying and estimating structural state dependence in the presence of unobservable heterogeneity developed in Chapter 1 and Chapter 2 is discussed in the context of consumer choice applications.

The simple binary choice model discussed here allows heterogeneity to enter only through an additive term (consumers respond to marketing mix variables in the same way and exhibit the same amount structural state dependence). The additive term can be interpreted as a consumer's unobservable preference for one brand relative to the other. This simplification makes it possible to highlight how the lagged choice variable can confound structural state dependence and unobservable heterogeneity, and explain the proposed solution in a concise manner. However, the solution proposed in Chapter 1 can be extended to models in which consumers have a heterogeneous response to marketing mix variables and exhibit different amounts of structural state dependence.

Related Literature This work is closely related to Chintagunta, Kyriazidou, and Perktold (2001). Chintagunta, Kyriazidou, and Perktold consider methods for estimating binary choice models with unobservable heterogeneity in the form of an additive fixed effect. They note that it is common in the marketing literature to include include the lagged choice variable as if it were an exogenous explanatory variable, and they discuss how this violates key assumptions of the methods and leads to inconsistent parameter estimates. In simulations and an application, they compare parameter estimates from these methods to estimates from a semiparametric method can accommodate both unobservable heterogeneity and a lagged choice variable. The semiparametric method, developed by Honorè and Kyriazidou (2000), uses a subset of the data to estimate model parameters using a strategy that differences out the fixed effects, which are allowed to be correlated with other

explanatory variables in an unspecified way. In an empirical application they find that the semiparametric method seems to produce different estimates of structural state dependence than the standard methods, and simulations show that the semiparametric method is estimating structural state dependence better than standard methods. However, as the method uses less data than other methods and kernel estimators, the standard errors are somewhat larger than with typical methods.

Proposed Solution The method discussed in this paper has several benefits over Honorè and Kyriazidou's method in the consumer choice context. It can be extended to choice models in which each household is allowed to have different responses to marketing mix variables. In the econometrics literature, Browning and Carro (2007) demonstrate that choice models allowing heterogeneity only through the additive fixed effect are generally insufficient to capture heterogeneity in consumer choice data. Many papers in the marketing literature have found that complex patterns of unobservable heterogeneity induce a better model fit than simpler heterogeneity structures (Keane (1997), Dubé, Hitsch, and Rossi (2010)). Another advantage of the method demonstrated in this paper is that it uses extra information to identify structural state dependence, rather than using only a subset of data for which explanatory variables do not change for identification.

The key assumption in Chapter 1 that facilitates identification of structural state dependence is an assumption that observable variables are independent of unobservable heterogeneity, conditional on other available information. The idea of using additional information to identify effects of interest in other dynamic choice models has been discussed by Manski (2004). Manski encourages researchers to augment statistical models with extra data in order to identify dynamic effects whenever possible, instead of relying assumptions. Manski suggests survey data in his examples. The method proposed in Chapter 1 allows for identification of structural state dependence with preference survey data, and sometimes with data that is usually available in consumer choice panels, like demographics. The method

uses the additional data as conditioning variables. These variables are required to explain the systematic portion of variation between observables and unobservables. In other words, conditional on the extra data, "enough" of the variation between observables and unobservable components of the model must be explained so that the remaining variation between observables and unobservables is not systematic. Examples of consumer choice scenarios that satisfy and do not satisfy this requirement are presented later in the paper.

Another advantage of the method proposed in Chapter 1 is that is allows for identification of state dependence in the presence of some types of serial correlation processes. Neither standard methods nor Honorè and Kyriazidou's estimator are credibly able to disentangle structural state dependence processes in the presence of serially correlated errors. This can be problematic if, for instance, it is thought that consumers might be able to predict variation in marketing mix variables. Recent work by Misra, Roberts, and Handel (2012) suggests that methods that are robust to serially correlated errors are important for empirical applications. Their proposed method, however, cannot accommodate a lagged choice variable.

Like Misra, Roberts, and Handel (2012) and Dubé, Hitsch, and Rossi (2010), the method presented here does not require distributional assumptions related to the distribution of error terms. This flexibility eliminates the possibility that misspecification of the error distribution will result in an incorrect estimate of structural state dependence.

Outline Section 3.2 will illustrate how the endogeneity of lagged choice can be problematic. Section 3.3 describes the semiparametric estimation strategy, and explains the meaning of the key assumption in the consumer choice context. Section 3.4 presents simulations results for the semiparametric estimator compared with standard methods and known structural parameters. Section 3.5 is an application of the estimator using IRI data.

3.2 Endogeneity of lagged choice

3.2.1 Binary model

Consider a typical consumer choice panel data set-up in which choices of $i = 1, ..., N$ individuals are observed over $t = 1, ..., T$ periods. Suppose that the choice made by household i in period t is $Y_{it} \in \{0, 1\}$. X_{it} can be a matrix of marketing mix components, and for simplicity here is restricted to be the difference between the natural logs of each alternative's price. Suppose that β and γ, the structural parameters describing the effect of X_{it} and the lagged choice variable Y_{it-1} , are the same across households. Let A_i denote an unobservable, time-invariant, household-specific effect. Let U_{it} denote an unobservable, household-time-specific shock. Assuming a single-index structure, the choice problem can be written as:

$$Y_{it} = 1\{X_{it}\beta + \gamma Y_{it-1} + A_i + U_{it} > 0\} \tag{3.1}$$

Note that the lagged choice variable Y_{it-1} is a function of A_i and γ in addition to the marketing mix variables and idiosyncratic shock from the previous period. When Equation (3.1) is rewritten, the problem with incorporating the lagged choice variable becomes evident:

$$Y_{it} = 1\{X_{it}\beta + \gamma Y_{it-1}(A_i, \gamma) + A_i + U_{it} > 0\} \tag{3.2}$$

When A_i is large in magnitude, lagged choices will be highly correlated with Y_{it}. A useful estimation strategy will distinguish between cases where there is persistence in the data because A_i is large in magnitude and cases where there is persistence in the data because γ is large in magnitude.

Endogeneity of lagged choice Consider the type of variation present in marketing mix variables compared to the lagged choice variable. Marketing mix vari-

ables are typically thought to be exogenous to each household's decision. Variation in price, feature and display induce households to switch between products. In contrast, variation in the lagged choice variable is caused in part by the same exogenous processes, but also by A_i and γ. Jointly, A_i and γ will determine how sensitive each household is to price variation. For instance, price variation will not induce a household with A_i that is large in magnitude relative to β to switch products. Even for a household with a small value of A_i such that it can be easily swayed by price promotions, there will still exist some correlation between current choices and lagged choices as long as A_i is not zero. It is important that a statistical method recognizes that the lagged choice variable is correlated with A_i in this way so it does not attribute the correlation between lagged choices and current choices entirely to state dependence.

More price variation reduces the correlation between current choices and lagged choices. However, it does not eliminate the problem that A_i has also influenced past choices. Some correlation between lagged choices and current choices is due to A_i and should not be attributed to structural state dependence.

Functional form Keane (1997) notes that it is not possible to nonparametrically identify unobservable heterogeneity and structural state dependence. He emphasizes that estimates of structural state dependence are conditional on the assumed functional form and discusses commonly-used specifications. The simulation and application sections of this paper consider only the lagged choice functional form of structural state dependence. This simplification is made in order to show that when even the functional form of structural state dependence is correctly specified and simple, it may not be identified from price variation alone. However, it is important to note that the endogeneity concern discussed here is problematic in all functional forms of structural state dependence. For example, it has been noted by Anderson (2002) that the Guadagni-Little form of structural state dependence confounds unobservable heterogeneity and state dependence. The identification method presented in Chapter 1 can be extended to various functional forms for

structural state dependence and allows for specification tests that can help distinguish between different parametric forms. For conciseness these topics are not addressed in this paper.

3.2.2 Evidence from simulations

As motivation, this section provides evidence from simulations that illustrate how standard methods can provide misleading estimates when the endogeneity of the lagged choice variable is not considered.

Joint distribution of observable data In the marketing literature, it is common practice to assume that it is possible to distinguish between unobservable heterogeneity and structural state dependence using the variation in lagged choices generated by variation in marketing mix variables. A household may be induced to switch away from their favored brand by another brand's price promotion. Then, if the household continues to purchase the less-favored brand after the price promotion ends, it is thought that the household exhibits structural state dependence. The difficulty in using price variation for identification arises because we do not observe the "favored brand" and "less-favored brand" in the data. Rather, we attempt to estimate the brand preferences and structural state dependence simultaneously.

The model is identified only if it is apparent from the observable data how much persistence in choices comes from structural state dependence and how much comes from preferences. Since preferences are unobserved and the lagged choice variable confounds structural state dependence and unobservable heterogeneity, it is not clear that the model is identified. Although we observe households switching between brands in response to price changes, it is not obvious how we would distinguish between preferences and structural state dependence with this information. For a particular realization of observed data, it may be the case that there are multiple ways to split up the sources of persistence that would be consistent with the observed choices. In other words, it is possible that different underlying

data generating processes can generate distributions of observable data that are very similar. For instance, it is possible that we observe data with persistence in choices because households have strong preferences. It is also possible that data with persistence could be the result of households having some preferences across brands, but also some structural state dependence.

To demonstrate this problem, it is possible to show that simulated data from models with different qualitative interpretations can generate very similar joint distributions of observable data. Figure 3.1 summarizes data that has been simulated for two qualitatively different specifications. In specification A, $\gamma = 1$ indicates structural state dependence in each household's choice. In specification B, $\gamma = 0$ indicates no structural state dependence. For specification A, the distribution of A_i is standard normal (Figure 3.1b). Many households have a draw of A_i close to zero, which gives them mild or no preference over the two alternatives. For specification B, the distribution of A_i is bimodal, with a cluster of households preferring one option, and another cluster of households preferring another option (Figure 3.1c). Few households are indifferent between the two choices. In the simulated data, overall transition probabilities and price variation are set to be similar to values observed in a sample of milk purchases from the IRI Academic Dataset used in Section 3.5.

Figure 3.1d and Figure 3.1e show that these data generating processes produce very similar distributions of observed choices, conditional on the past choice and price. A range possible values for x_{it} is ranging from 0.2 to 0.28 are along the horizontal axis. The vertical axis represents the observed probability of choosing $Y_{it} = 1$, conditional on a particular price and the previous choice. Figure 3.1d plots the probability of choosing $Y_{it} = 1$ for households with $Y_{it-1} = 0$, for a range of possible price differences. Figure 3.1e plots the same quantity for households with $Y_{it-1} = 1$. Data generated from both Specification A and Specification B are plotted. It is clear in both plots that the different specifications produce similar marginal distributions of observable variables.

Since different data generating processes generate very similar joint distributions of observables, assumptions behind the statistical methods used to analyze this data are crucial. Estimates of structural state dependence and unobservable heterogeneity will depend on how the statistical method separates these two distinct sources of persistence. A useful method would estimate γ close to 1 with data from Specification A and close to zero with data from Specification B. The next section will show that commonly-used methods that treat lagged choice as an exogenous variable does not do a good job estimating γ when the true value is close to zero, as in Specification B, even price variation increases.

Estimates with increasing price variation Since it is common practice to assume that price variation can be used to identify structural state dependence in choice models with unobservable heterogeneity, this section will present evidence from simulations showing that even as price variation increases, models that include a lagged choice variable without accounting for its endogeneity will tend to overestimate state dependence for realistic amounts of price variation.

Consider the simulations summarized in Figure 3.2. Here, data are simulated in accordance with Equation (3.1). Figure 3.2a summarizes the simulation specifications. The price parameter is set so that $\beta = -2$, and structural state dependence is set by $\gamma = 0$. This parameter value set the true amount of structural state dependence to be zero for all households, and since the errors are generated with an *iid* process, all persistence in the data comes through A_i. The question investigated here is whether or not more price variation leads to better estimates of structural state dependence, holding the distribution of unobservable heterogeneity in the data constant.

The difference in ln(price) between alternatives, x_{it}, is drawn from a normal distribution centered at 0 (the choices are on average the same price). The standard deviation of the distribution price is drawn from is varied between 0.05 and

0.45. This range was set to be similar to the amount of price variation observed in the IRI data used in Section 3.5 (standard deviation 0.18). For each price variance, 100 datasets are simulated.

The standard pooled logit (or random effects logit) is run treating the lagged choice variable as any other explanatory variable. This method, along with the conditional (fixed effects) logit approach, was discussed in Chintagunta, Kyriazidou, and Perktold (2001). They explain how the lagged choice variable violates necessary exogeneity conditions and how in the marketing literature, lagged choice variables are commonly included as exogenous variables despite violating these assumptions. In order for the pooled logit procedure to consistently estimate structural state dependence, it would be necessary to assume that the initial value of the lagged choice variable (y_{i0}) it is unrelated to unobservable components of the model (A_i). In commonly-purchased product categories, it seems very unlikely that each household's first purchase in a typical scanner panel dataset is unrelated to each household's value of A_i. It is much more plausible that each household's first purchase is correlated with A_i.

Figure 3.2b and Figure 3.2c show average coefficient estimates across the 100 simulated datasets at each level of price variation. For all price variation levels, the averaged estimated γ is larger than zero, the true value of γ. Moreover, the averaged estimated value of γ does not get substantially closer to the true value as the level of price variation increases. Note that as the amount of unobservable heterogeneity increases between Figure 3.2b and Figure 3.2c, the estimated amount of structural state dependence also increases even though the true value stays the same. Intuitively, this is because the correlation between past choices and current choices increases as draws of A_i increase and y_{i0} is more influenced by A_i. However, when estimating structural state dependence, we are not interested in knowing how strong the correlation between current choices and lagged choices is, but rather how much of the correlation can be explained by a structural state dependence mechanism rather than unobservable heterogeneity across households.

From these simulations, it is evident that identification of structural state dependence is a problem of practical importance. If the objective of estimation is to determine if a firm has a dynamic pricing incentive, a firm would be misled by estimates of structural state dependence using a method that does not incorporate the endogeneity of lagged choice. Section 3.3 presents a method for estimating Equation (3.1) with a clear strategy for disentangling structural state dependence from unobservable heterogeneity.

3.3 Accommodating the endogeneity of lagged choices

It is evident that estimating models with unobservable heterogeneity and structural state dependence must be done with care, in order to account for the correlation of lagged choices with unobservable quantities. Although the technique presented in Chintagunta, Kyriazidou, and Perktold (2001) seems to provide a feasible semiparametric method for estimating Equation (3.1), its appropriateness for the marketing context is limited since it cannot be extended to models with household-specific parameters. The method presented here is an alternative to their method in the context of Equation (3.1), yet it is more useful in the marketing literature since it can be extended to models with household-specific parameters.

This section will discuss the key identifying assumption in the context of consumer choice analysis, and other important aspects of the estimation procedure. Full details can be found in Chapter 1.

3.3.1 Key identifying assumption in practice

This method is able to distinguish between structural state dependence and unobservable heterogeneity though A_i using additional information, denoted Q_{it}. Q_{it} is selected by the researcher so that, conditional on Q_{it}, the joint distribution of

the unobservable elements of the model is independent from the joint distribution of observables. It is possible to decompose this assumption into two components.

$$U_{it} \perp X_{it}, Y_{it-1}, A_i \mid Q_{it} \tag{3.3}$$

$$A_i \perp X_{it}, Y_{it-1} \mid Q_{it} \tag{3.4}$$

In the marketing context, it is typically assumed that $U_{it} \perp X_{it}, Y_{it-1}, A_i$ holds without conditioning on Q_{it}. Household-specific shocks are assumed to be unrelated to marketing mix variables, and shocks are assumed to be *iid* across periods. For this reason, the discussion will focus on data available to marketing researchers that could be included in Q_{it} in order to satisfy Equation 3.4.

Intuitively, to satisfy Equation 3.4, Q_{it} must be selected so that "enough" of A_i is explained to disentangle the effect of A_i from structural state dependence. Unlike parametric methods, it is not necessary to model A_i explicitly by specifying a functional form for A_i as a function of Q_{it} — instead, Q_{it} enters only as a matrix of conditioning variables.

Lagged prices as conditioning variables In typical marketing applications, it is believed that variation in past prices generates variation in lagged choices that can disentangle state dependence from unobservable heterogeneity. As discussed above, this variation cannot be used to identify structural state dependence. However, it is an interesting exercise to consider what we would have to believe about the nature of unobservable heterogeneity in order to believe that Q_{it} composed of just prices could identify structural state dependence. The key condition this implies is:

$$A_i \perp Y_{it-1} \mid P_{it-1}, P_{it-2}, P_{it-3}, ... \tag{3.5}$$

For price variation alone to identify model parameters, it would be necessary

to believe that among households that were exposed to the same sequence of prices in the past, there is no systematic relationship between A_i and Y_{it-1}. This seems implausible since we might think that even among households that were exposed to the same sequence of past prices, households with a strong preference for Brand 1 (large A_i) would tend to have a realization of Y_{it-1} that is Brand 1. For this reason, it is necessary to consider other types of conditioning data.

Demographics Consumer choice datasets contain information about the households. These variables can range from basics like household size to more detailed information like household income and education levels. In the context of the model described in Equation (3.1), it is interesting to consider how this information can help to disentangle the effect of structural state dependence from unobservable heterogeneity. Suppose, as an example, information about the age of the primary shopper and household income is available as conditioning variables. The key condition then becomes

$$A_i \perp Y_{it-1} \mid age_{it}, income_{it} \tag{3.6}$$

Is it plausible that among households with the same age and income, A_i and Y_{it-1} are independent, or would we expect a systematic relationship between them? Suppose we consider all households with a shopper age 65-70 and income of \$150,000+. In this subset, do we still expect to see a systematic relationship between A_i and Y_{it-1}? If we believe that people with the same conditioning variables are similar enough in A_i that we expect to see a distribution of Y_{it-1} that is not systematically related to A_i, then the assumption is satisfied. Another way of thinking about this is that if we could observe Y_{it-1} for the group specified by the conditioning variable, would it provide any additional information about A_i? The answers to these questions depend on the particular application. To clarify this, two specific applications will be discussed - one in which it seems reasonable that this assumption is satisfied, and one in which it does not seem to be satisfied.

Suppose the application is a brand choice model in which consumers choose

between national brand products and private label products. Suppose that A_i denotes the unobservable relative propensity for household i to purchase the national brand over the private label brand. The key assumption is satisfied when we believe that for subsets of the population with the same observable characteristics Q_{it}, observing Y_{it-1} doesn't give us any information about A_i. There is evidence that older people tend to associate private label products with low quality, and younger people may not even be aware that a private label brand is not a national brand. So, if we consider households headed by a 65-70 year old shopper with a high income, we can assume that generally these people will be uninterested in private label brands. If in the data these people usually purchase a national brand, it does not tell us much about any individual's A_i. This does not mean we are assuming that all people in a particular age range and income level have the same preference — each household within the set of households specified by the conditioning variables has a draw of A_i from a random, unspecified distribution. However, the assumption is that observable information tells us enough about A_i that we cannot infer anything else about it from observing Y_{it-1}.

Variation in price and other marketing mix variables can aid identification here. When there is a lot of variation in these variables and households are switching frequently, Y_{it-1} will not provide much information about A_i since any value of Y_{it-1} may be due to price variation or preferences. The conditional exogeneity assumption will be satisfied more easily with more price variation.

Another type of demographic data that might be useful for satisfying the conditional exogeneity assumption is the type used in Bronnenberg, Dubè, and Gentzkow (forthcoming). Bronnenberg, Dubè, and Gentzkow find that a household's migration history seems to explain some of the household's brand preferences in the new location. Including this information in Q_{it} would help with identification of structural state dependence by alleviating some endogeneity.

One application where demographics are insufficient conditioning variables

is choice of flavor. Suppose households are choosing between strawberry and peach yogurt. It is unlikely that demographic data could explain enough of each household's A_i that it is plausible that Equation (3.4) holds. Looking at Equation (3.6), it is unlikely that age and income provide enough information about A_i that looking at a subpopulation we would feel comfortable assuming that A_i and Y_{it-1} are independent.

Survey data The flexibility in choosing Q_{it} enables this method to incorporate other types of marketing research that might be informative about unobservables. There are many ways in which marketing researchers have solicited information about preferences from consumers — from studies asking households for their willingness to buy or pay for products, to conjoint studies that evaluate consumers' trade-off between product features. If this type of preference information can be correlated with demographic information that is also available in scanner datasets, it would be possible to use it as a proxy for preferences in Q_{it} to satisfy the conditional exogeneity condition.

3.3.2 Serially correlated errors

The previous sections discussed how researchers can choose Q_{it} to satisfy the conditional exogeneity assumption when errors are *iid*. In some cases, it is possible to incorporate other information into Q_{it} so that some types of serial correlation can be accommodated. Incorporating serially correlated errors in choice models with structural state dependence is difficult. Even without serial correlation, it is difficult to distinguish between unobservable heterogeneity and the effect of a lagged choice variable. Allowing for another avenue for persistence over time complicates the problem substantially — intuitively, how can we distinguish between the effect of the lagged choice variable and the effect of a persistent shock? Typical methods that can identify structural state dependence in panel choice models do not allow for serial correlation (e.g. Honorè and Kyriazidou (2000)), and methods that allow for serial correlation typically do not allow for structural state dependence (e.g. Misra, Roberts, and Handel (2012)).

The method presented in Chapter 1 allows for a lagged choice variable and serially correlated errors concurrently, provided that a suitable Q_{it} can be found. If a shock that occurred in the $t-1$ period is still present in the t period, Equation (3.3) is violated if there is nothing in Q_{it} that provides some information about the serial correlation process. The cost of incorporating both structural state dependence and serial correlation into this model is that one must have some economic interpretation of the preference process, state dependence process and serial correlation process. However, researchers may prefer this to simply ruling out one or two routes of persistence and estimating the remaining process without many assumptions.

Although it is not possible to account for general types of serial correlation, there may be circumstances in which it is possible to account for certain types. For instance, holidays may be considered as conditioning variables since household behavior tends to change at that time, and subsequent weeks may be affected. For instance, households may purchase more candy than usual prior to Halloween, then less candy than usual for several weeks after in a deviation from typical behavior related to inventory and responsiveness to price promotions. Incorporating this information in Q_{it} can alleviate the need to assume that shocks like Halloween last only one week (an assumption required by current methods), or the need to throw out data around holidays. Another possibility for a variable containing information about serial correlation processes in scanner data could be when a household is observed starting to purchase diapers. The arrival of a newborn could have a lingering influence on purchase behavior in other product categories (for instance, organic products or convenience foods).

3.3.3 Implementation of the semiparametric method

A short summary of the semiparametric method presented in Chapter 1 and Chapter 2 follows. Suppose that $X_{it} = [X_{1it}\ X_{2it}]$ are two continuous explanatory variables with corresponding parameters $\beta = [\beta_1\ \beta_2]$.

Identification The key assumption, conditional exogeneity, described above, makes it possible to equate observable conditional probabilities, $Pr(Y_{it} = 1|Y_{it=1} = y, X_{it} = x, Q_{it} = q)$, with an unobservable quantity that is a function of both the distribution of error terms and the distribution of unobservable heterogeneity. Using methods from the transformation model literature and a large support regressor assumption, it can be shown that the ratio of coefficients (β_2/β_1) can be written as the derivative of the the conditional probability, evaluated at some (x, y, q). Then, using the sample of data for which $Y_{it-1} = 1$, it is possible to plot the unobservable quantity discussed above and, under mild regularity conditions on the joint distribution of unobservables, invert it to find an equation for γ as a function of observed conditional probabilities and β from the first step. Discrete X_{it} can be accommodated in the same way, if necessary.

Estimation The proposed estimation strategy follows the identification strategy closely. Conditional probabilities and derivatives of conditional probabilities are estimated with a local linear kernel estimator. Standard quantile estimation methods are used to trace out the unobserved component and estimate γ.

As with all methods employing kernel estimation strategies, results seem to be somewhat sensitive to the choice of bandwidth. The usual "rule of thumb" bandwidth is not a plausible option since the estimation strategy requires at least two continuous variables (the "rule of thumb" bins would be too small for reliable estimation). There is not an obvious "optimal" bandwidth given the two-stage estimation procedure. The difficulty in choosing a bandwidth is a limitation of the proposed method, although techniques like cross-validation have been show to produce reasonable guidance for selecting a bandwidth for similar problems. Another limitation is that the performance of kernel estimators tends to decline as more variables are included in X_{it} and Q_{it}.

3.4 Evidence from simulations

The simulations presented in Section 3.2.2 indicated that standard methods tend to overestimate structural state dependence. This section compares semiparametric estimates that account for endogeneity of lagged choices to standard pooled logit estimates that do not account for endogeneity.

Twenty-five datasets are simulated in accordance with Equation (3.1) for known parameter values. Since the semiparametric method requires at least two continuous variables, the simulation specification is changed slightly from Section 3.2. In addition to a continuous variable representing the difference in ln prices, another variable is added representing the difference in proportions of SKUs that are featured. In the application, this variable will include frequent shopper/members only features, small, medium and large ads, retailer coupons, and rebates.

The matrix of conditioning variables, Q_{it}, is generated from A_i. To get an idea of the semiparametric estimator's performance in actual applications, Q_{it} is generated in such a way that it is not a perfect proxy for A_i. Rather, values of Q_{it} are simulated to fall into 16 discrete values. This mimics the discrete form of demographic variables that is often found in scanner data, for instance in the IRI data used in the application in Section 3.5. In this set-up, Q_{it} provides much information about the size of A_i. However, if A_i is thought to be continuous, a discrete value of Q_{it} will not provide information about the magnitude of A_i within each discrete category. In this case, the conditional exogeneity assumption may not hold strictly. It is worthwhile to evaluate simulations for this type of data generating process since it is uncommon to have continuous data related to unobservable components of the model in real applications.

As described in the previous section, the semiparametric method and the parametric method use different scale normalizations, so only ratios of the coefficient estimates can be compared across models. To evaluate the estimated amount

of structural state dependence, the ratio of γ to β^p, the price parameter, will be compared across methods.

Figure 3.3 shows estimation results for the semiparametric estimator and the pooled logit estimator. The true value of γ/β^p in the underlying data generating process is 0, denoted by the vertical line. For each simulated dataset, coefficients are estimated using both the pooled logit method and the semiparametric method. For both methods, the estimated ratios of coefficients are reported in a histogram.

Figure 3.3a shows the estimated ratios of coefficients when data are simulated to have the same amount of price variation observed in the IRI data used in Section 3.5. The pooled logit estimator consistently overestimates structural state dependence. The semiparametric estimator sometimes overestimates structural state dependence, but to a lesser extent. The better estimate of structural state dependence is due to the incorporation of extra information from the conditioning variables. The semiparametric method estimates the ratio of coefficients well, even though the conditioning variable Q_{it} is not a perfect proxy for A_i.

Figure 3.3b and Figure 3.3c show the performance of the estimators as price variation increases. The data simulated to produce Figure 3.3b have five times more price variation than the IRI data used in Section 3.5. The performance of the pooled logit estimator improves, but it still overestimates structural state dependence. The semiparametric estimator still does well in the presence of additional price variation. Figure 3.3c shows that the pooled logit estimator does not estimate structural state dependence well, even in the presence of an unrealistic amount of price variation.

Although the pooled logit estimates improve as the amount of price variation increases, the semiparametric estimator performs well at all levels of price variation. From Figure 3.3, it is apparent that methods that ignore the endogene-

ity of lagged choices perform well only when there is so much price variation that past choices are uncorrelated with current choices — when brand preferences are insignificant. In realistic situations where past choices are correlated with current choices because brand preferences are important to consumers, the semiparametric method outperforms the parametric method.

3.5 Application to IRI data

As a demonstration, the estimator proposed in Chapter 1 is used to estimate the model described in Equation (3.1) for real brand choice data. Unlike the simulated data in previous sections, here it is not clear that the simple form of Equation (3.1) is sufficient to capture all the heterogeneity in household behavior. There is much evidence in the marketing literature that a richer model is necessary to capture relevant aspects of consumer behavior (for instance, varying price sensitivity across households). With this limitation in mind, estimates from the method described above will be compared with other methods as an approximation for the average of parameters across households.

Data The estimation method is applied to a sample of IRI data from the milk product category. The sample is take from a large store in Pittsfield, MA, between 2003 and 2005. For each household's purchase occasion, the brand choice variable is set to be 1 if a household chooses the national brand and 0 if a household chooses the private label brand. Milk is thought to be a suitable product category for this application since a substantial number of households purchase the private label brand and national brands seem to be fairly stable over the time period considered. The timing of purchases is assumed to be exogenous.

The two continuous explanatory variables considered are price (difference between ln(national brand price) and ln(private label price)) and feature (difference between the proportion of national brand SKUs featured and the proportion of private label SKUs featured). The price trends are graphed in Figure 3.4. Since

few milk products have promotional displays in this store, display information is not incorporated. Persistence in purchase behavior is allowed through a household-specific unobservable fixed effect and through the effect of the lagged brand choice variable.

In order to disentangle the effect of lagged brand choice and the fixed effect, age and income are used as conditioning variables. As described in Section 3.3, this information allows us to assume conditional exogeneity of unobservables and observables. In order to ensure that there are sufficient data for the kernel estimator to perform well, age and income categories with very few observations are combined with adjacent groups in order to construct 16 age-income combinations for use as conditioning variables.

For this application, it is assumed that there is not serial correlation in the error terms.

Results Three models are estimated using the data described above. Two models, the conditional logit and pooled logit, incorporate the lagged choice variable in the same way as the other explanatory variables, despite violating crucial exogeneity assumptions. The third method accommodates the endogeneity of lagged brand choice, as described above. Since the models are identified only to scale, the relative sizes of coefficients are presented.

The results in Figure 3.5 show that each method produces parameter estimates with the expected sign for the marketing mix parameters. The 'feature' variable has a positive effect on purchase probability, and higher price has a negative effect on purchase probability. All three methods estimate a positive amount of structural state dependence.

The relative magnitudes of coefficient estimates differ widely across estimation methods. The pooled logit method estimates β^p to be dramatically larger in

magnitude than β^f. The conditional logit, on the other hand, estimates β^p to be slightly smaller in magnitude than β^f. The semiparametric method estimates β^p to be around the same magnitude as β^f. The conditional logit method estimates that the structural state dependence and price coefficients have the same magnitude. However, the pooled logit and semiparameteric methods estimate that the state dependence magnitude is less than half the magnitude of the price effects. The semiparametric method estimates less state dependence than the other methods.

These results must be interpreted with several caveats. Most importantly, it is likely that there is much more heterogeneity in the data than the estimated model accounts for. Marketing models typically assume that β^p, β^f and γ differ across households. Also, it is possible that the true error term does not have a logistic distribution. The semiparametric method would be robust to this possibility, while the pooled logit and conditional logit methods would provide inconsistent estimates for the parameters if the logit assumption is violated. For this reason, the differences between the relatives magnitudes of coefficients should not be attributed entirely to assumptions related to the lagged brand choice variable. However, in light of the fact that the semiparametric method has more plausible assumptions about the lagged choice variable and is more robust to misspecification of unobservables, it seems that the method should be preferred to parametric methods.

3.6 Discussion

This paper argues that estimation of structural state dependence in binary choice models requires more than allowing for a flexible form of unobservable heterogeneity. Price variation in earlier periods that induces variation in lagged choices cannot disentangle structural state dependence from unobservable heterogeneity completely. Simulations show that one should take caution in interpreting results from previous studies suggesting that both structural state dependence and unobservable heterogeneity contribute to the observed persistence in consumer

choice data.

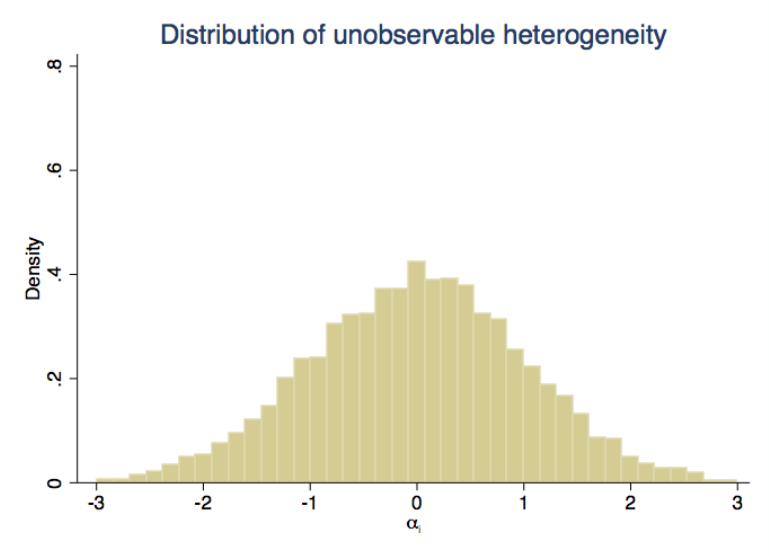
The semiparametric technique for binary choice models presented in this paper allows for both structural state dependence and unobservable heterogeneity. The method requires additional data and economic theory to satisfy the key assumption. For some marketing applications, researchers can incorporate information from other data sources to improve estimates of structural state dependence. Although computationally more difficult than standard techniques, this method can provide a much more accurate picture of structural state dependence than methods that treat the lagged choice variable as exogenous in the presence of unobservable heterogeneity.

The identification and estimation method presented here can be extended to accommodate household-specific price, feature and lagged choice parameters. However, comparing results from the semiparametric method to typical Bayesian procedures is difficult. Although both methods use shrinkage-style estimators, it is hard to compare the results because the semiparametric method discussed here produces point estimates of the parameters while the Bayesian method produces probabilistic distributions. Ongoing work aims to incorporate the structural state dependence identification strategy described here into a Bayesian estimation framework.

	Specification A	Specification B
β	-2	-2
γ	1	0

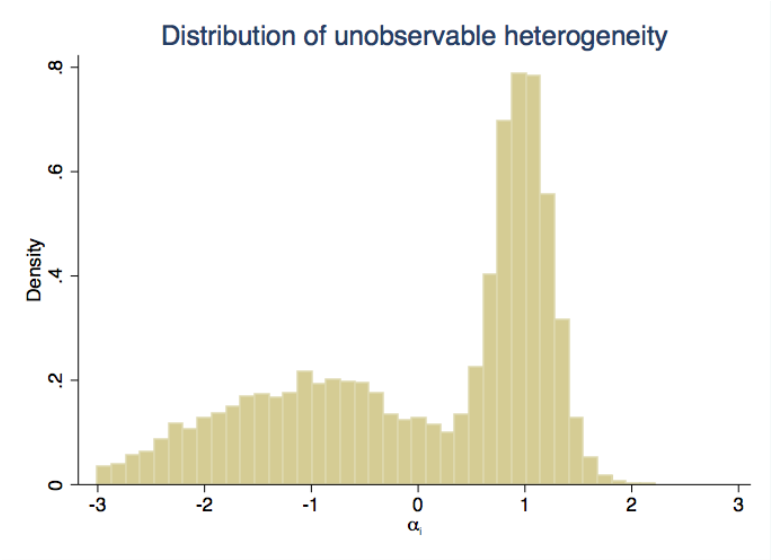
For both specifications, ln(price1) - ln(price2) $\sim$ U[0.18, 0.28]. T=4. Y_{i0} positively correlated with FE. Logit errors with var=1

(a) Simulation specifications

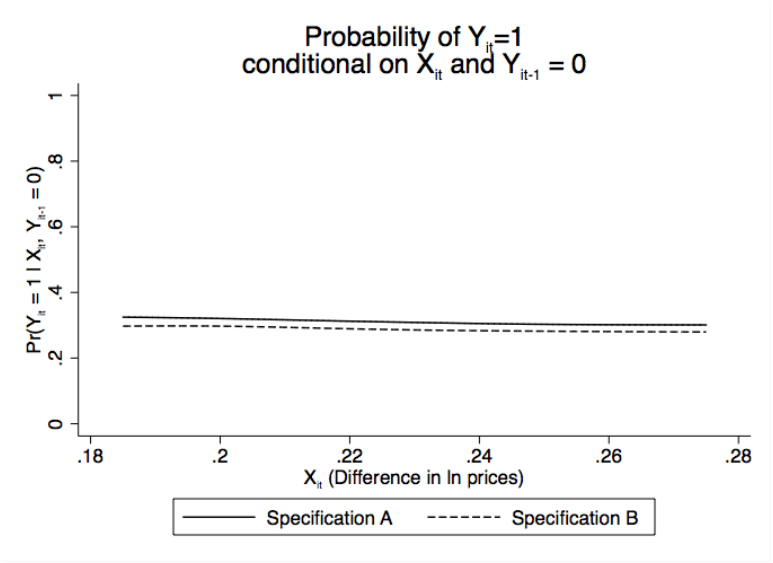


(b) Specification A - A_i

Figure 3.1: Different DGPs generate similar observable joint distributions

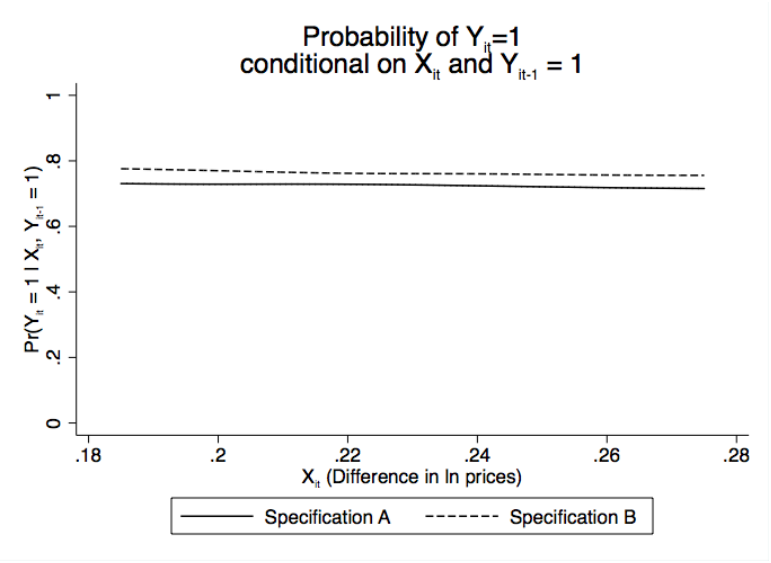


(c) Specification B - A_i



(d) Conditional distribution of Y_{it}

Figure 3.1: Different DGPs generate similar observable joint distributions, contd.



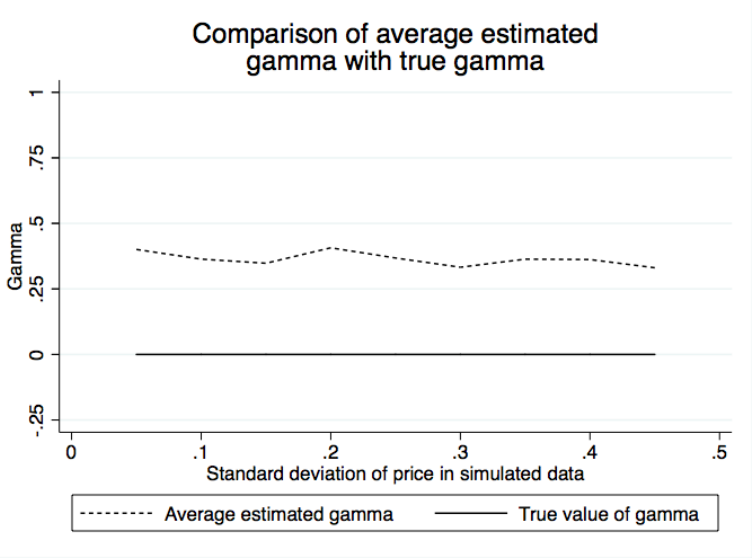
(e) Conditional distribution of Y_{it}

Figure 3.1: Different DGPs generate similar observable joint distributions, contd.

	panel (b)	panel (c)
β	-2	-2
γ	0	0
N	100	100
T	4	4
A_i	$\sim$U[-1,1]	$\sim$U[-2,2]
X_t	$\sim$N(0, Ω)	$\sim$N(0, Ω)
	for $\Omega \in [0.05, .45]$	for $\Omega \in [0.05, .45]$

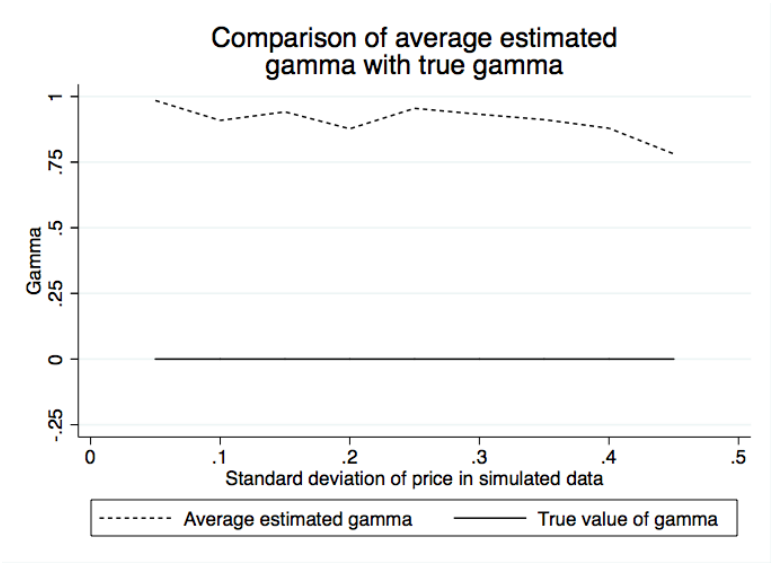
100 draws of data for each Ω, estimated with *xtlogit*

(a) Simulation specifications



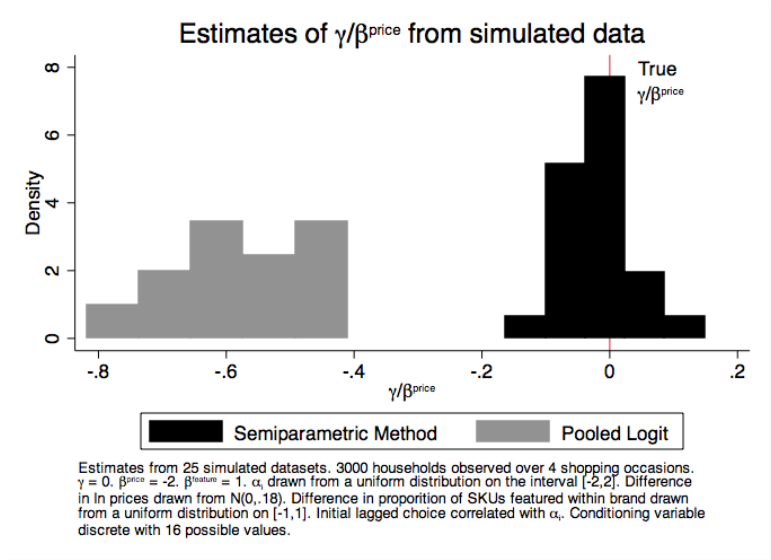
(b) Simulation results — variance of $A_i = 1/3$

Figure 3.2: Price variation when lagged choice is treated as exogenous

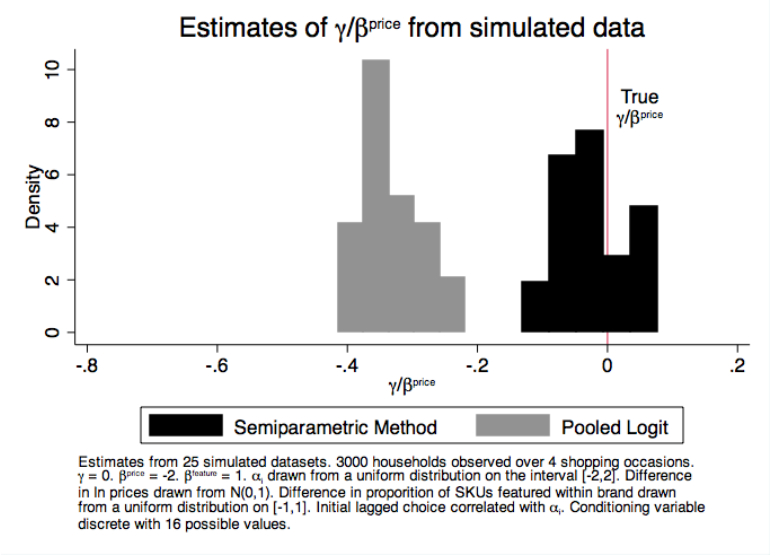


(c) Simulation results — variance of $A_i = 4/3$

Figure 3.2: Price variation when lagged choice is treated as exogenous, contd.

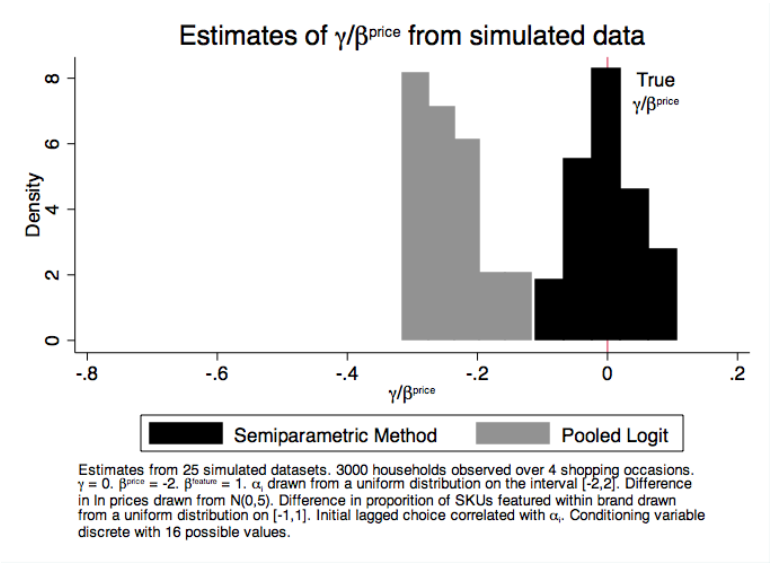


(a) Price variation matches IRI data



(b) Five times more price variation than IRI data

Figure 3.3: Estimated structural state dependence



(c) Twenty five times more price variation than IRI data

Figure 3.3: Estimated structural state dependence, contd.

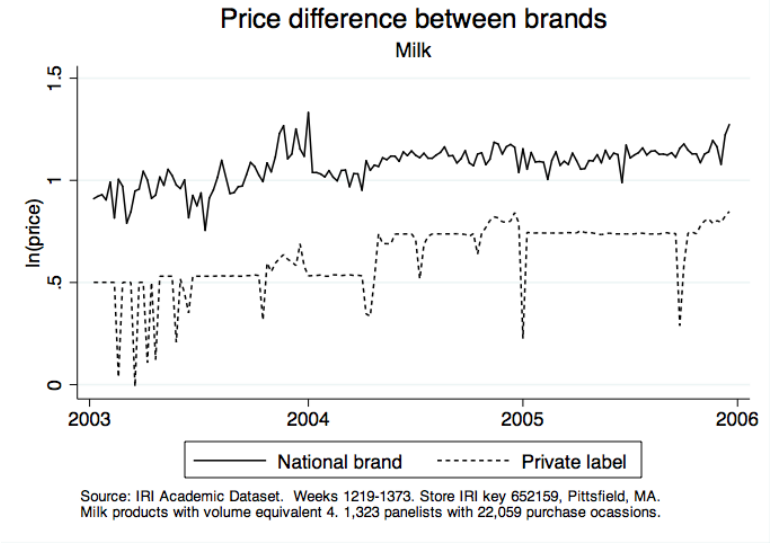


Figure 3.4: Price trends

	Pooled logit	Conditional logit	Semiparametric
β^p/β^f	-17.50	-0.86	-1.06
γ/β^p	-0.40	-1.00	-0.34
γ/β^f	7.04	0.86	0.36

Figure 3.5: Comparison of estimates

Appendix A

Proofs

A.1 Theorem 2

When additional assumptions hold, it is possible to show identification using a strategy that incorporates more data than the method described in Theorem 1. Consider an extension of the conditional exogeneity assumption:

A7 Additional assumptions related to conditioning instruments

(i) Q_{it} includes Q_{it-1}
(ii) $X^c_{it-1}, Y_{it-2}, A_i, U_{it-1} \perp X^c_{it} \mid Q_{it}$
(iii) $(Y_{it-1}, X^b_{it\ell}) \perp X^c_{it} \mid Q_{it}$

Theorem 2 Assume **A1 - A6**. β and γ can be identified using more information than Theorem 1 when

(i) in addition, **A7 (i)** and **A7 (ii)** holds. It is possible to incorporate information from both $Y_{it-1} = 1$ and $Y_{it-1} = 0$ into identification of β and γ.

(ii) in addition, **A7 (iii)** holds. It is possible to incorporate information from both $X^b_{it\ell} = 1$ and $X^b_{it\ell} = 0$ into identification of β and γ.

The proof of Theorem 2 closely follows the proof of Theorem 1, with the exception of identifying β^c. Section A.1.1 and Section A.1.2 illustrate adjustments

that must be made to the identification strategy proposed above in order to incorporate more information. Section A.1.3 discusses identification of $\tilde{G}$, γ and β^b for a general case where **A7** holds for all $X^b_{it\ell}$ in X^b_{it}.

A.1.1 Proof of Theorem 2(i)

Identification of β^c First consider the case in which the model contains only continuous variables and the lagged dependent variable (no other binary variables). It is possible to extend the identification strategy in order to average across both possible values of Y_{it-1}, incorporating more data than what is used when β^c is identified for any particular y when **A7(i)** holds. Note that

$$\begin{aligned}
&Pr(Y_{it} = 1 \mid X^c_{it} = x^c, Q_{it} = q) \\
&\quad = Pr(Y_{it} = 1 \mid X^c_{it} = x^c, Y_{it-1} = 1, Q_{it} = q) Pr(Y_{it-1} = 1 \mid X^c_{it} = x^c, Q_{it} = q) \\
&\qquad + Pr(Y_{it} = 1 \mid X^c_{it} = x^c, Y_{it-1} = 0, Q_{it} = q) Pr(Y_{it-1} = 0 \mid X^c_{it} = x^c, Q_{it} = q) \\
&\quad = \tilde{G}(x^c\beta^c + \gamma, q) Pr(Y_{it-1} = 1 \mid X^c_{it} = x^c, Q_{it} = q) \\
&\qquad + \tilde{G}(x^c\beta^c, q) Pr(Y_{it-1} = 0 \mid X^c_{it} = x^c, Q_{it} = q)
\end{aligned}$$

Taking the derivative as before,

$$\begin{aligned}
\frac{\partial}{\partial x^c_\ell} Pr(Y_{it} = 1 \mid X^c_{it} = x^c, Q_{it} = q) &= \frac{\partial}{\partial x^c_\ell} \left[\tilde{G}(x^c\beta^c + \gamma, q) Pr(Y_{it-1} = 1 \mid X^c_{it} = x^c, Q_{it} = q) \right] \\
&\quad + \frac{\partial}{\partial x^c_\ell} \left[\tilde{G}(x^c\beta^c, q) Pr(Y_{it-1} = 0 \mid X^c_{it} = x^c, Q_{it} = q) \right] \\
&= \tilde{G}(x^c\beta^c + \gamma, q) \frac{\partial}{\partial x^c_\ell} Pr(Y_{it-1} = 1 \mid X^c_{it} = x^c, Q_{it} = q) \\
&\quad + \frac{\partial}{\partial x^c_\ell} \tilde{G}(x^c\beta^c + \gamma, q) Pr(Y_{it-1} = 1 \mid X^c_{it} = x^c, Q_{it} = q) \\
&\quad + \tilde{G}(x^c\beta^c, q) \frac{\partial}{\partial x^c_\ell} Pr(Y_{it-1} = 0 \mid X^c_{it} = x^c, Q_{it} = q) \\
&\quad + \frac{\partial}{\partial x^c_\ell} \tilde{G}(x^c\beta^c, q) Pr(Y_{it-1} = 0 \mid X^c_{it} = x^c, Q_{it} = q)
\end{aligned}$$

When the derivatives of $Pr(Y_{it-1} = 1 \mid X^c_{it} = x^c, Q_{it} = q)$ and $Pr(Y_{it-1} = 0 \mid X^c_{it} = x^c, Q_{it} = q)$ do not depend on x^c, it will be possible to isolate β^c as in the proof of Theorem 1. Generally it is not the case that $Pr(Y_{it-1} = 1 \mid X^c_{it} = x^c, Q_{it} = q) = Pr(Y_{it-1} = 1 \mid Q_{it} = q)$. To proceed, it is necessary to show

$$Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) = Pr(Y_{it-1} = 1 \mid Q_{it} = q) \tag{A.1}$$

and

$$Pr(Y_{it-1} = 0 \mid X_{it}^c = x^c, Q_{it} = q) = Pr(Y_{it-1} = 0 \mid Q_{it} = q) \tag{A.2}$$

Note that

$$\begin{aligned} Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) &= Pr(X_{it-1}^c\beta^c + \gamma Y_{it-2} + A_i + U_{it-1} \geq 0 \mid X_{it}^c = x^c, Q_{it} = q) \\ &= Pr(X_{it-1}^c\beta^c + \gamma Y_{it-2} + A_i \geq -U_{it-1} \mid X_{it}^c = x^c, Q_{it} = q) \end{aligned}$$

In order to rewrite this in terms of $G(\cdot, q)$ as in Equation (1.4), it must be the case that **A4** holds for the $t-1$ period. This means that $(U_{it-1}, A_i) \perp X_{it-1}^c, Y_{it-2} \mid Q_{it-1}$ must hold. This holds when **A7(i)** holds so that Q_{it} includes Q_{it-1}. Additional conditioning variables, X_{it}, do not prevent **A7(i)** from inducing **A4** to hold in this situation.

In some cases, for instance when Q_{it} and Q_{it-1} contain only time-invariant variables, this it trivial and the cost of incorporating more information into the identification strategy is negligible. However, in other cases where more complicated, time-variant Q_{it} is required in order to satisfy **A4**, this is more restrictive.

So, by **A7(i)** and **A4**,

$$Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) = G(X_{it-1}^c\beta^c + \gamma Y_{it-2} + A_i \mid X_{it}^c = x^c, Q_{it} = q) \tag{A.3}$$

Assuming it is also the case that **A7(ii)** holds,

$$\begin{aligned} Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) &= G(X_{it-1}^c\beta^c + \gamma Y_{it-2} + A_i \mid X_{it}^c = x^c, Q_{it} = q) \\ &= G(X_{it-1}^c\beta^c + \gamma Y_{it-2} + A_i \mid Q_{it} = q) \\ &= Pr(Y_{it-1} = 1 \mid Q_{it} = q) \end{aligned} \tag{A.4}$$

Although possibly restrictive, **A7(ii)** allows Equation (A.1) and Equation (A.2) to hold. Then,

$$\frac{\partial}{\partial x_\ell^c} Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) = \frac{\partial}{\partial x_\ell^c} Pr(Y_{it-1} = 1 \mid Q_{it} = q) = 0$$

and

$$\frac{\partial}{\partial x_\ell^c} Pr(Y_{it-1} = 0 \mid X_{it}^c = x^c, Q_{it} = q) = \frac{\partial}{\partial x_\ell^c} Pr(Y_{it-1} = 0 \mid Q_{it} = q) = 0$$

These assumptions simplify the derivative of the conditional probability in such a way that β_ℓ^c can be isolated.

$$\begin{aligned}
\frac{\partial}{\partial x_\ell^c} Pr(Y_{it} = 1 \mid X_{it}^c = x^c, Q_{it} = q) &= \frac{\partial}{\partial x_\ell^c} \tilde{G}(x^c\beta^c + \gamma, q) Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) \\
&\quad + \frac{\partial}{\partial x_\ell^c} \tilde{G}(x^c\beta^c, q) Pr(Y_{it-1} = 0 \mid X_{it}^c = x^c, Q_{it} = q) \\
&= \beta_\ell^c \tilde{G}'(x^c\beta^c + \gamma, q) Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) \\
&\quad + \beta_\ell^c \tilde{G}'(x^c\beta^c, q) Pr(Y_{it-1} = 0 \mid X_{it}^c = x^c, Q_{it} = q) \\
&= \beta_\ell^c \big[\tilde{G}'(x^c\beta^c + \gamma, q) Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) \\
&\quad + \tilde{G}'(x^c\beta^c, q) Pr(Y_{it-1} = 0 \mid X_{it}^c = x^c, Q_{it} = q)\big]
\end{aligned}$$

Then, as before, it is possible to identify β_ℓ using the normalization $\beta_1 = 1$.

$$\frac{\frac{\partial}{\partial x_\ell^c} Pr(Y_{it} = 1 \mid X_{it}^c = x^c, Q_{it} = q)}{\frac{\partial}{\partial x_1^c} Pr(Y_{it} = 1 \mid X_{it}^c = x^c, Q_{it} = q)} =$$

$$= \frac{\beta_\ell^c \big[\tilde{G}'(x^c\beta^c + \gamma, q) Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) + \tilde{G}'(x^c\beta^c, q) Pr(Y_{it-1} = 0 \mid X_{it}^c = x^c, Q_{it} = q)\big]}{\beta_1^c \big[\tilde{G}'(x^c\beta^c + \gamma, q) Pr(Y_{it-1} = 1 \mid X_{it}^c = x^c, Q_{it} = q) + \tilde{G}'(x^c\beta^c, q) Pr(Y_{it-1} = 0 \mid X_{it}^c = x^c, Q_{it} = q)\big]}$$

$$= \frac{\beta_\ell^c}{\beta_1^c} = \beta_\ell^c \tag{A.5}$$

Note that it is not possible to average over Q_{it} without making further assumptions. Although it would be desirable to alleviate the need to chose a specific $Q_{it} = q$ to identify β, it is not possible without specifying a joint distribution for

X_{it}^c and Q_{it}. In applications where a specification of this sort is reasonable (for instance, in situations in which control function approaches involving parametric specification of the endogenous variables are considered appropriate), it would be straightforward to extend this method to use $Pr(Y_{it} = 1 \mid X_{it}^c = x^c)$ to identify β, rather than $Pr(Y_{it} = 1 \mid X_{it}^c = x^c, Q_{it} = q)$.

A.1.2 Proof of Theorem 2 (ii)

Identification of β^c When the model includes other binary variables, it may be possible to average over other binary variables, in addition to the lagged dependent variable. This averaging is desirable since it eliminates the need for the researcher to specify whether β^c is identified from $X_{it\ell}^b = 0$ or $X_{it\ell}^b = 1$, increasing the information used in identification.

This section will outline the strategy when there are two binary variables - the lagged dependent variable and one other binary variable. It is assumed that **A7(i)** and **A7(ii)** hold so it is possible to average across the lagged dependent variable, and also **A7(iii)** holds for so it is possible to average across the other binary variable. However, if **A7(i)** and **A7(ii)** did not hold, it would be possible to average across just the other binary variable as long as **A7(iii)** holds.

The equation for $Pr(Y_{it} = 1|X_{it}^c = x^c, Q_{it} = q)$ becomes:

$$\begin{aligned} Pr(Y_{it} = 1|X_{it}^c = x^c, Q_{it} = q) &= \sum_{j=0}^{1}\sum_{k=0}^{1} \tilde{G}(x^c\beta^c + \beta^b j + \gamma k) Pr(X_{it}^b = j \cap Y_{it-1} = k|X_{it}^c = x^c, Q_{it} = q) \\ &= \sum_{j=0}^{1}\sum_{k=0}^{1} \tilde{G}(x^c\beta^c + \beta^b j + \gamma k) Pr(X_{it}^b = j \cap Y_{it-1} = k|Q_{it} = q) \end{aligned}$$

The second equality is a consequence of **A7(iii)**. Taking the derivative with respect to x^c as before,

$$\begin{aligned} &\frac{\partial}{\partial x_\ell^c} Pr(Y_{it} = 1 \mid X_{it}^c = x^c, Q_{it} = q) \\ &\qquad = \beta_\ell^C \left[\sum_{j=0}^{1}\sum_{k=0}^{1} \tilde{G}'(x^c\beta^c + \beta^b j + \gamma k) Pr(X_{it}^b = j \cap Y_{it-1} = k \mid Q_{it} = q) \right] \end{aligned}$$

From here, as above, it is possible to identify β_ℓ^c using the ratio of derivatives using the normalization $\beta_1 = 1$.

$$\begin{aligned}
&\frac{\frac{\partial}{\partial x_\ell^c} Pr(Y_{it} = 1 \mid X_{it}^c = x^c, Q_{it} = q)}{\frac{\partial}{\partial x_1^c} Pr(Y_{it} = 1 \mid X_{it}^c = x^c, Q_{it} = q)} \\
&= \frac{\beta_\ell^c \left[\sum_{j=0}^{1}\sum_{k=0}^{1} \tilde{G}'(x^c\beta^c + \beta^b j + \gamma k) Pr(X_{it}^b = j \cap Y_{it-1} = k \mid Q_{it} = q)\right]}{\beta_1^c \left[\sum_{j=0}^{1}\sum_{k=0}^{1} \tilde{G}'(x^c\beta^c + \beta^b j + \gamma k) Pr(X_{it}^b = j \cap Y_{it-1} = k \mid Q_{it} = q)\right]} \\
&= \frac{\beta_\ell^c}{\beta_1^c} = \beta_\ell^c
\end{aligned} \tag{A.6}$$

Extensions of this strategy for more than one binary explanatory variable is straightforward.

A.1.3 Identification of $\tilde{G}$, β^b and γ

Identification results for β^b and γ follow immediately from Section 1.2.2.

A.2 Proof of Theorem 3

Denote $Z_{it} = (X_{it}, Y_{it-1}, Q_{it})$ and $z = (x, y, q)$. Let $\mathcal{Z}$ represent the support of Z_{it}.

Recall that $b_s(z)$ denotes the derivative of the conditional probability $Pr(Y_{it} = 1 | X_{it} = x, Y_{it-1} = y, Q_{it} = q)$ with respect to the s^{th} element of x^c, x_s^c.

Object of interest:

$$\begin{aligned}
\beta_s^c &= E\left[w(z)\beta_s^c(z)\right] \\
&= E\left[w(z)\frac{b_s(z)}{b_1(z)}\right]
\end{aligned}$$

Denote estimator:

$$\begin{aligned}\hat{\beta}_s^c &= \int_{\mathcal{Z}} w(z)\hat{\beta}_s^c(z)dz \\ &= \int_{\mathcal{Z}} w(z)\frac{\hat{b}_s(z)}{\hat{b}_1(z)}dz\end{aligned}$$

where $\hat{\beta}_s(z)$ and $\hat{\beta}_1(z)$ defined in Equation...

Approach

- Use a Taylor series expansion to approximate $\frac{\hat{\beta}_s(z)}{\hat{\beta}_1(z)}$ by a linear functional of kernel estimators.
- Show that the error made by the linear approximation is $o_P(n^{-1/2})$
- Use the uniform law of large numbers to show that the quantity observed by replacing $\frac{\hat{\beta}_s(z)}{\hat{\beta}_1(z)}$ with the linear approximation is asymptotically equivalent to an empirical process after centering and normalization (i.e. it is asymptotically equivalent to a sum of the form $n^{-1/2}\sum_{i=1}^{N}[\Phi(Z_{it} - E\Phi(Z)]$ for a suitable function Φ)
- Show that the empirical process converges in distribution

Expand $\hat{\beta}$ around β.

$$\begin{aligned}\hat{\beta}_s^c - \beta_s^c &= \int_{\mathcal{Z}} w(z)\left[\frac{\hat{b}_s(z)}{\hat{b}_1(z)} - \frac{b_s(z)}{b_1(z)}\right]dz \\ &= \int_{\mathcal{Z}} w(z)\frac{1}{b_1(z)}\left(\hat{b}_s(z) - b_s(z)\right)dz - \int_{\mathcal{Z}} w(z)\frac{b_s(z)}{b_1(z)^2}\left(\hat{b}_1(z) - b_1(z)\right)dz \\ &\quad + O(||\hat{b}_s(z) - b_s(z)||_\infty^2) + O(||\hat{b}_1(z) - b_1(z)||_\infty^2)\end{aligned}$$

where $||\cdot||_\infty$ denotes the supremum norm over $\mathcal{Z}$.

As shown in Lemma 1 and Lemma 2, the first two terms can be rewritten as a U-statistic.

$$\int_{\mathcal{Z}} w(z)\frac{1}{b_1(z)}\left(\hat{b}_s(z) - b_s(z)\right)dz - \int_{\mathcal{Z}} w(z)\frac{b_s(z)}{b_1(z)^2}\left(\hat{b}_1(z) - b_1(z)\right)dz$$
$$= U_1(Z) + U_2(Z) = U(Z)$$

Lemma 3 shows that $O(||\hat{b}_s(z) - b_s(z)||_\infty^2) = o_P(1/\sqrt{n})$ and $O(||\hat{b}_1(z) - b_1(z)||_\infty^2) = o_P(1/\sqrt{n})$.

Then,

$$\sqrt{N}(\hat{\beta} - \beta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T} U(Z_{it}) + o_P(1) \tag{A.7}$$

Pointwise weak convergence follows from the CLT for iid sequences.

Lemma 1

$$\frac{1}{b_1(z)}\left(\hat{b}_s(z) - b_s(z)\right)dz - \frac{b_s(z)}{b_1(z)^2}\left(\hat{b}_1(z) - b_1(z)\right)dz$$
$$= \nabla_f\beta(z)\{\hat{f}(z) - f(z)\} + \nabla_g\beta(z)\{\hat{g}(z) - g(z)\} + R$$

Proof:

Recall that $b_s(z)$ is the local linear estimator of the derivative of the conditional probability with respect to x_s^c. This proof will use the local constant estimator to derive the asymptotic distribution and variance.

- The formula for the $s+1^{th}$ element of the local linear estimator is intractable.
- Although Hansen (2008) presents convergence results for the local constant estimator and the local linear estimator, the local linear estimator results pertain only to $a(z)$, not $b(z)$. It is not clear how to derive results for $b(z)$ due to the first point.

Asymptotically, the local constant and local linear estimators have the same distribution and variance. This approach requires some additional notation.

Suppose $g(z)$ denotes the joint distribution of Y_{it}, X_{it}, Y_{it-1} and Q_{it}. Let $f(z)$ denote the marginal distribution of X_{it}, Y_{it-1} and Q_{it}. Then,

$$Pr(Y_{it} = 1|X_{it} = x, Y_{it} = y, Q_{it} = q) = \frac{g(z)}{f(z)}$$

Then, the analytical derivative of the conditional probability with respect to x_s^c

$$b_s(z) = \frac{\partial}{\partial x_s^c} Pr(Y_{it} = 1|X_{it} = x, Y_{it} = y, Q_{it} = q) = \frac{1}{f(z)} \left[\frac{\partial g(z)}{\partial x_s^c} - \frac{g(z)}{f(z)} \frac{\partial f(z)}{\partial x_s^c} \right] \tag{A.8}$$

The kernel estimator for $g(z)$ is

$$\hat{g}(z) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} Y_{it} K(Z_{it}, z) \tag{A.9}$$

with derivative

$$\hat{g}_s(z) = \frac{\partial \hat{g}(z)}{\partial x_s^c} = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} Y_{it} \frac{\partial K(Z_{it}, z)}{\partial x_s^c} \tag{A.10}$$

The kernel estimator for $f(z)$ is

$$\hat{f}(z) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} K(Z_{it}, z) \tag{A.11}$$

with derivative

$$\hat{f}_s(z) = \frac{\partial \hat{f}(x, y, q)}{\partial x_s^c} = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{\partial K(Z_{it}, z)}{\partial x_s^c} \tag{A.12}$$

The local constant estimator for $b_s(z)$ is defined by substituting Equation

(A.9), Equation (A.10), Equation (A.11) and Equation (A.12) into Equation (A.8).

Expand $\hat{b}_s(z)$ around $b_s(z)$:

$$\begin{aligned}\hat{b}_s(z) - b_s(z) &= \frac{1}{\hat{f}(z)}\left[\hat{g}_s(z) - \frac{\hat{g}(z)}{\hat{f}(z)}\hat{f}_s(z)\right] - \frac{1}{f(z)}\left[g_s(z) - \frac{g(z)}{f(z)}f_s(z)\right] \\ &= \left[\frac{2g(z)f_s(z)}{f(z)^3} - \frac{g_s(z)}{f(z)^2}\right]\{\hat{f}(z) - f(z)\} + \left[-\frac{f_s(z)}{f(z)^2}\right]\{\hat{g}(z) - g(z)\} \\ &\quad + \left[-\frac{g(z)}{f(z)^2}\right]\{\hat{f}_s(z) - f_s(z)\} + \left[\frac{1}{f(z)}\right]\{\hat{g}_s(z) - g_s(z)\} \\ &\quad + O(|\hat{f}(z) - f(z)|^2) + O(|\hat{g}(z) - g(z)|^2) + O(|\hat{f}_s(z) - f_s(z)|^2) + O(|\hat{g}_s(z) - g_s(z)|^2)\end{aligned}$$

This holds for all $s = 1, ..., q^{x^c}$.

From the previous expansion, note that

$$\begin{aligned}&\frac{1}{b_1(z)}\left[\hat{b}_s(z) - b_s(z)\right] \\ &\quad = \frac{1}{b_1(z)}\left[\frac{2g(z)f_s(z)}{f(z)^3} - \frac{g_s(z)}{f(z)^2}\right]\{\hat{f}(z) - f(z)\} + \frac{1}{b_1(z)}\left[-\frac{f_s(z)}{f(z)^2}\right]\{\hat{g}(z) - g(z)\} \\ &\qquad + \frac{1}{b_1(z)}\left[-\frac{g(z)}{f(z)^2}\right]\{\hat{f}_s(z) - f_s(z)\} + \frac{1}{b_1(z)}\left[\frac{1}{f(z)}\right]\{\hat{g}_s(z) - g_s(z)\} \\ &\qquad + O(|\hat{f}(z) - f(z)|^2) + O(|\hat{g}(z) - g(z)|^2) + O(|\hat{f}_s(z) - f_s(z)|^2) + O(|\hat{g}_s(z) - g_s(z)|^2)\end{aligned}$$

and

$$\begin{aligned}\frac{b_s(z)}{b_1(z)^2}\left[\hat{b}_1(z) - b_1(z)\right] &= \frac{b_s(z)}{b_1(z)^2}\left[\frac{2g(z)f_1(z)}{f(z)^3} - \frac{g_1(z)}{f(z)^2}\right]\{\hat{f}(z) - f(z)\} + \frac{b_s(z)}{b_1(z)^2}\left[-\frac{f_1(z)}{f(z)^2}\right]\{\hat{g}(z) - g(z)\} \\ &\quad + \frac{b_s(z)}{b_1(z)^2}\left[-\frac{g(z)}{f(z)^2}\right]\{\hat{f}_1(z) - f_1(z)\} + \frac{b_s(z)}{b_1(z)^2}\left[\frac{1}{f(z)}\right]\{\hat{g}_1(z) - g_1(z)\} \\ &\quad + O(|\hat{f}(z) - f(z)|^2) + O(|\hat{g}(z) - g(z)|^2) + O(|\hat{f}_1(z) - f_1(z)|^2) + O(|\hat{g}_1(z) - g_1(z)|^2)\end{aligned}$$

Combining :

$$\begin{aligned}&\frac{1}{b_1(z)}\left[\hat{b}_s(z) - b_s(z)\right] - \frac{b_s(z)}{b_1(z)^2}\left[\hat{b}_1(z) - b_1(z)\right] = \\ &\quad \frac{1}{b_1(z)}\left[\frac{2g(z)f_s(z)}{f(z)^3} - \frac{g_s(z)}{f(z)^2}\right]\{\hat{f}(z) - f(z)\} + \frac{1}{b_1(z)}\left[-\frac{f_s(z)}{f(z)^2}\right]\{\hat{g}(z) - g(z)\} \\ &\qquad + \frac{1}{b_1(z)}\left[-\frac{g(z)}{f(z)^2}\right]\{\hat{f}_s(z) - f_s(z)\} + \frac{1}{b_1(z)}\left[\frac{1}{f(z)}\right]\{\hat{g}_s(z) - g_s(z)\} \\ &\qquad - \frac{b_s(z)}{b_1(z)^2}\left[\frac{2g(z)f_1(z)}{f(z)^3} - \frac{g_1(z)}{f(z)^2}\right]\{\hat{f}(z) - f(z)\} - \frac{b_s(z)}{b_1(z)^2}\left[-\frac{f_1(z)}{f(z)^2}\right]\{\hat{g}(z) - g(z)\} \\ &\qquad - \frac{b_s(z)}{b_1(z)^2}\left[-\frac{g(z)}{f(z)^2}\right]\{\hat{f}_1(z) - f_1(z)\} - \frac{b_s(z)}{b_1(z)^2}\left[\frac{1}{f(z)}\right]\{\hat{g}_1(z) - g_1(z)\} \\ &\qquad + O(|\hat{f}(z) - f(z)|^2) + O(|\hat{g}(z) - g(z)|^2) + O(|\hat{f}_1(z) - f_1(z)|^2) + O(|\hat{g}_1(z) - g_1(z)|^2)\end{aligned}$$

Collecting terms :

$$\begin{aligned}
&\frac{1}{b_1(z)}\left[\hat{b}_s(z)-b_s(z)\right]-\frac{b_s(z)}{b_1(z)^2}\left[\hat{b}_1(z)-b_1(z)\right]= \\
&\quad\left\{\frac{1}{b_1(z)}\left[\frac{2g(z)f_s(z)}{f(z)^3}-\frac{g_s(z)}{f(z)^2}\right]-\frac{b_s(z)}{b_1(z)^2}\left[\frac{2g(z)f_1(z)}{f(z)^3}-\frac{g_1(z)}{f(z)^2}\right]\right\}\{\hat{f}(z)-f(z)\} \\
&\qquad+\left\{\frac{1}{b_1(z)}\left[-\frac{f_s(z)}{f(z)^2}\right]-\frac{b_s(z)}{b_1(z)^2}\left[-\frac{f_1(z)}{f(z)^2}\right]\right\}\{\hat{g}(z)-g(z)\} \\
&\qquad+\frac{1}{b_1(z)}\left[-\frac{g(z)}{f(z)^2}\right]\{\hat{f}_s(z)-f_s(z)\}+\frac{1}{b_1(z)}\left[\frac{1}{f(z)}\right]\{\hat{g}_s(z)-g_s(z)\} \\
&\qquad-\frac{b_s(z)}{b_1(z)^2}\left[-\frac{g(z)}{f(z)^2}\right]\{\hat{f}_1(z)-f_1(z)\}-\frac{b_s(z)}{b_1(z)^2}\left[\frac{1}{f(z)}\right]\{\hat{g}_1(z)-g_1(z)\} \\
&\qquad+O(|\hat{f}(z)-f(z)|^2)+O(|\hat{g}(z)-g(z)|^2)+O(|\hat{f}_1(z)-f_1(z)|^2)+O(|\hat{g}_1(z)-g_1(z)|^2) \\
&\qquad+O(|\hat{f}_s(z)-f_s(z)|^2)+O(|\hat{g}_s(z)-g_s(z)|^2)
\end{aligned}$$

Define functionals for some function $m(z)$:

$$\begin{aligned}
\nabla_f\beta(z)\{m(z)\} &= \left(\frac{1}{\alpha_1(z)}\left[\frac{2g(z)f_s(z)}{f(z)^3}-\frac{g_s(z)}{f(z)^2}\right]-\frac{\alpha_s(z)}{\alpha_1(z)^2}\left[\frac{2g(z)f_1(z)}{f(z)^3}-\frac{g_1(z)}{f(z)^2}\right]\right)\{m(z)\} \\
&\quad-\frac{g(z)}{\alpha_1(z)f(z)^2}\left\{\frac{\partial}{\partial x_s}m(z)\right\}+\frac{\alpha_s(z)g(z)}{\alpha_1(z)^2f(z)^2}\left\{\frac{\partial}{\partial x_1}m(z)\right\} \\
&=: \zeta(z)m(z)+\zeta_s(z)m_s(z)+\zeta_1(z)m_1(z)
\end{aligned}$$

$$\begin{aligned}
\nabla_g\beta(z)\{m(z)\} &= \left[\frac{\alpha_s(z)f_1(z)}{\alpha_1(z)^2f(z)^2}-\frac{f_s(z)}{\alpha_1(z)f(z)^2}\right]\{m(z)\}+\frac{1}{\alpha_1(z)f(z)}\left\{\frac{\partial}{\partial x_s}m(z)\right\} \\
&\quad-\frac{\alpha_s(z)}{\alpha_1(z)^2f(z)}\left\{\frac{\partial}{\partial x_1}m(z)\right\} \\
&=: \xi m(z)+\xi_s(z)m_s(z)+\xi_1(z)m_1(z)
\end{aligned}$$

Define the remainder:

$$\begin{aligned}
R = &O(|\hat{f}(z)-f(z)|^2)+O(|\hat{g}(z)-g(z)|^2)+O(|\hat{f}_1(z)-f_1(z)|^2)+O(|\hat{g}_1(z)-g_1(z)|^2) \\
&+O(|\hat{f}_s(z)-f_s(z)|^2)+O(|\hat{g}_s(z)-g_s(z)|^2)
\end{aligned}$$

Then

$$= \nabla_f\beta(z)\{\hat{f}(z)-f(z)\}+\nabla_g\beta(z)\{\hat{g}(z)-g(z)\}+R$$

Assumptions imply that $R = o_P(1/\sqrt{n})$.

Assumption (Weakly dependent data)

$\{Y_{it}, X_{it}, Y_{it-1}, Q_{it}\}$ is strictly stationary and strong mixing with mixing coefficients α_m that satisfy

$$\alpha_m \leq Am^{-\delta} \tag{A.13}$$

where $A < \infty$ and $\delta > 2$. Also, the marginal density of Z_{it}, f(z), must satisfy

$$\sup_z f(z) \leq B_0 < \infty \tag{A.14}$$

The joint density of $z_{i_1t_1}$ and $z_{i_2t_2}$ is also bounded.

Assumption (Differentiable kernel) Univariate kernel k is differentiable. There are constants $C, \eta > 0$ such that, for the i^{th} derivative $k^i(z)$

- $|k(z)| \leq C|z|^{-\eta}$
- $|k'(z)| \leq C|z|^{-\eta}$ (bounded derivative)
- $|k(z) - k(z')| \leq C|z - z'|$
- $|k'(z) - k'(z')| \leq C|z - z'|$
- $\int_{-\infty}^{\infty} k(z)dz = 1$
- $\int_{-\infty}^{\infty} z^j k(z)dz = 0$ when$1 \leq j \leq m - 1$ (higher order kernel)
- $\int_{-\infty}^{\infty} |z|^m k(z)dz < \infty$

Assumption (Sufficiently smooth data) The joint density $g(z)$ is

- bounded
- m times differentiable with respect to each component
- with bounded derivatives

- the m^{th} order partial derivatives are uniformly continuous
- $\sup_{z\in\mathcal{Z}}||z||^b g(z) < \infty$ for some constant $b > 0$ Sufficient conditions:
  - $E[||X||^b] < \infty$
  - $E[||Q||^b] < \infty$
  - No condition necessary on Y since it is binary.

These two assumptions allow us to apply standard results from the literature on iid kernel density smoothers. The following results are taken from Hansen (2008), Theorem 6.

$$||\hat{f}(z) - f(z)||_\infty = O_P(h^m) + O_P\left(\sqrt{\frac{\log n}{nh^{d^{x^c}+d^{q^c}}}}\right)$$
$$||\hat{g}(z) - g(z)||_\infty = O_P(h^m) + O_P\left(\sqrt{\frac{\log n}{nh^{d^{x^c}+d^{q^c}}}}\right)$$
$$||\hat{f}_1(z) - f_1(z)||_\infty = O_P(h^m) + O_P\left(\sqrt{\frac{\log n}{nh^{d^{x^c}+d^{q^c}+2}}}\right)$$
$$||\hat{g}_1(z) - g(z)||_\infty = O_P(h^m) + O_P\left(\sqrt{\frac{\log n}{nh^{d^{x^c}+d^{q^c}+2}}}\right)$$
$$||\hat{f}_s(z) - f_s(z)||_\infty = O_P(h^m) + O_P\left(\sqrt{\frac{\log n}{nh^{d^{x^c}+d^{q^c}+2}}}\right)$$
$$||\hat{g}_s(z) - g_s(z)||_\infty = O_P(h^m) + O_P\left(\sqrt{\frac{\log n}{nh^{d^{x^c}+d^{q^c}+2}}}\right)$$

Note: this is written for common bandwidths. In the more realistic case where different bandwidths are selected for each variable, the first component would look like $O_P(\max(h_{x_1^c}, ..., h_{x_d^c}, h_{q_1^c}, ..., h_{q_d^c})^m)$ instead of $O_P(h^m)$.

Assumption (bandwidth/n rate) Suppose for ease of exposition that the same bandwidth is appropriate for all variables: $h_{x_1^c} = ... = h_{q_1^c} = ... = h$. Data-

driven bandwidth selection methods are discussed in Section 2.3.1. Bandwidth h refers to a sequence of bandwidth h_n that shrinks to zero as $n \to \infty$.

- $\sqrt{nh^2} \to 0$
- $\sqrt{nh^{d+2}}/log(n) \to 0$

  This implies that $R = o_P(1/\sqrt{n})$, and therefore

$$= \nabla_f \beta(z)\{\hat{f}(z) - f(z)\} + \nabla_g \beta(z)\{\hat{g}(z) - g(z)\} + o_P(1/\sqrt{n})$$

Lemma 2

$$\nabla_f \beta(z)\{\hat{f}(z) - f(z)\} + \nabla_g \beta(z)\{\hat{g}(z) - g(z)\} = U_1(Z_{it}) + U_2(Z_{it}) = U(Z_{it}) \tag{A.15}$$

Recall:

$$\begin{aligned}\nabla_g \beta(z)\{m(z)\} = &\left[\frac{\alpha_s(z) f_1(z)}{\alpha_1(z)^2 f(z)^2} - \frac{f_s(z)}{\alpha_1(z) f(z)^2}\right]\{m(z)\} + \frac{1}{\alpha_1(z) f(z)}\left\{\frac{\partial}{\partial x_s} m(z)\right\} \\ &- \frac{\alpha_s(z)}{\alpha_1(z)^2 f(z)}\left\{\frac{\partial}{\partial x_1} m(z)\right\} \\ =: &\nabla_g^{(1)} \beta(z)\{m(z)\} + \nabla_g^{(2)} \beta(z)\{m(z)\} + \nabla_g^{(3)} \beta(z)\{m(z)\}\end{aligned}$$

For each component, substitute kernel estimator in, then integrate over z across entire support $\mathcal{Z}$, using weighting function $w(z)$.

First component

$$\begin{aligned}\nabla_g^{(1)} \beta(z)\{\hat{g}(z)\} &= \left[\frac{\alpha_s(z) f_1(z)}{\alpha_1(z)^2 f(z)^2} - \frac{f_s(z)}{\alpha_1(z) f(z)^2}\right]\{\hat{g}(z)\} \\ &= \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} Y_{it} K(Z_{it}, z) \left[\frac{\alpha_s(z) f_1(z)}{\alpha_1(z)^2 f(z)^2} - \frac{f_s(z)}{\alpha_1(z) f(z)^2}\right]\end{aligned}$$

Averaging:

$$\int_{\mathcal{Z}} w(z)\nabla_g^{(1)}\beta(z)\{\hat{g}(z)\}dz = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\int_{\mathcal{Z}} w(z)K(Z_{it},z)\left[\frac{\alpha_s(z)f_1(z)}{\alpha_1(z)^2 f(z)^2} - \frac{f_s(z)}{\alpha_1(z)f(z)^2}\right]dz$$
$$= \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}w(Z_{it})\left[\frac{\alpha_s(Z_{it})f_1(Z_{it})}{\alpha_1(Z_{it})^2 f(Z_{it})^2} - \frac{f_s(Z_{it})}{\alpha_1(Z_{it})f(Z_{it})^2}\right] \times [1 + O_P(h_z^m)]$$

Second component

$$\nabla_g^{(2)}\beta(z)\{\hat{g}(z)\} = \frac{1}{\alpha_1(z)f(z)}\{\hat{g}_s(z)\}$$
$$= \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\frac{\partial K(Z_{it},z)}{\partial x_s^c}\frac{1}{\alpha_1(z)f(z)}$$

Averaging:

$$\int_{\mathcal{Z}} w(z)\nabla_g^{(2)}\beta(z)\{\hat{g}(z)\}dz = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\int_{\mathcal{Z}} w(z)\frac{\partial K(Z_{it},z)}{\partial x_s^c}\frac{1}{\alpha_1(z)f(z)}dz$$

change of variables

$$= \frac{-1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\int_{\mathcal{Z}} K(Z_{it},z)\frac{\partial}{\partial x_s^c}\left[\frac{w(z)}{\alpha_1(z)f(z)}\right]dz$$
$$= \frac{-1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\frac{\partial}{\partial x_s^c}\left[\frac{w(Z_{it})}{\alpha_1(Z_{it})f(Z_{it})}\right][1 + O_P(h_z^m)]$$
$$= \frac{-1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\frac{\alpha_1(Z_{it})f(Z_{it})w_s(Z_{it}) - w(Z_{it})\frac{\partial}{\partial x_s^c}[\alpha_1(Z_{it})f(Z_{it})]}{\alpha_1(Z_{it})^2 f(Z_{it})^2}[1 + O_P(h_z^m)]$$
$$= \frac{-1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\frac{\alpha_1(Z_{it})f(Z_{it})w_s(Z_{it}) - w(Z_{it})\alpha_1(Z_{it})f_s(Z_{it}) - w(Z_{it})\alpha_1 s(Z_{it})f(Z_{it})}{\alpha_1(Z_{it})^2 f(Z_{it})^2}[1 + O_P(h_z^m)]$$

Third component

$$\nabla_g^{(3)}\beta(z)\{\hat{g}(z)\} = -\frac{\alpha_s(z)}{\alpha_1(z)^2 f(z)}\{\hat{g}_1(z)\}$$
$$= \frac{-1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\frac{\partial K(Z_{it},z)}{\partial x_1^c}\frac{\alpha_s(z)}{\alpha_1(z)^2 f(z)}$$

Averaging:

$$\int_{\mathcal{Z}} w(z)\nabla_g^{(3)}\beta(z)\{\hat{g}(z)\}dz = \frac{-1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\int_{\mathcal{Z}} w(z)\frac{\partial K(Z_{it},z)}{\partial x_1^c}\frac{\alpha_s(z)}{\alpha_1(z)^2 f(z)}dz$$
$$= \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\int_{\mathcal{Z}} w(z)K(Z_{it},z)\frac{\partial}{\partial x_1^c}\left[\frac{\alpha_s(z)}{\alpha_1(z)^2 f(z)}\right]dz$$
$$= \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}\frac{\partial}{\partial x_1^c}\left[\frac{\alpha_s(Z_{it})w(Z_{it})}{\alpha_1(Z_{it})^2 f(Z_{it})}\right][1+O_P(h_z^m)]$$
$$= \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} Y_{it}[\frac{\alpha_1(Z_{it})^2 f(Z_{it})(\alpha_s(Z_{it})w_1(Z_{it})+w(Z_{it})\alpha_{s1}(Z_{it})}{\alpha_1(Z_{it})^4 f(Z_{it})^2}$$
$$\cdot\frac{-\alpha_s(Z_{it})w(Z_{it})(\alpha_1(Z_{it})^2 f_1(Z_{it})+2f(Z_{it})\alpha_1(Z_{it})\alpha_{11}(Z_{it})}{\alpha_1(Z_{it})^4 f(Z_{it})^2}][1+O_P(h_z^m)]$$

So, under appropriate conditions,

$$\nabla_g\beta(z)\{\hat{g}(z)\} =$$

Since $\sqrt{N}[h_x^m] = o(1)$

Lemma 3 $||\hat{\alpha}_s(x,y,q)-\alpha_s(x,y,q)||^2 = o_P(1/\sqrt{n})$ and $||\hat{\alpha}_1(x,y,q)-\alpha_1(x,y,q)||^2 = o_P(1/\sqrt{n})$

Bibliography

Altonji, J., and R. Matzkin (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 73(4), 1053–1102.

Andersen, E. (1973): *Conditional Inference and Models for Measuring*. Mentalhygiejnisk Forlag.

Anderson, E. (2002): "A Guadagni-Little likelihood can have multiple maxima," *Marketing Letters*, 13(2), 135–159.

Arellano, M., and R. Carrasco (2003): "Binary choice panel data models with predetermined variables," *Journal of Econometrics*, 115, 125–157.

Arulampalam, W., A. Booth, and M. Taylor (2000): "Unemployment Persistence," *Oxford Economic Papers*, 52, 24–50.

Blundell, R., and J. Powell (2004): "Endogeneity in Semiparametric Binary Response Model," *Review of Economic Studies*, 71, 655–679.

Bronnenberg, B., J.-P. Dubè, and M. Gentzkow (forthcoming): "The Evolution of Brand Preferences: Evolution from Consumer Migration," *American Economic Review*.

Browning, M., and J. Carro (2007): *Heterogeneity and Microeconometrics Modelling*vol. 3. Cambridge University Press.

——— (2009): "Heterogeneity in dynamic discrete choice models," *Working paper*.

Chalak, K., and H. White (2006): "An Extended Class of Instrumental Variables for the Estimation of Causal Effects," *Working Paper*.

Chamberlain, G. (1984): *Handbook of Econometrics*vol. 2, chap. 22 - Panel data. Elsevier Science Publishers.

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78(1), 159–168.

Chernozhukov, V., I. Fernàndez-Val, J. Hahn, and W. Newey (2010): "Average and Quantile Effects in Nonseparable Panel Models," *Working Paper*.

Chiappori, P.-A., I. Komunjer, and D. Kristensen (2011): "Nonparametric Identification and Estimation of Transformation Models," *Working Paper*.

Chintagunta, P., E. Kyriazidou, and J. Perktold (2001): "Panel data analysis of household brand choices," *Journal of Econometrics*, 103(1-2), 111–153.

Croda, E., and E. Kyriazidou (2002): "Intertemporal Labor Force Participation of Married Women in Germany: A Panel Data Analysis," *Working Paper*.

Dubé, J.-P., G. Hitsch, and P. Rossi (2009): "Do Switching Costs Make Markets Less Competitive," *Journal of Marketing Research*, 46(4).

——— (2010): "State dependence and alternative explanations for consumer inertia," *The RAND Journal of Economics*, 41(3), 417–445.

Dubé, J.-P., G. Hitsch, P. Rossi, and M. A. Vitorino (2008): "Category pricing with state-dependent utility," *Marketing Science*, 27(3), 417–429.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*. CRC Press.

Hardle, W., and T. M. Stoker (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84(408), 986–995.

Heckman, J. (1981): "Heterogeneity and State Dependence," in *Studies in Labor Markets*, ed. by S. Rosen, NBER Chapters, chap. 31 Heterogeneity and State Dependence, pp. 91–140. University of Chicago Press.

Hoderlein, S. (2008): "Endogeneity in Semiparametric Binary Random Coefficient Models," *Working Paper*.

Hoderlein, S., and E. Mammen (2007): "Identification of Marginal Effects in Nonseparable Models without Monotonicity," *Econometrica*, 75(1513-1519).

Hoderlein, S., and H. White (2010a): "Nonparametric identification in nonseparable panel data models with generalized fixed effects," *Working Paper*.

——— (2010b): "Nonparametric Identification in nonseparable panel data models with unobservable heterogeneity," *Working paper*.

Honorè, B., and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68(4), 839–874.

Honorè, B., and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models without Strictly Exogenous Regressors," *Econometrica*, 70(5), 2053–2063.

Honorè, B., and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74(3), 611–629.

Horowitz, J. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64(1), 103–137.

Hyslop, D. (1999): "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67(6), 1255–1294.

Imbens, G., and W. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5), 1481–1512.

Keane, M. (1997): "Current Issues in Discrete Choice Modeling," *Marketing Letters*, 8(3), 307–322.

Khan, S., and E. Tamer (2009): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Working Paper*.

Knight, S., M. Harris, and J. Loundes (2002): "Dynamic Relationships in the Australian Labour Market: Heterogeneity and State Dependence," *Economic Record*, 78, 284–298.

Lee, M.-J., and Y.-H. Tae (2005): "Analysis of Labour Participation Behavior of Korean Women with Dynamic Probit and Conditional Logit," *Oxford Bulletin of Economics and Statistics*, 67(1).

Lewbel, A. (1997): "Semiparametric estimation of location and other discrete choice moments," *Econometric Theory*, 13, 32–51.

——— (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97, 145–1777.

Magnac, T. (2004): "Panel binary variables and sufficiency: generalizing conditional logit," *Econometrica*, 72(6), 1859–1876.

Manski, C. (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica*, 55(2), 357–362.

——— (2004): "Measuring Expectations," *Econometrica*, 72(5), 1329–1376.

Misra, K., J. Roberts, and B. Handel (2012): "Robust firm pricing with panel data," *Working paper*.

Powell, J., J. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica*, 57, 1403–1430.

Rasch, G. (1960): *Probabalistic Models for Some Intelligence and Attainment Tests*. Paedagogiske Institut, Copenhagen, Denmark.

Ridder, G. (1990): "The Nonparametric Identification of Generalized Accelerated Failure-Time Models," *Review of Economic Studies*, 57, 167–182.

Wooldridge, J. (2005): "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," *Journal of Applied Econometrics*, 20, 39–54.