Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects

Stefan Hoderlein* Halbert White
Brown University UC San Diego

March 25, 2010

Abstract

This paper is concerned with extending the familiar notion of fixed effects to nonlinear setups with infinite dimensional unobservables like preferences. The main result is that a generalized version of differencing identifies local average structural derivatives (LASDs) in very general nonseparable models, while allowing for arbitrary dependence between the persistent unobservables and the regressors of interest, even if there are only two time periods. These quantities specialize to well known objects like the slope coefficient in the semiparametric panel data binary choice model with fixed effects. We extend the basic framework to include time trends and dynamics in the regressors, and we show how distributional effects as well as average effects are identified. In addition, we show how to handle endogeneity in the transitory component. Finally, we adapt our results to the semiparametric binary choice model with interaction between observables and persistent unobservables, and we establish that average structural marginal probabilities are identified. We conclude this paper by applying the last result to a real-world data example. Using the PSID, we analyze the way in which the lending restrictions for mortgages eased between 2000 and 2004.

Keywords: Nonseparable Models, Identification, Panel Data, Semiparametric, Binary Choice.

*Stefan Hoderlein: Brown University, Department of Economics, Box B, Providence, RI 02921, USA, email: stefan_hoderlein@yahoo.com. Halbert White, University of California, San Diego, Department of Economics, 9500 Gilman Drive, La Jolla, CA 92093-0508, USA, email: hwhite@weber.ucsd.edu. We are indebted to Brendan Beare, Bryan Graham, Jinyong Hong, and to seminar participants at Brown ("Econometrics of Demand"), CEMFI (Madrid), Chicago, Hohenheim, UC San Diego and at UT Austin for helpful comments. Excellent research assistance by Olga Koshevaya at Brown is gratefully acknowledged.

1 Introduction

In linear structures, panel data allow one to deal with persistent but unobservable sources of endogeneity. In many microeconomic data applications, such unobservables include traits that are specific to the individual, e.g., their background or ability, that are clearly correlated with many individual-specific regressors of interest, e.g., income. In linear settings, these invariant factors are typically modelled using a scalar additive unobservable (the "fixed effect"); this is typically removed by taking first differences or by quasi-differencing (see any standard textbook, e.g., Wooldridge, 2008).

This paper establishes that a particular form of differencing can be applied to a large class of nonseparable models. As a special case, we consider binary choice models with unobserved additive scalar transitory effects and infinite-dimensional nonseparable time-invariant unobservables.

Specifically, we first consider the general class of nonseparable panel structures of the form

$$Y_t = \phi(X_t, Z_t, U_t, A), \qquad t = 1, ..., T, \tag{1.1}$$

where, for $i = 1, 2, ...$, $Y_t = Y_{it} \in \mathcal{Y}$ is an observable real-valued random scalar, $(X_t, Z_t) = (X_{it}, Z_{it}) \in \mathcal{X} \times \mathcal{Z} \subseteq \mathbb{R}^{K+L}$ are observable real-valued random K- and L-vectors, respectively, and $(U_t, A) = (U_{it}, A_i) \in \mathcal{U} \times \mathcal{A}$ denote unobservables, respectively time varying and time invariant, both of which are allowed to be of countably infinite dimension. For example, $\mathcal{A}$ may be a Borel space whose elements are piecewise continuous utility functions, whereas $\mathcal{U}$ may be a Borel space whose elements represent piecewise continuous belief functions. The idea is that the first two arguments of ϕ denote drivers of Y_t that we can observe without error, whereas the latter two denote genuinely unobservable causes and characteristics determining Y_t. We assume that interest centers on the effect of X_t on Y_t, whereas we only want to account or control for the influence of all other variables, whether observed like Z_t or unobserved – we are not primarily interested in their effects.

Note further that we will not assume any type of monotonicity of ϕ in U_t or A. This will imply that the function ϕ itself and its derivatives are not identified. Because of this fact, interest centers on averages or conditional averages of this function. In the binary treatment effect literature, $X \in \{0, 1\}$, and the objects of interest are of the form $\mathbb{E}\left[\phi(1, U_t, A) - \phi(0, U_t, A)|\mathcal{F}\right]$, where we have omitted the dependence on Z_t. These are often denoted $\mathbb{E}\left[Y_1 - Y_0|\mathcal{F}\right]$; examples of $\mathcal{F}$ include the trivial sigma algebra, in which case one obtains the average treatment effect, and $X = 1$, in which case one obtains the average effect of treatment on the treated; see Heckman and Vytlacil (2008) for a discussion. In the continuous case, starting with Chamberlain (1982), researchers focused on the average partial effect (APE), where the word "partial" refers to the partial derivative. To denote these derivatives, we let $D_x f$ denote the row vector

of partial derivatives of f with respect to the elements of x; we also let $D_{xx}f$ be the $K \times K$ Hessian of f. Examples of the APEs include in particular Altonji and Matzkin (2005), equation (2.1), who analyze in our notation $\mathbb{E}\left[D_x\phi(X_t, U_t, A)|X_t = x\right]$, under additional restrictions on the dependence between X_t and A that are closer in spirit to "random effects" assumption. But exactly the same effect is also analyzed in Graham and Powell (2009) who consider the "correlated random coefficients" model, a submodel of our model where X_t enters linearly, i.e. $Y_t = \beta(U_t, A)X_t$, in which case the effect of interest becomes $\mathbb{E}\left[\beta(U_t, A)\right]$. These APEs are derivatives of the average structural function of Blundell and Powell (2004)

In this paper, we are concerned with the APE as well. Due to the similarity with the Blundell and Powell ASF, we interpret this effect as a local average structural derivative (LASD). For simplicity, our focus for most of the paper is on the case where $T = 2$, i.e., we consider the two period case[1]. In the following, we let $\Delta Y := Y_2 - Y_1$, $X = (X_1', X_2')'$, $\Delta X := X_2 - X_1$, $Z = (Z_1', Z_2')'$ and $\Delta Z := Z_2 - Z_1$. Our main theorem establishes that

$$D_\xi \mathbb{E}\left[\Delta Y|\Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} \\ = \mathbb{E}\left[D_x\phi(X_1, Z_1, U_1, A)|\Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right].$$

The expression on the left-hand side involves only observable quantities. The expression on the right is the LASD. For a fixed value $\Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z$, this gives the average structural derivative for the subpopulation characterized by those values of the regressors (hence the term "local", meaning local to these values. Of course, the global average effect can be obtained straightforwardly). Similar quantities are analyzed in the cross-section case with endogeneity in Altonji and Matzkin (2005), Hoderlein (2005, 2008), Hoderlein and Mammen (2007), Imbens and Newey (2009), and Schennach, White, and Chalak (2008), and in the panel case by Chernozhukov, Fernandez-Val, Hahn, and Newey (2009) and Graham and Powell (2009). LASDs are also related to the average structural function of Blundell and Powell (2004). Note the different roles of X and Z. Whereas we differentiate with respect to first differences of the former, we only condition on the latter. Note, moreover, that if Z is time invariant then $\Delta Z = 0$ is automatically satisfied.

We show below that this LASD is identified under mild assumptions. Indeed, we require only conditional independence between U_t and X, conditional on A and Z, and a mild stationarity condition on the error U. To emphasize, Z can be arbitrarily correlated with A and U; we also allow arbitrary correlation between A and U as well as between A and X. Thus, we can indeed say that panel data allow one to correct for the influence of potentially endogenous persistent unobserved heterogeneity in a very general class of models.

[1]Indeed, we consider the fact that our approach can work in such a simple setup a major advantage, compared to some of the other semiparametric approaches.

There are two caveats to our analysis: First, as it stands, our approach does not allow for lagged dependent variables. Second, under the weakest set of assumptions, we can only identify effects for the subpopulation for which $\Delta X = 0$ and $\Delta Z = 0$, a subpopulation which Chamberlain (1982) calls the "stayers". As argued by Graham and Powell (2009), in many applications (e.g., Card (1996)) this is a natural subpopulation and comprises, at least in an approximate sense, large parts of the data. In other applications, this may be restrictive. In this sense, our approach complements Graham and Powell (2009), who exclude the "stayers" from their analysis. The same subpopulation like ours is use as base of the analysis in Evdokimov (2009), and in parts in Chernozhukov, Fernandez-Val, Hahn, and Newey (2009), where some of the variables are held constant.

In the case where interests centers on features that characterize the entire population, there are, however, two possible alternatives within our framework: functional form restrictions and less weak restrictions on the dependence between A and X. An example of the former is imposition of an index structure on the data generating process, e.g., $\phi(X_t, Z_t, U_t, A) = \psi(X_t'\beta_o, Z_t, U_t, A)$. In this case, the subpopulation of stayers suffices to identify the index coefficient β_o up to scale (and hence the ratio of marginal effects). We will use this fact later when discussing the binary choice model. An example for the strengthening of the dependence assumptions is restricting the correlation between increments of the X process and A, given X_1. In this case we obtain

$$D_\xi \mathbb{E}\left[\Delta Y | \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} = \mathbb{E}\left[D_x \phi(X_1, Z_1, U_1, A) | X_1 = x, Z_1 = z\right],$$

implying that we can learn the LASD for the entire population by just considering the subpopulation with $\Delta X = 0$ and $\Delta Z = 0$. Consequently, we do not consider this a fundamental limitation of our approach. Moreover, in addition to average marginal effects, we are also able to describe effects on the conditional distribution of Y, given regressors and fixed effects. As such, our approach extends to reveal interesting additional structural features of the model.

As mentioned earlier, the general approach proposed here can be applied to the semiparametric binary choice panel structure, as in Arellano and Carrasco (2003)[2], e.g.,

$$Y_t = \mathbb{I}\left\{X_t'\beta_o + Z_t'\gamma_o + U_t + A > 0\right\}, \qquad t = 1, ..., T, \tag{1.2}$$

where all variables are as before, but now $\mathcal{U} \times \mathcal{A} \subseteq \mathbb{R}^2$, i.e., A corresponds to the classical notion of fixed effect, $\mathbb{I}\{\cdot\}$ denotes the indicator function, and β_o and γ_o denote unknown coefficients.

[2]Though the model is the same, Arellano and Carrasco (2003) allow for predetermined regressors, which we exclude.

If we let $T = 2$, by arguments similar to those for the general nonseparable case, we can show that under similarly unrestrictive assumptions,

$$\beta_o \propto \mathbb{E}\left[\{D_\xi \mathbb{E}\left[\Delta Y | \Delta X = \xi, X_1, \Delta Z = 0, Z_1\right]|_{\xi=0}\} \, b(X_1, Z_1)\right],$$

where b denotes a user-specified weighting function affecting only the constant of proportionality. As is standard in semiparametric index models, identification is only up to scale. In fact, we obtain identification results for the more general case in which unobserved heterogeneity is present in the coefficients by replacing β_o in this structure with $\beta(A, Z)$.

Perhaps more importantly, we establish how to obtain average marginal probabilities in the binary choice model which are structural in the sense that they control for the infinite dimensional unobservable A even without imposing any index structure on the influence of X_t. These applications and extensions illustrate the wide applicability of our approach.

Related Literature: Analyzing nonlinear panel data models has a long tradition, dating back to the conditional ML approach by Rasch (1960, 1961); see also Andersen (1970) and Chamberlain (1982, 1984). Nonlinear parametric panel data models have frequently been analyzed. For an overview of work related to discrete choice models, see Arellano (2003). Closely related to our work is that of Manski (1987), who considers semiparametric estimation of a non-dynamic binary choice panel data model via a median restriction. Chamberlain (1992) discusses the identification of the dynamic panel data binary choice model, and why the logistic distribution assumption is required for identification of β_o, unless one is willing to assume unbounded support for one of the regressors, as is the case in Manski (1987). For other nonlinear fixed effects models, see also Hausman, Hall, and Griliches (1984) for panel count data, Honore (1992) for panel censored regression, and Kyriazidou (1997) for a panel sample selection model.

Like all of this work, our approach assumes a fixed number of time periods. Indeed, it is one of the appealing features of our approach that we only require $T = 2$. This distinguishes our approach from some of the work on the dynamic binary choice model that requires several time periods more and that focuses only on a very restrictive subpopulation; see Honore and Kyriazidou (2000).

All of the work just described is concerned with a specific semiparametric model, e.g., the dynamic binary choice model. Approaches that are closer in spirit to our work are those of Chernozhukov, Fernandez-Val, Hahn, and Newey (2009), who consider discrete variation, whereas we consider derivatives, and Graham and Powell (2009) who focus on a linear heterogeneous population (i.e., the structure is linear in the coefficients, with coefficients that vary across the population), and not on a fully nonseparable structure. Graham and Powell (2009) also require at least as many time periods as regressors (resp. parameters) to estimate, while we require only two time periods. Altonji and Matzkin (2005) treat models satisfying an exchangeability

assumption more closely related to random effects than the fixed effects treated here. Finally, Bester and Hansen (2009) consider a fixed-effects model where the regressors enter through an index structure, and are weakly separable from the correlated unobservable, whereas we can allow for random coefficients in, e.g., binary choice models.

Outline of the Paper: After this introduction, we focus directly on the main identification result. We start with a discussion of the precise assumptions we require, and present and discuss the main result, which establishes the identification of LASDs by generalized differencing. We provide heuristics for the arguments involved and discuss a number of extensions: We show how to treat dynamics in the regressors, as well as distributional effects. We also provide guidance regarding the introduction of time trends and discuss how to deal with endogeneity in X_t beyond that already permitted by our assumptions. To conclude this section, we give a brief discussion of estimation under our assumptions, and we also show how to identify average effects. In the third section, we show how the semiparametric panel data binary choice model can be identified, using the proof of the main theorem as foundation. We discuss two special cases, one where the regression coefficients are random, and one where they are fixed. We show that in both cases, average marginal conditional probabilities are identified (conditional on unobservables A). We also provide constructive identification for the coefficient of interest, β_o, in the fixed-parameter case and we give a closed form expression for the coefficient β_o that can be used to construct a sample counterpart estimator. We conclude this section with a discussion of the estimator's properties. All these concepts are put to the test in an application, using data from the PSID to study the relation between income and the probability of home ownership. The final section contains a summary and concluding remarks.

2 Identification of Marginal Effects in Nonseparable Functions via Differencing

Assumptions and Notations: To keep the exposition as transparent as possible, we consider the simplest possible panel data generating process, in which there are just two time periods, $t = 1, 2$. In this case, we have

$$\begin{aligned} Y_1 &= \phi(X_1, Z_1, U_1, A) \\ Y_2 &= \phi(X_2, Z_2, U_2, A). \end{aligned} \tag{2.1}$$

We assume that Y_t is a scalar random variable, and that X_t and Z_t are random vectors of finite dimension K and L, respectively. On the other hand, the unobservables U_t and A are random vectors of possibly countably infinite dimension.

In addition, we impose sufficient regularity on the conditional cumulative distribution function (CDF) $F_{A|\Delta X,X_1,\Delta Z,Z_1}(a \mid \xi, x, 0, z)$ to ensure that it has a density representation

$$f(a \mid \xi, x, 0, z)\ \mu(da \mid x, 0, z). \tag{2.2}$$

Here, we understand f to be the Radon-Nikodým derivative; e.g., it is a conditional density if A is continuous (so A is absolutely continuous with respect to Lebesgue measure μ), or a conditional probability if A is discrete (i.e., μ is counting measure). In stating our assumptions, we understand that conditions that hold "almost everywhere$-\mu$" ($a.e.-\mu$) are with respect to $\mu(\cdot \mid x, 0, z)$. Further, we let $\nu(u, a \mid x, z)$ denote the product measure defined by $F(u \mid a, 0, z) \times \mu(a \mid x, 0, z)$. Functions that are "$\nu-$integrable" are understood to be integrable with respect to $\nu(\cdot, \cdot \mid x, z)$. We let $||\xi|| := [\xi'\xi]^{1/2}$ denote the Euclidean norm, and we define the neighborhood $\mathcal{N}_\varepsilon = \{\xi : \|\xi\| < \varepsilon\}$. Finally, we write $|M| := \max_{i,j} |M_{i,j}|$ for any matrix $M := [M_{i,j}]$.

We impose the following assumptions:

Assumption 1. *Let $(\Omega, \mathcal{F}, P)$ be a complete probability space on which are defined the random vectors $A : \Omega \to \mathcal{A}$, $\mathcal{A} \subseteq \mathbb{R}^\infty$, and $(Y_t, X_t, Z_t, U_t) : \Omega \to \mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \times \mathcal{U}$, $\mathcal{Y} \subseteq \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^K$, $\mathcal{Z} \subseteq \mathbb{R}^L$, $\mathcal{U} \subseteq \mathbb{R}^\infty$, $t = 1, 2$, with K and L finite integers, such that for $t = 1, 2$, (i) $\mathbb{E}(|Y_t|) < \infty$; (ii)*

$$Y_t = \phi(X_t, Z_t, U_t, A),$$

where $\phi : \mathcal{X} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{A} \to \mathcal{Y}$ is a Borel measurable function; and (iii) realizations of (Y_t, X_t, Z_t) are observable, whereas those of (U_t, A) are not.

Assumption 2. *There exists $\varepsilon > 0$ such that*

$$U_t \perp (\mathbb{I}\{\|\Delta X\| < \varepsilon\}\, \Delta X, X_1) \mid A, \Delta Z = 0, Z_1 \quad t = 1, 2. \tag{2.3}$$

Assumption 3. *U_t is conditionally stationary: $F_{U_1|A,\Delta Z=0,Z_1} = F_{U_2|A,\Delta Z=0,Z_2}$.*

Assumption 4. *$\mathcal{X}$ is an open convex set, and for each $(z, u, a) \in \mathcal{Z} \times \mathcal{U} \times \mathcal{A}$, $\phi(\cdot, z, u, a)$ is twice continuously differentiable on $\mathcal{X}$. Further, $\mathbb{E}[D_x\phi(X_1, Z_1, U_1, A)] < \infty$ and $\mathbb{E}[D_{xx}\phi(X_1, Z_1, U_1, A)] < \infty$.*

Assumption 5. *For each $(x, z) \in \mathcal{X} \times \mathcal{Z}$, there exists a $\sigma-$finite measure $\mu(\cdot \mid x, 0, z)$ absolutely continuous with respect to $F(\cdot \mid \xi, x, 0, z)$ for all $\xi \in \mathcal{N}_\varepsilon$, so that there exists a Radon-Nikodým density f such that for each $(a, \xi) \in \mathcal{A} \times \mathcal{N}_\varepsilon$, $F(da \mid \xi, x, 0, z) = f(a \mid \xi, x, 0, z)\ \mu(da \mid x, 0, z)$.*

Assumption 6. *For each $(x, z) \in \mathcal{X} \times \mathcal{Z}$, $D_\xi f(a \mid \xi, x, 0, z)$ exists $a.e.-\mu$ for all $\xi \in \mathcal{N}_\varepsilon$.*

Assumption 7. *For each $(x,z) \in \mathcal{X} \times \mathcal{Z}$, there exists a $\nu-$integrable dominating function $(u,a) \to D(u,a \mid x,z)$ such that*

$$\begin{aligned} \sup_{\xi \in \mathcal{N}_\varepsilon} |D_{xx}\phi(x+\xi,z,u,a)\ f(a \mid \xi,x,0,z)| &\leq D(u,a \mid x,z) \\ \sup_{\xi \in \mathcal{N}_\varepsilon} |D_x\phi(x+\xi,z,u,a)'\ D_\xi f(a \mid \xi,x,0,z)| &\leq D(u,a \mid x,z). \end{aligned}$$

Assumption 8. *$f(a \mid 0,x,0,z) = f(a \mid x,z)$ for all $(a,x,z) \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$.*

Discussion of Assumptions: These assumptions merit some discussion. First, assumption **A1** formally specifies the data generating process discussed at the beginning of this section. The fact that ϕ is time invariant rules out unrestricted time trends; however, we can include trends when they enter in a specific fashion, as discussed below.

Next, assumption **A2** specifies the sense in which X is exogenous. Conditional on A and $Z_1, \Delta Z = 0$, U_1 is independent of X_1 and ΔX, and similarly for U_2. Note that for the differences ΔX, the independence condition only has to hold for small values of the increment, leaving larger magnitude values out of account. Note in addition that Z_1 and ΔZ may be arbitrarily correlated with U_t, and that the U_t process may exhibit time series dependence. Below, we will discuss an extension addressing the case when this assumption does not hold.

What does this condition mean in economic terms? Suppose we have a data set involving individual-specific information on demand for some good, and on the income and the household characteristics of each of a set of individuals; and assume for simplicity that all household characteristics are time invariant. In addition, assume that we only want to control for the influence of the household characteristics on the income (X) effect, but we are not interested in the impact of household characteristics per se (i.e., these play the role of Z). Assume for the moment that A and Z_t are discrete, and that we stratify the population according to individual values (z,a), which we call a "cell". Then we require that the errors (i.e., transitory shocks) U_1 and U_2 are marginally, but not necessarily jointly, independent of income X_1 and its increments ΔX for small values of the increment, within every cell. Suppose we have data on gender and type of occupation in Z, and A is (unobserved) ability. Then this means that U_t is distributed independently of income and small income changes for, e.g., all high-ability female workers in the iron industry. Nevertheless, income and transitory shocks U_t are allowed to be dependent unconditionally, i.e., ignoring the (z,a) values.

If we are interested in the effects of all regressors, there are effectively no Z's to condition on, and the condition becomes closely related to the strict exogeneity condition in textbook linear panel data models, with the only (important) exceptions that we can allow for correlation between A and U and that A and U can have arbitrarily large dimension. This already weak condition could be weakened further, if we have additional conditioning instruments (e.g., past values of X); see the discussion below.

The next assumption, **A3**, is a mild conditional stationarity requirement for the unobservable drivers. It essentially says that the conditional distribution of U_t is time invariant. As we will see below, this assumption rules out lagged dependent variables as regressors. **A4** specifies that the function ϕ is differentiable in the directions of interest, so that it admits a mean-value expansion. Moreover, the integrability conditions ensure that the needed expectations are well defined. We also suppose that $f(a \mid \xi, x, 0, z)$ is continuously differentiable at $\xi = 0$ **(A6).** Differentiability, combined with the fact that we are considering a neighborhood of zero in the changes of the X variable (for all values of X), implies that we are effectively requiring X to be continuously distributed. Hence, this approach rules out discrete random variables X. Thus, our world is one of continuous variables and differentiation[3]. Note, however, that we require neither condition for the covariates Z, and we do not impose any restriction on the correlation of Z with all the unobserved variables. This parallels discussions in the cross-section case (see Hoderlein (2005, 2008) and Schennach, White, and Chalak (2008)).

In contrast to these material assumptions, assumptions **A5** and **A7** can be seen as regularity conditions. The latter allows one to interchange integration and differentiation, and only the former has some binding content in an economic sense: It allows the conditional probability of A, Z to depend on realized values of $\Delta X, X_1$, but it does not permit the *possible* values for A, Z to depend on these realized values. For example, the support cannot be discrete for some values of x and continuous for others.

Finally, the last assumption **A8** is again material: It restricts the correlation between the increments of the X_t and Z_t processes and A, conditional on X_1, Z_1. This condition is discussed in detail in a companion paper, Hoderlein and White (2009). Since, as already mentioned, we do not require this for the main result, we only mention here that it is fulfilled with correlated A, for example if the influence of A is additively separable from other drivers of X, say, $X_t = \psi(X_{t-1}, Z_t, U_t) + \lambda(A)$. Nevertheless, this type of restriction need not hold generally. Observe again that the conditioning on Z admits arbitrary dependence, as in **A2** above.

The Main Result: The assumptions introduced above now allow us to identify the object of interest, the LASD. Our result is as follows:

Theorem 1. *Let assumptions **A1**–**A7** hold. Then*

$$D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} = \mathbb{E}\left[D_x \phi(X_1, Z_1, U_1, A) \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right],$$

[3]If interest centers on the effect of discrete variables, then we refer the interested reader to Chernozhukov, Fernandez-Val, Hahn, and Newey (2009), which provides a complement to our approach.

*with probability one. If in addition assumption **A8** holds, then*

$$D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} = \mathbb{E}\left[D_x \phi(X_1, Z_1, U_1, A) \mid X_1 = x, Z_1 = z\right], \tag{2.4}$$

with probability one.

Remark: 2.1 - Discussion of Theorem 1: Our main result establishes that certain conditional averages of structural derivatives are identified. The left hand side involves only observables: It is simply the derivative of the nonparametric regression of ΔY on $\Delta X, X_1, \Delta Z, Z_1$ with respect to the first arguments (i.e., ΔX), evaluated at arbitrary positions $X_1 = x, Z_1 = z$ and at $\Delta X = 0, \Delta Z = 0$. The right hand side is exactly the LASD introduced above.

What is this effect, and why is it economically relevant? Consider the demand example we introduced above, but assume now that we have a time-varying covariate, say, years of education. Then we can determine the average marginal effect of income on, say, food demand for all female workers in the iron industry earning \$50,000 and having 10 years of education, whose income and years of education did not change between the periods. But we are not able to identify the marginal effect for every single individual woman. Note that we may allow for omitted persistent factors like preferences, which may be arbitrarily correlated with income, occupation, or years of education.

The difference between the first and the second statement of this theorem is that under the stronger assumption **A8**, we may actually determine the average marginal effect for *all* female workers in the iron industry earning \$50,000 and having 10 years of education, regardless of whether their income or years of education change or not.

These quantities are similar to those considered in Altonji and Matzkin (2005), Chernozhukov, Fernandez-Val, Hahn, and Newey (2009) and to Graham and Powell (2009), and are closely related to the LASD of Hoderlein (2005, 2008) and Hoderlein and Mammen (2007), to the covariate-conditioned average effects of White and Chalak (2008), and to derivatives of the average structural function of Blundell and Powell (2004). They reduce to well known quantities like β_o in the linear model; see also the binary choice model below.

One might object that focusing on the effect of a continous variable for which there is no time variation ($\Delta X = 0$) is not of interest, as such cases form a set of measure zero. But this overlooks several important considerations. First, identification and estimation at points for continuous random variables is common; this is precisely what kernel estimation does. Second, to refute the null hypothesis of no effect, it only necessary to reject the hypothesis at a single point; such identification and estimation is thus valuable for hypothesis testing. Further, however, because economic relations are typically smooth, identification or hypothesis rejection is not literally just for a set of measure zero; knowledge about effects at a point provides useful

information at least about the local neighborhood and possibly more[4]. That is, by learning what the effect is for $\Delta X = 0$, we also learn approximately what the effect is for individuals with a somewhat stable X. A related point is that this local information is precisely what permits estimation of the effect at a point. We only require observations in the neighborhood of $\Delta X = 0$, not just *at* $\Delta X = 0$. Given a sufficiently large sample, estimation of such effects is not only possible, but we can estimate these effects with some precision, giving power to tests of no effect. Finally, one should expect to pay some price for learning anything at all about effects when, as is true here, there may be endogeneity of X outside the neighborhood of $\Delta X = 0$ and arbitrary interactions between X and unobservable attributes A. The need to restrict attention to effects in the neighborhood of $\Delta X = 0$ is the price paid here for this generality. For examples where effects have similarly been identified for $\Delta X = 0$, see Evdokimov (2009), or, for constant covariates, $\Delta Z = 0$, Chernozhukov, Fernandez-Val, Hahn, and Newey (2009).

Moreover, if we have several time periods, we can perform further pairwise comparisons of the above form. Since, according to the first part of theorem 1, in general the population with $Z_2 - Z_1 = 0$ and, say, $Z_3 - Z_2 = 0$ differ in terms of A, we can use a large panel to make statements about large parts of the population. The precise technical conditions required to identify the LASD for the entire population in the absence of assumption **A8** – but using many time periods instead – is left for future research.

Remark 2.2 - Extensions: Theorem 1 admits a number of interesting extensions. First, Theorem 1 also allows us to accommodate lagged regressors. To see this, consider the three period case:

$$\begin{aligned} \mathcal{Y}_1 &= \phi(\mathcal{X}_1, \mathcal{X}_0, \mathcal{U}_1, A) \\ \mathcal{Y}_2 &= \phi(\mathcal{X}_2, \mathcal{X}_1, \mathcal{U}_2, A) \\ \mathcal{Y}_3 &= \phi(\mathcal{X}_3, \mathcal{X}_2, \mathcal{U}_3, A). \end{aligned} \tag{2.5}$$

Essentially, the same identification strategy as above goes through when we employ time differences in the dependent variable that are further apart than the order of lags of the dependent variable. Specifically, if we rewrite $X_2 = (\mathcal{X}_3', \mathcal{X}_2')'$, $X_1 = (\mathcal{X}_1', \mathcal{X}_0')'$, $Y_2 = \mathcal{Y}_3$, $Y_1 = \mathcal{Y}_1$, $U_2 = \mathcal{U}_3$, and $U_1 = \mathcal{U}_1$, the above structure fits exactly into the framework above. This means that we use $\Delta\tilde{Y} = Y_2 - Y_1 = \mathcal{Y}_3 - \mathcal{Y}_1$, and $\Delta X = (\mathcal{X}_3' - \mathcal{X}_1', \mathcal{X}_2' - \mathcal{X}_0')'$, i.e. wider time differences, and under trivial modifications the result continues to hold.

[4]If we assume that the second derivative of ϕ and $D_\xi f(a \mid \xi, x, 0, z)$ are uniformly bounded, then it follows straightforwardly that the bias is at most of order of the difference, i.e. ξ, implying that the bias vanishes smoothly and that we may expect only a small bias in the neighborhood of $\xi = 0$. The same conditions can be used to obtain bounds.

Second, although unrestricted time trends are excluded, trends can be admitted when

$$Y_t = \phi(t, X_t, Z_t, U_t, A) = \phi_0(X_t, Z_t, U_t, A) + \phi_1(t, U_t, Z_t),$$

say. Note the additive separability between X_t, A, and t. This generalizes the commonly used additive time trend, but is restrictive in that the marginal effects $D_x\phi$ do not depend on t.

Next, analysis parallel to that above permits us to identify not just the average marginal effect in the presence of generalized fixed effects, but also marginal causal effects on essentially any aspect of the conditional response distribution that may be of interest. For example, Heckman, Smith, and Clements (1997) draw attention to these effects in the context of programme evaluation. Imbens and Newey (2009) discuss a variety of such measures. We discuss two examples: The first generalizes the above result to known differentiable transformations (e.g., higher moments), the second uses the conditional CDF.

To illustrate, we first let g denote some known differentiable transformation of Y, e.g., $g(y) = F_Y(y)$, where F_Y is the CDF of Y Then, by the same reasoning as in Theorem **1**,

$$\begin{aligned} &D_\xi \mathbb{E}\left[\Delta g(Y) \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} \\ &\quad = \mathbb{E}\left[D_x \left(g[\phi(X_1, Z_1, U_1, A)) \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right]\right. \\ &\quad\quad = \mathbb{E}\left[D_x\phi(X_1, Z_1, U_1, A) g'(Y) \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right]. \end{aligned}$$

This means in particular that weighted averages of the form $\mathbb{E}\left[D_x\phi f_Y(Y) \mid \cdot\right]$ or a weighting scheme that allows focusing on a subset of Y only, are identified. This is potentially interesting for policy considerations, when it is not just the average marginal effect that one is interested in, but the focus is on the marginal effects for those at particular values of the Y distribution.

In fact, g does not have to be differentiable, although in this case a little different analysis is required. Specifically, consider the conditional CDF, obtained by taking $g_y(\phi) = \mathbb{I}\{\phi \leq y\}$; this is not differentiable. In this case, we can derive the result in a manner entirely parallel to that used next in our treatment of the binary dependent variable. We obtain:

$$\begin{aligned} &D_\xi \mathbb{E}\left[\Delta g_y(Y) \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} \\ &\quad\quad = \mathbb{E}\left[D_x \Psi_y(X_1, Z_1, A) \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right], \end{aligned}$$

where

$$\begin{aligned} \Psi_y(x, z, a) &= \int \mathbb{I}\{\phi(x, z, u, a) \leq y\} \; F_{U|A,\Delta Z=0,Z_1}(du \mid a, 0, z) \\ &= \mathbb{P}\left[Y_1 \leq y | X_1 = x, Z_1 = z; \Delta Z = 0, A = a\right] \end{aligned}$$

is assumed differentiable in x. Note that this expression admits again an LASD interpretation. The structural derivative of interest is $D_x\Psi_y(x, z, a)$, and the average effect for this given covariates is exactly the effect that is obtained.

Marginal effects on (a vector of) aspects $a(x, z)$ of the conditional response distribution defined by implicit moments can be similarly analyzed, using the implicit function theorem, as in Chalak and White (2008). These equations can define distributional aspects that optimize a quasi log-likelihood function (e.g., a conditional quantile), or a generalized moment.

Remark 2.3 - Conditional endogeneity of X. Even after isolating the marginal effects of interest from the influence of individual-specific persistent heterogeneity, there could still be dependence between U_t and X_t. That is, assumption **A2** may not hold when conditioning is restricted to the specified conditioning variables. For instance, if X_t is a choice variable, and U_t represents new information revealed to the decision maker, there may well be correlation, even conditional on the information specified in assumption **A2**.

Given suitable additional structure, this issue can be resolved with the use of control variables. Specifically, suppose that X_t is structurally generated as

$$X_t = \chi(W_t, V_t, A), \quad t = 1, 2,$$

where W_t and V_t are observable and unobservable drivers of X_t, respectively, and χ is an unknown measurable function of its arguments such that χ is suitably invertible in V_t, similar to Imbens and Newey (2009). An advantage of panel data is that there are usually natural candidates for W_t, such as past X_t's. In the case where endogeneity arises because of the use of the same information in both decisions (X_t and Y_t), we may well assume that past choice variables reflect past information only and are hence independent of future information.

With suitable structure, we can recover V_t for use as a control variable. For example, suppose that $X_t = \chi_0(W_t) + \chi_1(W_t, A)V_t$ with $V_t \perp (W_t, A)$, and impose the normalizations $\mathbb{E}(V_t) = 0$, $Var(V_t) = 1$. This permits us to solve for V_t as $V_t = Var(X_t|W_t)^{-1/2} [X_t - \mathbb{E}(X_t|W_t)]$, where $\mathbb{E}(X_t|W_t)$ and $Var(X_t|W_t)$ can be straightforwardly estimated.

Under exogeneity conditions for W_t analogous to those ensuring the validity of standard instrumental variables, we can now use V_1 and $\Delta V := V_2 - V_1$ as control variables. Specifically, suppose that

$$(U_t, \Delta V, V_1) \perp (W_1, W_2) \mid A, \Delta Z = 0, Z_1 \quad t = 1, 2.$$

Applying Dawid (1979), lemma 4.2(ii), together with lemmas 4.1 and 4.2(i), we obtain

$$U_t \perp (\mathbb{I}\{\|\Delta X\| < \varepsilon\} \Delta X, X_1) \mid A, \Delta Z = 0, Z_1, \Delta V, V_1 \quad t = 1, 2.$$

We recognize this as a version of **A2** in which Z_1 is augmented by control variables $\Delta V, V_1$.

Remark 2.4 - Chamberlain's impossibility theorem revisited: As mentioned at the outset, Theorem **1** will not cover the case of a lagged dependent variable. This parallels the

discussion in Chamberlain (1992). To see this, consider the system of equations

$$\begin{aligned} Y_1 &= \phi(Y_0, X_1, U_1, A) \\ Y_2 &= \phi(Y_1, X_2, U_2, A). \end{aligned}$$

At first glance it may appear that we can still identify local average structural effects by treating Y_{t-1} as another cause of interest. This, however, violates condition **A2,** as is immediate by simple substitution. The alternative is to treat Y_{t-1} as a conditioning variable, i.e. $Z_t = Y_{t-1}$. Unfortunately, this is not compatible with our assumptions either. Indeed, the problem stems in this case from the otherwise innocuous assumption **A3.** Specifically, it is not possible that

$$U_2 \mid A, Y_1, Y_0 \sim U_1 \mid A, Y_1, Y_0, \tag{2.6}$$

because although U_2 can be plausibly assumed to be independent of Y_1, Y_0 (e.g., if there is no serial correlation amongst the U_t), the condition fails because U_1 helps determine Y_1.

Remark 2.5 - Estimation: Estimation of the quantities of interest here is straightforward. Standard nonparametric regression techniques can be employed, e.g., kernel or series-based methods. As interest attaches to a specific partial derivative or collection of partial derivatives, evaluated at a specified value for the conditioning variables, together with averages of these, it is especially convenient to use kernel methods.

Specifically, we recommend the use of local polynomial regression (as proposed, e.g., by Cleveland (1979)), as these methods are well understood and deliver consistent and asymptotically normal derivative estimators under mild conditions (see, e.g., Fan (1992), Ruppert and Wand (1994), Fan and Gijbels (1996), and Masry (1997)). Further, local polynomials can readily accommodate the empirically significant mixed data case in which some variables are continuous and others are discrete (e.g., Li and Racine (2004)), and they avoid the boundary problems that arise with the use of standard kernel (polynomial of degree zero) methods.

We assume we have data on a panel of individuals, $i = 1, ..., n$, where, for convenience, we may assume the observations are independent and identically distributed (IID). We obtain parameter estimators from local polynomial regression as

$$\hat{\theta}_n(w) = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} [\Delta Y_i - g_p(W_i, \theta)]^2 K_{h_n}(W_i - w),$$

where $g_p(W, \theta)$ defines a polynomial of degree p in $W := (\Delta X', X_1', \Delta Z', Z_1')'$ with parameters θ; K_{h_n} is a multivariate kernel with suitably chosen bandwidth h_n, e.g., the product kernel

$$K_{h_n}(W_i - w) = h_n^{-d} \prod_{\ell=1}^{d} \kappa(\frac{W_{i,\ell} - w_\ell}{h_n}),$$

where $d := 2K+2L$, κ is a univariate kernel, and W_i has elements $W_{i,\ell}$; and $w_0 := (\xi, x, 0, z), \xi \in \mathcal{N}_\varepsilon$, defines the covariate values of interest. We provide further specifics concerning the choice of the kernel and bandwidth in our empirical application below. The considerations involved are entirely standard.

Given an estimator $\hat{\theta}_n$, we estimate the effect of interest,

$$\begin{aligned}\delta^*(x,z) &:= D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} \\ &= \mathbb{E}\left[D_x\phi(X_1, Z_1, U_1, A) \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right],\end{aligned}$$

as

$$\hat{\delta}_n(x,z) := D_\xi g_p(w_0, \hat{\theta}_n(w_0)) \mid_{\xi=0}. \tag{2.7}$$

Under mild conditions (e.g., see Fan & Gijbels (1996), Masry (1997), or Hoderlein (2005)),

$$\sqrt{nh^{d+2}}(\hat{\delta}_n(x,z) - \delta^*(x,z) - h^p\mathcal{B}(x,z)) \xrightarrow{d} \mathcal{N}(0, \Sigma(x,z)),$$

where $\mathcal{B}(x,z)$ and $\Sigma(x,z)$ are the asymptotic bias and covariance matrix respectively.

Remark 2.6 - Average Effects: Interest may also focus on average measures of these conditional effects. Specifically, the assumptions that yield the identifying equation (2.4) also permit us to identify average effects of the form

$$\mathbb{E}_F[D_x\phi(X, Z, U_1, A)] := \int \mathbb{E}[D_x\phi(x, z, U_1, A) \mid X = x, Z = z] \; F(dx, dz),$$

where F is some density of interest specified by the researcher. In particular, we have

$$\delta_0^* := \int D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} F(dx, dz) =: \mathbb{E}_F[D_x\phi(X, Z, U_1, A)],$$

which is the overall average effect across the population for which $\Delta X = 0$ and $\Delta Z = 0$.

This quantity is a partial mean, and it can therefore be estimated by the average

$$\hat{\delta}_{0,n} := n^{-1} \sum_{i=1}^{n} \hat{\delta}_n(X_i, Z_i),$$

where $\{X_i, Z_i\}$ is IID with joint distribution F. Newey (1994) gives conditions under which

$$n^{1/2} h_n^\alpha (\hat{\delta}_{0,n} - \delta_0^*) \xrightarrow{d} \mathcal{N}(0, \Sigma_0),$$

where α is a constant depending on the dimensions of X and Z, and Σ_0 is a covariance matrix whose specific form depends on the choice of g_p. In the appendix, we derive the estimator for Σ_0 for the local linear polynomial, g_1. The derivation for $p > 1$ will be obvious from this.

3 The Endogenous Binary Choice Model

Now consider the case of a binary dependent variable, Y_t, with potential correlation of X_t with U_t and A. As already mentioned, this case can be treated by similar arguments, but not exactly in the same fashion as above. To obtain results for this case, we modify our previous assumptions appropriately. In particular, we specify the structure of interest as follows.

Assumption 9. *Assumption **A1** holds with*

$$\phi(X_t, Z_t, U_t, A) = \mathbb{I}\{\phi_o(X_t, Z_t, A) + U_t > 0\},$$

where ϕ_o is an unknown measurable function and U_t is a random scalar.

Note that this reduces to the textbook binary choice fixed effects case if $\phi_o(X_t, Z_t, A) = X_t'\beta_o + Z_t'\gamma_o + \alpha(A)$ for some unknown vectors β_o and γ_o and some unknown function α.

Assumption 10. *Assumption **A3** holds, and for each $(a, z) \in \mathcal{A}\times\mathcal{Z}$, $u \to F_{U|A,\Delta Z,Z_1}(u \mid a, 0, z)$ is twice continuously differentiable in u for all $u \in \mathcal{U}$, with*

$$\sup_{u\in\mathcal{U}} |D_u F_{U|A,\Delta Z,Z_1}(u \mid a, 0, z)| \leq K_{a,z} < \infty.$$

Assumption 11. *For each $(x, a, z) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Z}$, let*

$$\Psi(x, a, z) := 1 - F_{U|A,\Delta Z,Z_1}(-\phi_o(x, z, a) \mid a, 0, z).$$

For each $(x, z) \in \mathcal{X} \times \mathcal{Z}$ there exists a $\mu-$integrable function $a \to D(a \mid x, z)$ such that

$$\begin{aligned} \sup_{\xi\in\mathcal{N}_\varepsilon} |D_{xx}\Psi(x+\xi, a, z)\ f(a \mid \xi, x, 0, z)| &\leq D(a \mid x, z) \\ \sup_{\xi\in\mathcal{N}_\varepsilon} |D_x\Psi(x+\xi, a, z)'\ D_\xi f(a \mid \xi, x, 0, z)| &\leq D(a \mid x, z). \end{aligned}$$

Assumption 12. *The weighting function $b : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^+$ is measurable such that*

$$0 < \int \{ \int D_x\Psi(x, z, a)\ F(da \mid 0, x, 0, z)\}\ b(x, z)\ F(dx, dz) < \infty.$$

Remark 3.1. Discussion of Assumptions: These assumptions merit some discussion. First, assumption **A9** formally specifies the data generating process. In particular, we consider a latent variable determined by a separable structure; however, the effect of X_t may depend on its own level and may also vary across the population as a function of both the persistent unobservable A (e.g., think of A as preferences) and the observable Z_t. We restrict U_t to enter in an additively separable fashion. In view of Theorem 1, this is not necessary, but we refrain from treating the greatest possible generality here. Instead, we specify a structure that immediately nests the textbook case where $Y_t = \mathbb{I}\{X_t'\beta_o + Z_t'\gamma_o + U_t + A > 0\}$, with β_o

nonrandom and A a scalar. We also provide results for this important special case in Theorem 3. The nonlinear case is nevertheless useful, as it nests random coefficients structures (e.g., $Y_t = \mathbb{I}\{X_t'\beta(A) + Z_t'\gamma(A) + U_t + \alpha(A) > 0\}$), allowing us to treat applications in, e.g., consumer demand or empirical industrial organization, where individual consumers have heterogeneous parameters, or other fields where heterogeneity in individual responses is crucial.

Next, assumption **A10** modifies the differentiability assumptions in **A3** for the binary choice setup. Although the indicator function is obviously not differentiable, we just require differentiability of the conditional CDF of U_t. All other conditions are regularity conditions that ensure that all expectations exist and that interchanging integration and differentiation is warranted. In particular, the domination conditions of assumption **A7** are modified to account for the specific setup here (see **A11**). Finally, the weighting function is formally defined in assumption **A12,** which also ensures that the weighting function is suitably integrable.

To state the nonparametric identification result for this structure, let

$$\begin{aligned} \beta^*(x,z) &:= D_\xi \mathbb{E}[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z]|_{\xi=0} \quad \text{and} \\ \beta_b^* &:= \mathbb{E}\left[\, \beta^*(X_1, Z_1)\; b(X_1, Z_1)\right]. \end{aligned}$$

Both of these involve only the joint distribution of observable random variables and are therefore empirically accessible.

Theorem 2. *Let assumptions **A2**, **A5**, **A6**, and **A9**–**A11** hold.* (*i*) *Then*

$$\begin{aligned} \beta^*(x,z) &= \int D_x\Psi(x,z,a)\; f(a \mid 0,x,0,z)\; \mu(da \mid x,0,z) \\ &= \mathbb{E}\left[\, D_x\mathbb{P}\left[Y_1 = 1 | X_1, \Delta X, Z_1, \Delta Z, A\right] \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right]. \end{aligned}$$

(*ii*) *Suppose that **A12** also holds and that* $\phi_o(x,z,a) = x'\beta(a,z) + \gamma(a,z)$, *where* β *and* γ *are unknown measurable functions. Then* $D_x\Psi(x,z,a) = \beta(a,z)\,\psi(x,z,a)$, *where*

$$\psi(x,z,a) := D_u F_{U|A,\Delta Z,Z_1}(-x'\beta(a,z) - \gamma(a,z) \mid a,0,z) \geq 0,$$

and

$$\beta_b^* = \mathbb{E}(\beta(A,Z_1)\; \psi(X_1,Z_1,A)\; b(X_1,Z_1)).$$

Remark 3.2. Discussion of Theorem 2: This result provides constructive identification of the average structural marginal probabilities for the general case of a population with heterogeneous effects of X_t, including the case of random coefficients, $\beta(A,Z)$. In part (*i*) we establish that the derivative of the conditional expectation provides the best approximation to the derivative of the heterogeneous probabilities for an individual, given the information set $\sigma(X_1, \Delta X, Z_1, \Delta Z, A)$, which means in particular conditioning on the unobserved and high

dimensional A This is close in spirit to the average structural function of Blundell and Powell (2004) for the case of the control function solution to the endogenous binary choice problem. Key steps in the proof of this result follow arguments similar to the general nonseparable case. Note further that the extensions previously discussed continue to be feasible, in particular, the introduction of lagged regressors and the conditional endogeneity of X_t.

In part (ii), we identify weighted averages of the underlying coefficients $\beta(A, Z_1)$, involving partially unknown positive weights $\psi(X_1, Z_1, A)\ b(X_1, Z_1)$. Because interest usually centers on the average marginal probability, we view this as a minor limitation.

This limitation disappears completely when $\phi_o(x, z, a) = x'\beta_o + \gamma(a, z)$, e.g., the population is homogeneous in marginal effects with an additive correlated fixed effect. To state the next result, define

$$\bar{\psi}(x, z) := E(\psi(X_1, Z_1, A) \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z).$$

The precise result is as follows:

Theorem 3. *Suppose the assumptions of Theorem 2(ii) hold for $b(x, z) \equiv 1$. If in addition for all $(a, z) \in \mathcal{A} \times \mathcal{Z}$, $\beta(a, z) = \beta_o$, then β_o is identified up to scale as:*

$$\beta_o = D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \ |_{\xi=0} / \bar{\psi}(x, z),$$

for any (x, z), and as a consequence also by the average partial derivative

$$\beta_o \propto \mathbb{E}\left[D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1, \Delta Z = 0, Z_1\right] \ |_{\xi=0}\ b(X_1, Z_1)\right]$$

*for any b satisfying assumption **A12**.*

Remark 3.3. Discussion of Theorem 3: This result provides constructive identification of the coefficients β_o in the panel data binary choice model. Note, however, that, as is standard, the index structure allows us to identify the object of interest, namely β_o, only up to scale; or, put differently, the ratio of two coefficients is identified. Because of the generality afforded by **A2,** we may allow again for arbitrary dependence between all observed variables and the unobserved components. In particular, our assumptions are weaker than standard strict exogeneity notions in the nonlinear model literature (again, see Arellano (2003)).

Remark 3.4 Average Effects: In the absence of random coefficients, a natural sample counterparts estimator to

$$\begin{aligned} \beta_o &\propto \mathbb{E}\left[D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1, \Delta Z = 0, Z_1\right] \ |_{\xi=0}\ b(X_1, Z_1)\right] \\ &= \mathbb{E}\left[\beta^*(X_1, Z_1)\ b(X_1, Z_1)\right] \end{aligned}$$

is given by

$$\hat{\beta}_{0,n} = n^{-1} \sum_{i=1}^{n} \hat{\delta}_n(X_i, Z_i)\ b(X_i, Z_i),$$

where $\hat{\delta}_n$ is the same local polynomial-based estimator as defined above in equation (2.7).

This $\hat{\beta}_{0,n}$ is a partial means estimator entirely analogous to $\hat{\delta}_{0,n}$ discussed above, except that now a weighting by $b(X_i, Z_i)$ explicitly appears. As above, Newey's (1994) conditions ensure

$$n^{1/2} h_n^{\alpha} (\hat{\beta}_{0,n} - \beta_b^*) \xrightarrow{d} \mathcal{N}(0, \Sigma_{0,b}),$$

where α is a constant depending on the dimensions of X and Z, and $\Sigma_{0,b}$ is a covariance matrix whose specific form depends on the choice of g_p. The derivation for Σ_0 in the appendix applies with obvious modifications to yield an estimator for $\Sigma_{0,b}$.

Even in the presence of random coefficients, this estimator may provide useful information about the direction (sign) of effects of interest.

4 Application: The Vanishing Liquidity Constraint

In this section we demonstrate the ability of our framework to address economically and politically important real-world questions. The specific question we address in our application is the extent of the easing of liquidity constraints between the years 1999 and 2005, which for reasons discussed below was likely a main driver of the current financial crisis. We structure this discussion as follows: We first provide some background. Then we describe the data at hand that allow us to tackle this question. Next, we discuss how the specific question and the data fit into our approach. Finally, we present the results, which suggest that liquidity constraints essentially disappeared in the time period under study.

4.1 The Empirical Question

U.S. subprime mortgage lending has been identified as the likely main culprit of the current global financial crisis. This in turn has caused a downturn in the real economy of a magnitude not seen in the U.S., Europe, or Japan since World War II and the Great Depression. The main chain of the argument for why the crisis that started in the housing market was so harmful runs as follows (see, e.g., "The Subprime Panic" by Gordon (2008)). First, banks changed their lending policy towards previously uncreditworthy customers. This change in lending policy had several drivers: First, it was encouraged by policy makers who wanted to see their view of an "ownership society" established throughout the economy, and who initiated policies and subsidies to encourage home ownership. Second, banks and the wider public held the myopic belief that an ongoing boom in the economy, and in particular the housing market, would

continue for the foreseeable future, whereas by objective criteria the increase in housing prices was already beyond historical precedent (e.g, see the Case-Shiller index[5], for the astonishing run up of prices). This distorted assessment of the risks associated with the housing market then spread across the wider economy, because risky housing loans were made marketable and were actively traded in huge volumes between banks, leading eventually to the collapse or near collapse of historically viable institutions.

Our methods allow us to gather some evidence about the root cause of this crisis. More specifically, we can shed light on the extent of the easing of liquidity constraints, affording a more detailed picture of the roots of the crisis. Using a panel data set, we determine the average marginal effect of income on the probability of owning a home. If lower income individuals are more likely to be liquidity constrained, then we should expect this marginal probability to be positive for lower and mid-level incomes. For sufficiently high income levels, we would expect the marginal probability to decline to near zero as the probability of home ownership stabilizes, other things equal. On the other hand, in the absence of a liquidity constraint, we would expect the marginal probability to lie near zero. By applying our approach, we can account for covariates that are correlated with X (income) whether they are observable (Z in our notation, e.g., age) or unobservable (denoted U (time-varying) or A (persistent)).

This is important, because the probability of buying a house depends both on factors that we observe, like age (younger households are believed to be more mobile, and hence more reluctant to buy a house, other things equal, because of the associated fixed costs (Campbell and Cocco (2007), Li and Yao (2007)), as well as on unobservable but relatively persistent factors like the credit score or other criteria banks use to make their decisions. Another unobservable factor that may well assumed to be constant is initial wealth, a variable which is notoriously hard to measure and also impacts the liquidity constraint. Generally, these factors will be highly correlated with the (transitory) labor income, making a direct regression of home ownership on transitory income potentially highly confounding. In contrast, our approach allows us to determine the average marginal effect of the transitory income on the home ownership, controlling for both observable and persistent unobserved correlated drivers of the decision to buy a house. We will now explain how we isolate the changing influence of transitory income, and hence shed light on one of the specific factors leading to this housing price bubble.

4.2 Data Description

The Panel Study of Income Dynamics (PSID) is a longitudinal sample of U.S. individuals and their families. It is largely representative; however, minorities are oversampled. Although it

[5]http://www2.standardandpoors.com/portal/site/sp/en/us/page.topic/indices_csmahp/0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0.html

is available for a longer period, we use only the years 1999 and 2001, as well as 2003 and 2005, to construct two data sets, each of which comprises its own two-period panel. We then compare our results for average marginal effects across these data sets, to assess changes in the marginal income effect over time. This allows us to isolate the easing of the liquidity constraint with respect to a different attitude towards transitory income risks from other changes in the liquidity constraint.

For the dependent variable, we use information about whether an individual owns a house. We use both the directly elicited variable, as well as an indicator of whether someone pays property tax, without much effect on the results. As regressors, we take income and age of household head. Income is gross yearly income of the entire household, including social security income. For large parts of the population, income varies only little, generating exactly the variation around zero that we require for estimation. Age of the household head in contrast varies between any given two years (e.g., between 1999 and 2001), but in a deterministic and uniform fashion across the population. Hence we use "age in the first period" as a regressor. Note also that we only want to control for age and are not interested in its effect, so that "age in the first period" takes the role of Z, and $\Delta Z = 0$. An alternative would be to use the average age over the respective two year periods. Since in our sample age changes in a completely deterministic fashion (two years later everybody is simply two years older), no added information is employed by conditioning on age in both periods[6]. Taking age into account ensures that we control for the fact that younger individuals may be more mobile for the same income.

We select a sample for which there is information about transitory income in all four years. This reduces the number of households to some 1079 per year. We have experimented with excluding outliers, but it does not materially affect our results.

4.3 Econometric Modelling

Since our dependent variable is binary, we employ the framework of Section 3. Specifically, we apply Theorem 2, which gives

$$\begin{aligned} &D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right]|_{\xi=0} \\ &= \mathbb{E}\left[D_x \mathbb{P}\left[Y_1 = 1 | X_1, \Delta X, Z_1, \Delta Z, A\right] \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right] \end{aligned} \tag{4.1}$$

In our application, $Y_t = 1$ indicates that someone owns a home in period t (as proxied by paying property tax). Consequently, ΔY essentially denotes the change in home ownership between

[6]The implied assumption here is that the equal change in age does not have a differential impact on housing in the second period conditioning on age in the first period, whether you are 25 or 45, This is clearly an approximation, but, we believe, a valid one in the small time interval under consideration.

1999 and 2001, as well as between 2003 and 2005. X_1 is income in either 1999 or in 2003, and ΔX denotes the change in income between 1999 and 2001, or 2003 and 2005, respectively. Z_1 is "age in the first period" (which is either 1999 or 2003); by definition this does not change. Age and income form the complete set of variables that we use (and require) to estimate the left hand side of equation (4.1). However, we do implicity control for further variables by considering the subsample of the population for which health is good in both periods (i.e., we condition on their being no negative health shocks; good health is defined as being in the health categories "excellent", "very good", or "good").

Note, moreover, from the right-hand side of equation (4.1) that we are also implicitly conditioning on all individual-specific (but time-invariant) variables A, and then averaging over these. Thus, we are implicitly controlling for persistent variables like education or race. But we are also controlling for all persistent causes of, e.g., a bad credit score. Controlling for all time-invariant factors is a consequential advantage of our nonparametric approach: If we are not interested in the influence of a time-invariant regressor, we can simply omit it, thus dramatically mitigating the curse of dimensionality and allowing a very parsimonious analysis.

We implement our approach directly using kernel methods, as discussed above, by estimating

$$D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right]|_{\xi=0}, \tag{4.2}$$

as the first derivative of a local quadratic estimator of the regression of the change in home ownership on income and age in the respective first period, and on the change in income, evaluated at income changes equal to zero. We employ a standard Epanechnikov kernel. The bandwidth is chosen by selecting a smaller bandwidth than the cross-validated optimum. Changes in the bandwidth impact our results only marginally; in particular they do not greatly affect the point estimates.

4.4 Results

Applying the methods outlined above to the PSID housing data, we obtain results that can be best summarized graphically. Consider Figure 1, which shows the marginal income effect from the local quadratic estimator for (4.2) applied to the first subset of data, namely 1999/2001 (i.e., around 2000, hence the title of the figure). We show the marginal effect for a value of $z = 40$, which is near the sample mean, and across a 95% window of the log income range.

—-Fig. 1 approx here—-

This graph shows the point estimate of the marginal effect along with a bootstrap-based 95% confidence band. As is clear from this graph, the marginal probability of owning a home with

respect to income is positive everywhere. This means that the probability of home ownership is increasing everywhere, with a noticeable acceleration at the lower to lower-mid income levels (associated with the increase at the left). This is in line with liquidity constraints binding at the lower and middle range of the income distribution for parts of the population. Recall that we are averaging over a heterogeneous population, so at each income level it may be binding for some, but not for others. The effect is highly statistically significant and is almost insensitive to reasonable variations of the bandwidth. In particular, what remains robust at all plausible levels of bandwidth is a positive marginal effect, with some indication of smaller values at the lower end of the income distribution. Since we control for age and implicitly for all other time invariant unobservables that might impact the liquidity constraint, e.g., bad credit history, or wealth in the beginning of the period, we conclude that there is evidence for liquidity constraints, or – from another perspective – more cautious lending by banks in the period around the year 2000 with respect to low transitory income.

This contrasts sharply with the later period, see Figure 2, which depicts exactly the same quantities in the period 2003/2005:

—-Fig. 2 approx here—-

The situation has now changed fundamentally: We do not find a significantly positive effect of income on the marginal probability of owning a home. While the point estimate is still positive it is much closer to zero, and we can only safely infer that the average probability of home ownership did not change across the income range, conditional on other unobservables A, e.g. a bad credit history. The overall ownership rate increased from 0.64 to 0.69 between 1999 and 2005. Assuming that the wealthy did not on net sell their houses in order to rent a house, this means that those individuals at the lower and middle range of the income distribution experienced some catching up in terms of their ownership rate between the 1999/2001 and 2003/2005 periods, with most of the catching up done by those at the low end of the income range. Their lower transitory income alone was simply not an important factor in obtaining a mortgage any longer.

—-Fig. 3 approx here—-

Figure 3 contrasts the two periods, to demonstrate the difference. At the 90% significance level, the 2000 function is outside the pointwise confidence bands of the 2004 function. At the 95% level depicted in Figure 3, the evidence is somewhat inconclusive. However, it is worthwhile emphasizing that these are pointwise tests, which are notorious for their lack of power. We conjecture that L_2 distance tests would very likely reject the hypothesis of equality

at any conventional significance level, if pointwise tests already indicate rejection. As already mentioned, the results do not change if we confine ourselves to a sample of people in constant good health, in employment, or if we consider only the non-Afro-American population[7].

In summary, our findings are consistent with changing U.S. attitudes towards borrowing by consumers and lending by banks leading to an "ownership society," where the differences in income risks associated with different levels of income – especially the risk of default – no longer mattered for home ownership. We emphasize that what we have isolated here is the changing attitude towards transitory income, which we can separate from all time invariant and correlated factors, e.g., a bad credit history, due to generality of our approach. While it is entirely possible that attitudes towards these factors may also have changed in the subprime mortgage crisis, our analysis only (and clearly) establishes this change in attitudes for transitory income.

5 Summary and Conclusions

This paper demonstrates the usefulness of panel data for controlling for individual-specific persistent and potentially correlated unobserved heterogeneity under mild assumptions. We demonstrate that a particular form of first differencing is widely applicable, allowing the recovery of effects of interest in general nonseparable nonparametric structures with general forms of unobserved heterogeneity, e.g., in preferences and beliefs. Moreover, it also allows recovery of effects of interest in certain semiparametric nonlinear panel data models, like the binary correlated random coefficients model, for which no estimation strategy has previously been proposed. The approach is flexible, and admits a variety of extensions: time trends, endogenous transitory components, distributional effects, and lagged regressors are some of these. As it stands, the only major limitation of this approach concerns the inclusion of lagged dependent variables. We leave this issue to future research.

One key feature of our approach is its general nonparametric structure, which provides constructive identification results that can be employed directly to construct nonparametric estimators that have a straightforward economic interpretation. We demonstrate this with an application to housing data, where we find that even after accounting for differences in age and persistent unobserved factors, there was a clear dissociation between income and the probability of owning a home between 1999/2001 and 2003/2005. This easing of liquidity constraints was at least in part a major cause underlying the current economic crisis. Our ability to investigate this cause, controlling for both observed and unobserved factors, underscores the usefulness of our approach as much as its theoretical advantages do.

[7]For the Afro-American population the effect appears to be weaker, but we have too little observations.

6 Appendix

6.1 Proof of Theorem 1

First, we establish

$$D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0}$$
$$= \mathbb{E}\left[D_x \phi(X_1, Z_1, U_1, A) \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right].$$

To see this, we start by using assumption **A1** to write

$$\mathbb{E}\left[Y_2 - Y_1 \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right]$$
$$= \int \left[\phi(x+\xi, z, u_2, a) - \phi(x, z, u_1, a)\right] F_{U_2,U_1,A|\Delta X,X_1,\Delta Z,Z_1}(du_2, du_1, da \mid \xi, x, 0, z).$$

To simplify the notation in what follows, we let the argument list implicitly specify the relevant random variables. Thus, we write

$$F(u_2, u_1, a \mid \xi, x, 0, z) := F_{U_2,U_1,A|\Delta X,X_1,\Delta Z,Z_1}(u_2, u_1, a \mid \xi, x, 0, z).$$

Applying successive conditioning and rearranging, we have

$$\int \left[\phi(x+\xi, z, u_2, a) - \phi(x, z, u_1, a)\right] \; F(du_2, du_1, da \mid \xi, x, 0, z)$$
$$= \int \left[\int \phi(x+\xi, z, u_2, a) F(du_2 \mid a, \xi, x, 0, z) - \int \phi(x, z, u_1, a) F(du_1 \mid a, \xi, x, 0, z)\right] \; F(da \mid \xi, x, 0, z).$$

Next, we use the conditional independence assumption **A2**. This assumption ensures that for all ξ in $\mathcal{N}_\varepsilon = \{\xi : \|\xi\| < \varepsilon\}$ and all other admissible function arguments

$$\begin{aligned} F(u_2 \mid a, \xi, x, 0, z) &= F(u_2 \mid a, 0, z) \quad \text{and} \\ F(u_1 \mid a, \xi, x, 0, z) &= F(u_1 \mid a, 0, z), \end{aligned}$$

so that for all ξ in $\mathcal{N}_\varepsilon$ and all admissible x and z,

$$\mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right]$$
$$= \int \left[\int \phi(x+\xi, z, u_2, a) F(du_2 \mid a, 0, z) - \int \phi(x, z, u_1, a) F(du_1 \mid a, 0, z)\right] \; F(da \mid \xi, x, 0, z).$$

Next, apply the conditional stationarity ensured by assumption **A3.** This gives $\int \phi(x, z, u_2, a)$ $F(du_2 \mid a, 0, z) = \int \phi(x, z, u_1, a) F(du_1 \mid a, 0, z) = \int \phi(x, z, u, a) F(du \mid a, 0, z)$, so that

$$\mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right]$$
$$= \int \left[\int \phi(x+\xi, z, u, a) F(du \mid a, 0, z) - \int \phi(x, z, u, a) F(du \mid a, 0, z)\right] \; F(da \mid \xi, x, 0, z).$$

Assumption **A4** ensures that ϕ is sufficiently smooth to admit a mean value expansion in its first argument:

$$\phi(x+\xi,z,u,a) = \phi(x,z,u,a) + D_x\phi(\bar{x},z,u,a)\xi.$$

The mean value for $D_x\phi(\bar{x},z,u_1,a)$ is given by $\bar{x} = \lambda(x+\xi) + (1-\lambda)x = x + \lambda\xi$, where λ depends on (z,u,a) and takes values in $[0,1]$. For convenience, we reflect these dependencies by writing

$$J(\xi,x,z,u,a) := D_x\phi(\bar{x},z,u,a);$$

we note that $J(0,x,z,u,a) = D_x\phi(x,z,u,a)$. Under the finite expectations assumed in assumption **A4**, we can write

$$\int \phi(x+\xi,z,u,a)F(du \mid a,0,z) = \int [\phi(x,z,u,a) + J(\xi,x,z,u,a)\xi]\ F(du \mid a,0,z),$$

from which it follows that

$$\begin{aligned}
&\mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \\
&\qquad = \xi' \int \left[\int J(\xi,x,z,u,a)'\ F(du \mid a,0,z)\right]\ f(a \mid \xi,x,0,z)\ \mu(da \mid x,0,z),
\end{aligned}$$

where we use assumption **A5** to write $F(da \mid \xi,x,0,z) = f(a \mid \xi,x,0,z)\ \mu(da \mid x,0,z)$.

The next step is to differentiate this expression with respect to ξ. Using the regularity imposed in **A6** and **A7** to ensure the valid interchange of derivative and integral, we obtain

$$\begin{aligned}
&D_\xi\mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z = z\right] \\
&\qquad = \int \left[\int J(\xi,x,z,u,a)\ F(du \mid a,0,z)\right]\ f(a \mid \xi,x,0,z)\ \mu(da \mid x,0,z) \\
&\qquad + \xi' \int \left[\int D_\xi J(\xi,x,z,u,a)\ F(du \mid a,0,z)\right]\ f(a \mid \xi,x,0,z)\ \mu(da \mid x,0,z) \\
&\qquad\qquad + \xi' \int \left[\int J(\xi,x,z,u,a)'\ F(du \mid a,0,z)\right]\ D_\xi f(a \mid \xi,x,0,z)\ \mu(da \mid x,0,z),
\end{aligned}$$

where $D_\xi J(\xi,x,z,u,a)$ is the $K \times K$ matrix with elements $(\partial/\partial\xi_j)J_k(\xi,x,z,u,a)$.

Evaluating this expression at $\xi = 0$, we obtain

$$\begin{aligned}
&D_\xi\mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right]\ |_{\xi=0} \\
&\qquad = \mathbb{E}\left[D_x\phi(X_1,Z_1,U_1,A) \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right],
\end{aligned}$$

yielding the first part of the desired result.

If we strengthen our assumptions by imposing **A8**, i.e., $f(a \mid 0,x,0,z) = f(a \mid x,z)$ for all admissible (a,x,z), then we obtain

$$\begin{aligned}
&D_\xi\mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right]\ |_{\xi=0} \\
&\qquad\qquad = \mathbb{E}\left[D_x\phi(X_1,Z_1,U_1,A) \mid X_1 = x, Z_1 = z\right],
\end{aligned}$$

completing the proof. *Q.E.D.*

6.2 Proof of Theorem 2

(*i*) We again start again by considering $\mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right]$. By **A9**, we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta Y \mid X_1 = x, \Delta X = \xi, \Delta Z = 0, Z_1 = z\right] \\
&\qquad = \int \left[\mathbb{I}\left\{\phi_o(x+\xi, z, a) + u_2 > 0\right\} - \mathbb{I}\left\{\phi_o(x, z, a) + u_1 > 0\right\}\right] \\
&\qquad\qquad \times F(du_1, du_2, da \mid \xi, x, 0, z) \\
&\qquad = \int \Big[\int \mathbb{I}\left\{\phi_o(x+\xi, z, a) + u_2 > 0\right\} F(du_2 \mid a, \xi, x, 0, z) \\
&\qquad\quad - \int \mathbb{I}\left\{\phi_o(x, z, a) + u_1 > 0\right\} F(du_1 \mid a, \xi, x, 0, z)\Big]\ F(da \mid \xi, x, 0, z) \\
&\qquad = \int \Big[\int \mathbb{I}\left\{\phi_o(x+\xi, z, a) + u_2 > 0\right\} F(du_2 \mid a, 0, z) \\
&\qquad\quad - \int \mathbb{I}\left\{\phi_o(x, z, a) + u_1 > 0\right\} F(du_1 \mid a, 0, z)\Big]\ F(da \mid \xi, x, 0, z),
\end{aligned}$$

where, as before, we make use of successive conditioning and assumption **A2**.

Applying the conditional stationarity ensured by **A10**, we can write the first inner integral in the last expression as

$$\begin{aligned}
\Psi(x+\xi, z, a) \quad &:= 1 - F_{U|A,\Delta Z,Z_1}(-\phi_o(x+\xi, z, a) \mid a, 0, z) \\
&= \int \mathbb{I}\left\{\phi_o(x+\xi, z, a) + u > 0\right\} F(du \mid a, 0, z).
\end{aligned}$$

The second inner integral is $\Psi(x, z, a)$, so we can write the final expression above as

$$\int \left[\Psi(x+\xi, z, a) - \Psi(x, z, a)\right]\ F(da \mid \xi, x, 0, z). \tag{6.1}$$

Assumption **A10** ensures that Ψ is sufficiently smooth to admit a mean value expansion in its first argument, so that

$$\Psi(x+\xi, z, a) = \Psi(x, z, a) + D_x\Psi(\bar{x}, z, a)\ \xi.$$

The mean value $\bar{x}$ is given by

$$\bar{x} = \lambda(x+\xi) + (1-\lambda)x = x + \lambda\xi,$$

where λ depends on (z, a) and takes values in $[0, 1]$. For convenience, we reflect these dependencies by writing

$$J(\xi, x, z, a) = D_x\Psi(\bar{x}, z, a);$$

we note that $J(0,x,z,a) = D_x\Psi(x,z,a)$.

The bound imposed in assumption **A10** ensures that we can write

$$\mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] = \xi' \int J(\xi,x,z,a)' \; f(a \mid \xi,x,0,z) \; \mu(da \mid x,0,z),$$

where we use assumption **A5** to write $F(da \mid \xi,x,0,z) = f(a \mid \xi,x,0,z) \; \mu(da \mid x,0,z)$.

Differentiating with respect to ξ, we have

$$\begin{aligned} D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] &= \int J(\xi,x,z,a) \; f(a \mid \xi,x,0,z) \; \mu(da \mid x,0,z) \\ &\quad + \xi' \int D_\xi J(\xi,x,z,a) \; f(a \mid \xi,x,0,z) \; \mu(da \mid x,0,z) \\ &\quad + \xi' \int J(\xi,x,z,a)' \; D_\xi f(a \mid \xi,x,0,z) \; \mu(da \mid x,0,z), \end{aligned}$$

where $D_\xi J(\xi,x,z,a)$ is the $K \times K$ matrix with elements $(\partial/\partial\xi_j)J_k(\xi,x,z,a)$ and the interchange of integral and derivative is justified by assumptions **A6** and **A11**. Evaluating at $\xi = 0$ produces

$$\begin{aligned} \beta^*(x,z) := D_\xi \mathbb{E}\left[\Delta Y \mid \Delta X = \xi, X_1 = x, \Delta Z = 0, Z_1 = z\right] \mid_{\xi=0} &= \int D_x\Psi(x,z,a) \; f(a \mid 0,x,0,z) \; \mu(da \mid x,0,z). \end{aligned}$$

The second result follows because

$$\int D_x\Psi(x,z,a) \; f(a \mid 0,x,0,z) \; \mu(da \mid x,0,z) = \mathbb{E}\left[D_x\mathbb{P}\left[Y = 1 | X_1, \Delta X, Z_1, \Delta Z, A\right] \mid \Delta X = 0, X_1 = x, \Delta Z = 0, Z_1 = z\right].$$

(*ii*) Multiply $\beta^*(X_1,Z_1)$ by the scalar $b(X_1,Z_1)$, and take expectations over the joint distribution of (X_1,Z_1) to obtain

$$\begin{aligned} \mathbb{E}\left[\; \beta^*(X_1,Z_1) \; b(X_1,Z_1)\right] &= \int \{ \int D_x\Psi(x,z,a) \; f(a \mid 0,x,0,z) \; \mu(da \mid x,0,z)\} \; b(x,z) \; F(dx,dz) \\ &= \mathbb{E}(D_x\Psi(X_1,Z_1,A) \; b(X_1,Z_1)). \end{aligned}$$

Because $\phi_o(x,z,a) = x'\beta(a,z) + \gamma(a,z)$,

$$\begin{aligned} D_x\Psi(X_1,Z_1,A) &= -D_x F_{U|A,\Delta Z,Z_1}(-x'\beta(a,z) - \gamma(a,z) \mid a,0,z) \\ &= \beta(a,z) \, \psi(x,z,a), \end{aligned}$$

where

$$\psi(x,z,a) := D_u F_{U|A,\Delta Z,Z_1}(-x'\beta(a,z) - \gamma(a,z) \mid a,0,z) \geq 0,$$

as $D_u F_{U|A,\Delta Z,Z_1}$ is the conditional density corresponding to $F_{U|A,\Delta Z,Z_1}$. It follows that

$$\mathbb{E}[\ \beta^*(X_1,Z_1)\ b(X_1,Z_1)] = \mathbb{E}(\beta(A,Z_1)\,\psi(X_1,Z_1,A)\ b(X_1,Z_1)),$$

and the proof is complete. *Q.E.D.*

6.3 Proof of Theorem 3

Immediate from Theorem 2(*ii*). *Q.E.D.*

Derivation of Newey's (1994) estimator for Σ_0 :

Newey's partial means analysis involves a function m depending on observations z_i (here $(1, W_i', \Delta Y_i)'$), parameters β (here δ_0) and a vector function h whose "true value," h_0, is estimated by a kernel estimator, $\hat{h}$. Here, the analog of $\hat{h}$, say $\hat{\boldsymbol{h}}_n$, determines $\hat{\theta}_n$. An estimator of the partial mean of interest β_0 (here δ_0^*), is given by $\hat{\beta}$ (here $\hat{\delta}_{0,n}$) satisfying

$$n^{-1}\sum_{i=1}^{n} m(z_i, \hat{\beta}, \hat{h}) = 0.$$

For concreteness and simplicity, we work here with the local linear polynomial, $g_1(w) = \theta_{00}(w) + \xi'\theta_{10}(w) + v'\theta_{20}(w)$, where we write $w' := (\xi', v')'$ and $\theta_0 := (\theta_{00}, \theta_{10}', \theta_{20}')'$. Letting $w_{00} := (0', x', 0', z')'$, it follows that

$$\hat{\delta}_n(x,z) = s_\xi\ \hat{\theta}_n(w_{00}),$$

where s_ξ is the $L \times d$ selection matrix that selects the ξ components of w, so that $s_\xi\ \hat{\theta}_n = \hat{\theta}_{1n}$.

With the choice g_1, $\hat{\theta}_n$ is the weighted least squares estimator

$$\hat{\theta}_n = [\hat{\boldsymbol{h}}_{n,1}]^{-1}\ \hat{\boldsymbol{h}}_{n,2},$$

where, with $K_{h_n,i,w} := K_{h_n}(W_i - w)$,

$$\hat{\boldsymbol{h}}_{n,1}(w) := \begin{bmatrix} n^{-1}\sum_{i=1}^n K_{h_n,i,w} & n^{-1}\sum_{i=1}^n K_{h_n,i,w}\ (W_i - w)' \\ n^{-1}\sum_{i=1}^n K_{h_n,i,w}\ (W_i - w) & n^{-1}\sum_{i=1}^n K_{h_n,i,w}\ (W_i - w)(W_i - w)' \end{bmatrix},$$

and

$$\hat{\boldsymbol{h}}_{n,2}(w) := \begin{bmatrix} n^{-1}\sum_{i=1}^n K_{h_n,i,w}\ \Delta Y_i \\ n^{-1}\sum_{i=1}^n K_{h_n,i,w}\ (W_i - w)\ \Delta Y_i \end{bmatrix}.$$

Letting $W_{i,00} := (0', X_i', 0', Z_i')'$ and $\hat{\boldsymbol{h}}_n = (vech'[\hat{\boldsymbol{h}}_{n,1}], \hat{\boldsymbol{h}}_{n,2}')'$, it follows that Newey's $m(z_i, \hat{\beta}, \hat{h})$ corresponds to

$$m(W_{i,00}, \hat{\delta}_{0,n}, \hat{\boldsymbol{h}}_n) = s_\xi\ \{[\hat{\boldsymbol{h}}_{n,1}]^{-1}\hat{\boldsymbol{h}}_{n,2}\}(W_{i,00}) - \hat{\delta}_{0,n}.$$

(In the binary dependent variable case, the weighting by $b(X_i, Z_i)$ is accommodated by instead taking

$$m(W_{i,00}, \hat{\delta}_{0,n}, \hat{\boldsymbol{h}}_n) = b(X_i, Z_i)\ s_\xi\ \{[\hat{\boldsymbol{h}}_{n,1}]^{-1}\hat{\boldsymbol{h}}_{n,2}\}(W_{i,00}) - \hat{\delta}_{0,n},$$

but we leave this implicit in what follows for simplicity.)

Because $(\partial/\partial\delta_0)m(w, \delta_0, \boldsymbol{h}) = \boldsymbol{I}_d$ (the $d \times d$ identity matrix), Newey's (1994) estimator of Σ_0 is given by

$$\hat{\Sigma}_n := n^{-1}\sum_{i=1}^n \hat{\psi}_i\hat{\psi}_i',$$

where

$$\hat{\psi}_i := m(W_{i,00}, \hat{\delta}_{0,n}, \hat{\boldsymbol{h}}_n) + \hat{\eta}_i - n^{-1}\sum_{j=1}^n \hat{\eta}_j,$$

with

$$\hat{\eta}_i := (\partial/\partial\zeta)[n^{-1}\sum_{j=1}^n m(W_{j,00}, \hat{\delta}_{0,n}, \hat{\boldsymbol{h}}_n + \zeta\boldsymbol{\rho}_{n,i,0})]\ |_{\zeta=0}$$

and $\boldsymbol{\rho}_{n,i,0} := (vech'[\boldsymbol{\rho}_{n,i,0,1}], \boldsymbol{\rho}_{n,i,0,2}')'$, where, letting $K_{h_n,i,0}(w) := K_{h_n}(W_{i,00} - w)$,

$$\boldsymbol{\rho}_{n,i,0,1}(w) := \begin{bmatrix} K_{h_n,i,0}(w) & K_{h_n,i,0}(w)\ (W_{i,00} - w)' \\ K_{h_n,i,0}(w)\ (W_{i,00} - w) & K_{h_n,i,0}(w)\ (W_{i,00} - w)(W_{i,00} - w)' \end{bmatrix},$$

and

$$\boldsymbol{\rho}_{n,i,0,2}(w) := \begin{bmatrix} K_{h_n,i,0}(w)\ \Delta Y_i \\ K_{h_n,i,0}(w)\ (W_{i,00} - w)\ \Delta Y_i \end{bmatrix}.$$

By the chain rule and the formula for the derivative of the matrix inverse, we obtain

$$(\partial/\partial\zeta)m(W_{j,00}, \hat{\delta}_{0,n}, \hat{\boldsymbol{h}}_n + \zeta\boldsymbol{\rho}_{n,i,0}) = s_\xi\ (\partial/\partial\zeta)\{[\hat{\boldsymbol{h}}_{n,1} + \zeta\boldsymbol{\rho}_{n,i,0,1}]^{-1}(\hat{\boldsymbol{h}}_{n,2} + \zeta\boldsymbol{\rho}_{n,i,0,2})\}(W_{j,00})$$
$$= s_\xi\ \{[\hat{\boldsymbol{h}}_{n,1} + \zeta\boldsymbol{\rho}_{n,i,0,1}]^{-1}\boldsymbol{\rho}_{n,i,0,2} - [\hat{\boldsymbol{h}}_{n,1} + \zeta\boldsymbol{\rho}_{n,i,0,1}]^{-1}\boldsymbol{\rho}_{n,i,0,1}[\hat{\boldsymbol{h}}_{n,1} + \zeta\boldsymbol{\rho}_{n,i,0,1}]^{-1}(\hat{\boldsymbol{h}}_{n,2} + \zeta\boldsymbol{\rho}_{n,i,0,2})\}(W_{j,00}).$$

Evaluating this expression at $\zeta = 0$ gives

$$(\partial/\partial\zeta)m(W_{j,00}, \hat{\delta}_{0,n}, \hat{\boldsymbol{h}}_n + \zeta\boldsymbol{\rho}_{n,i,0})\ |_{\zeta=0}$$
$$= s_\xi\ \{[\hat{\boldsymbol{h}}_{n,1}]^{-1}(\boldsymbol{\rho}_{n,i,0,2} - \boldsymbol{\rho}_{n,i,0,1}[\hat{\boldsymbol{h}}_{n,1}]^{-1}\hat{\boldsymbol{h}}_{n,2})\}(W_{j,00}),$$

so

$$\hat{\eta}_i := s_\xi\ n^{-1}\sum_{j=1}^n \{[\hat{\boldsymbol{h}}_{n,1}]^{-1}(\boldsymbol{\rho}_{n,i,0,2} - \boldsymbol{\rho}_{n,i,0,1}[\hat{\boldsymbol{h}}_{n,1}]^{-1}\hat{\boldsymbol{h}}_{n,2})\}(W_{j,00}).$$

7 References

References

[1] Andersen, E. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society Series B*, 32, 283-301.

[2] Altonji, J., and R. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053–1103.

[3] Arellano, M. (2003): "Discrete Choice with Panel Data," *Investigaciones Economicas*, 27, 423-458.

[4] Arellano, M. and R. Carrasco (2003): "Binary Choice Panel Data Models with Predetermined Variables," *Journal of Econometrics*, 115, 125-157.

[5] Bester, A. and C. Hansen (2008),"Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model," *Journal of Business and Economic Statistics,* forthcoming.

[6] Blundell, R. and J. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655-679.

[7] Campbell, J. and J. Cocco (2007), "How Do House Prices Affect Consumption? Evidence from Micro Data," *Journal of Monetary Economics,* 54, 591-627.

[8] Card, D. (1996), "The Effects of Unions on the Structures of Wages: a Longitudinal Analysis", *Econometrica*, 69, 957-979.

[9] Chamberlain, G. (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics,* 18(1), 5 - 46.

[10] Chamberlain, G. (1984): "Panel Data," in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics,* Vol. 2. New York: North Holland.

[11] Chamberlain, G. (1992): "Binary Response Models for Panel Data: Identification and Information," Harvard University Working Paper.

[12] Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2009): "Identification and Estimation of Marginal Effects in Nonlinear Panel Models," MIT Working Paper.

[13] Cleveland, W.S. (1979): "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

[14] Dawid, A.P. (1979): "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society Series B*, 41, 1-31.

[15] Evdokimov, K. (2009). ""Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity", Yale University, Working paper.

[16] Fan, J. (1992): "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998-1004.

[17] Fan J., and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

[18] Gordon, G. (2008), "The Subprime Panic," Yale University Working Paper.

[19] Graham, B., and J. Powell (2008): "Identification and Estimation of 'Irregular' Correlated Random Coefficient Models," NBER Working Paper 14469.

[20] Hausman, J., B. Hall, and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52, 909-938.

[21] Heckman, J. J. Smith, and N. Clements (1997): "Making The Most Out of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487-535.

[22] Heckman, J. and E.Vytlacil (2008): "Econometric Evaluation of Social Programs," in: *Handbook of Econometrics*, Vol. 6b, Heckman, J. and E. Leamer (eds.), North Holland.

[23] Hoderlein, S. (2005): "Nonparametric Demand Systems, Instrumental Variables and a Heterogeneous Population," Brown University Working Paper.

[24] Hoderlein, S. (2009): "How Many Consumers are Rational?" Brown University Working Paper.

[25] Hoderlein, S. and H. White (2009): "Dynamic Nonseparable Panel Data Models," Brown University Working Paper.

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

[27] Honore, B. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60, 533-565.

[28] Honore, B. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica,* 68, 839-874.

[29] Honore, B. and E. Tamer (2006): "Bounds on Parameters in Dynamic Discrete Choice Models without Strict Exogeneity," *Econometrica*, 74, 611-629.

[30] Imbens, G. and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481-1512.

[31] Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335-1364.

[32] Li, Q. and J. Racine (2004): "Cross-Validated Local Linear Nonparametric Regression," *Statistica Sinica*, 14, 485-512.

[33] Li, W. and R. Yao (2007): "The Life-Cycle Effects of House Price Changes," *Journal of Money, Credit and Banking*, 39, 1375-1409.

[34] Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, 357-62.

[35] Newey, W. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233-253.

[36] Rasch. G. (1960): "Probabilistic Models for some Intelligence and Attainment Tests," *Denmarks Paedagogiske Institut*, Copenhagen.

[37] Rasch, G. (1961): "On the General Law and the Meaning of Measurement in Psychology," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. Berkeley: UC Press.

[38] Ruppert, D. and M. P. Wand (1994): "Multivariate Locally Weighted Least Squares Regression," *Annals of Statistics*, 22, 1346-1370.

[39] Schennach, S., H. White, and K. Chalak (2008): "Estimating Average Marginal Effects of Endogenous Variables in the Nonseparable Case," UCSD Working Paper.

[40] White, H. and K. Chalak (2008): "Identifying Structural Effects in Nonseparable Models Using Covariates," UCSD Working Paper.

[41] Wooldridge, J. (2008). *Econometrics of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

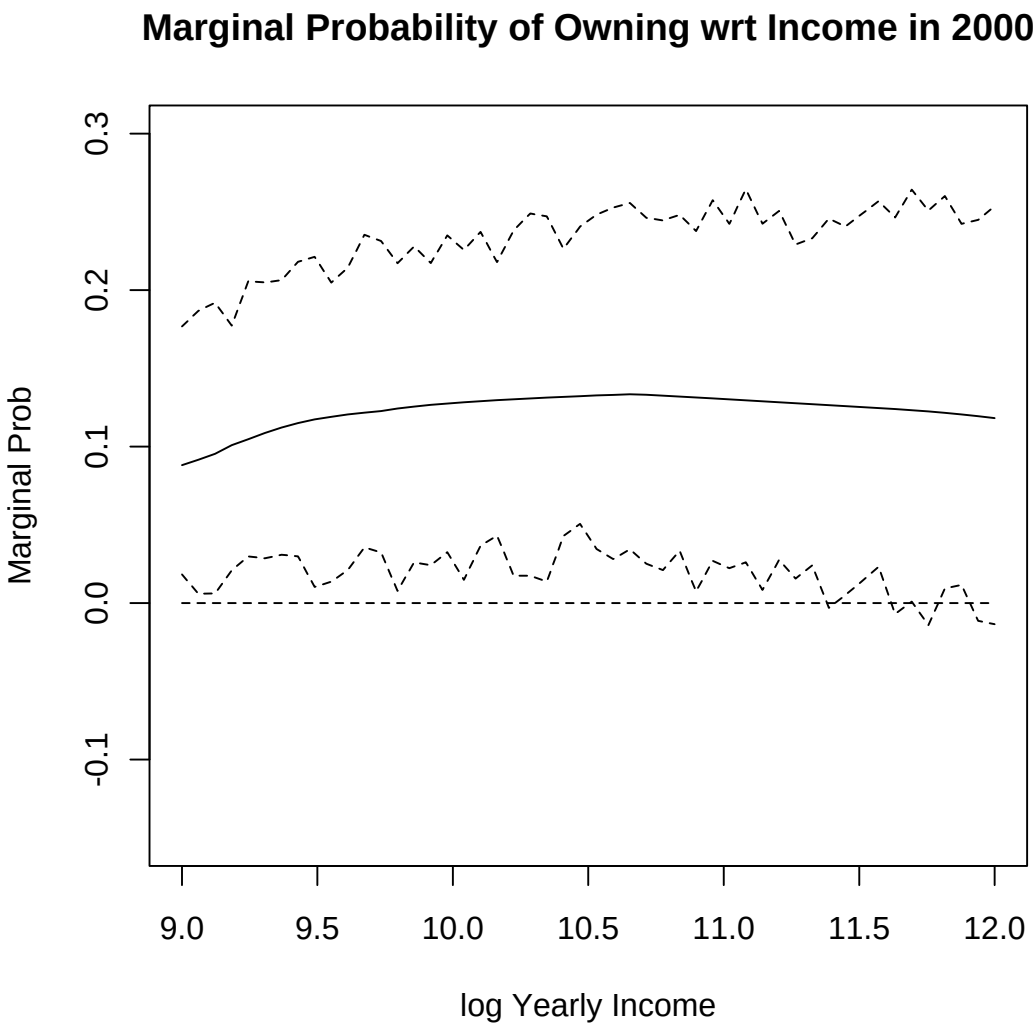
Marginal Probability of Owning wrt Income in 2000
Marginal Prob
0.3
0.2
0.1
0.0
-0.1
9.0
9.5
10.0
10.5
11.0
11.5
12.0
log Yearly Income

Figure 1

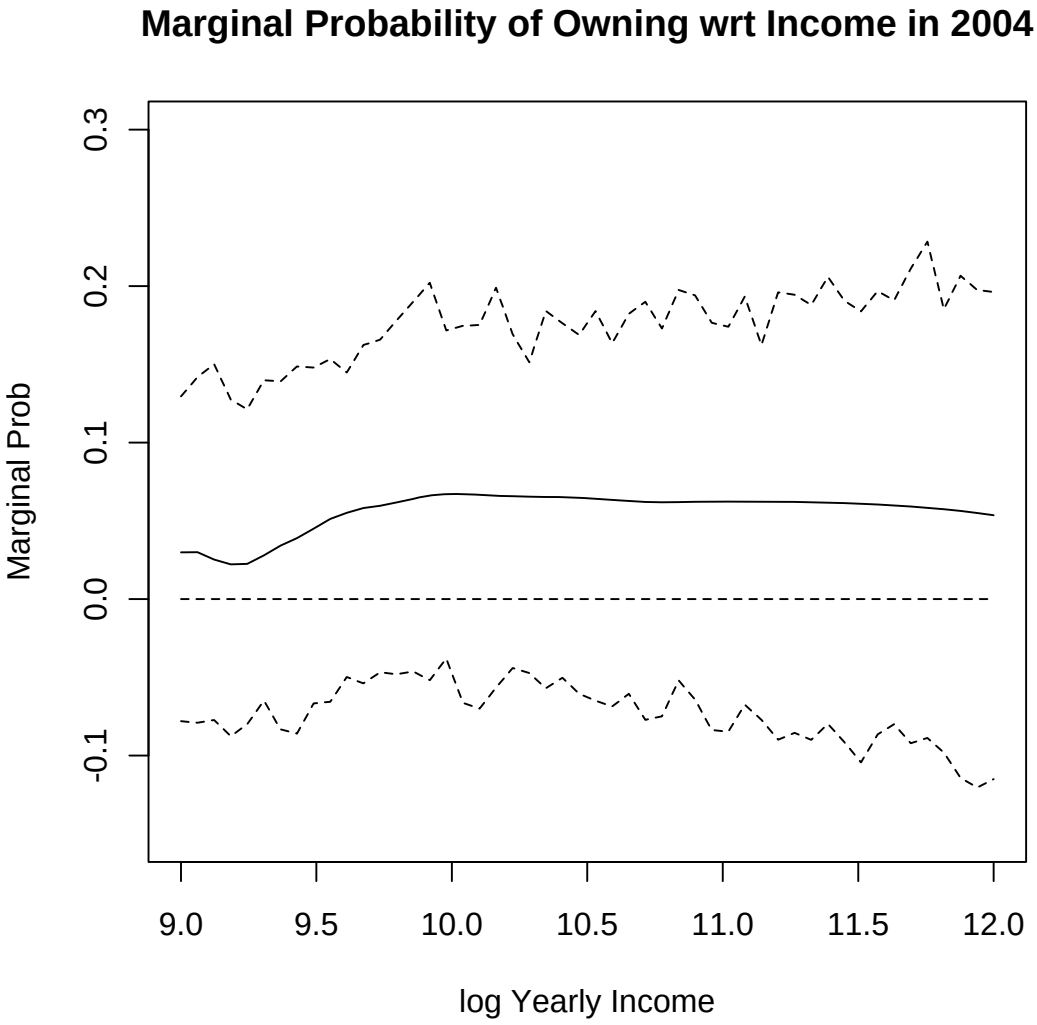
Marginal Probability of Owning wrt Income in 2004
Marginal Prob
0.3
0.2
0.1
0.0
-0.1
9.0
9.5
10.0
10.5
11.0
11.5
12.0
log Yearly Income

Figure 2

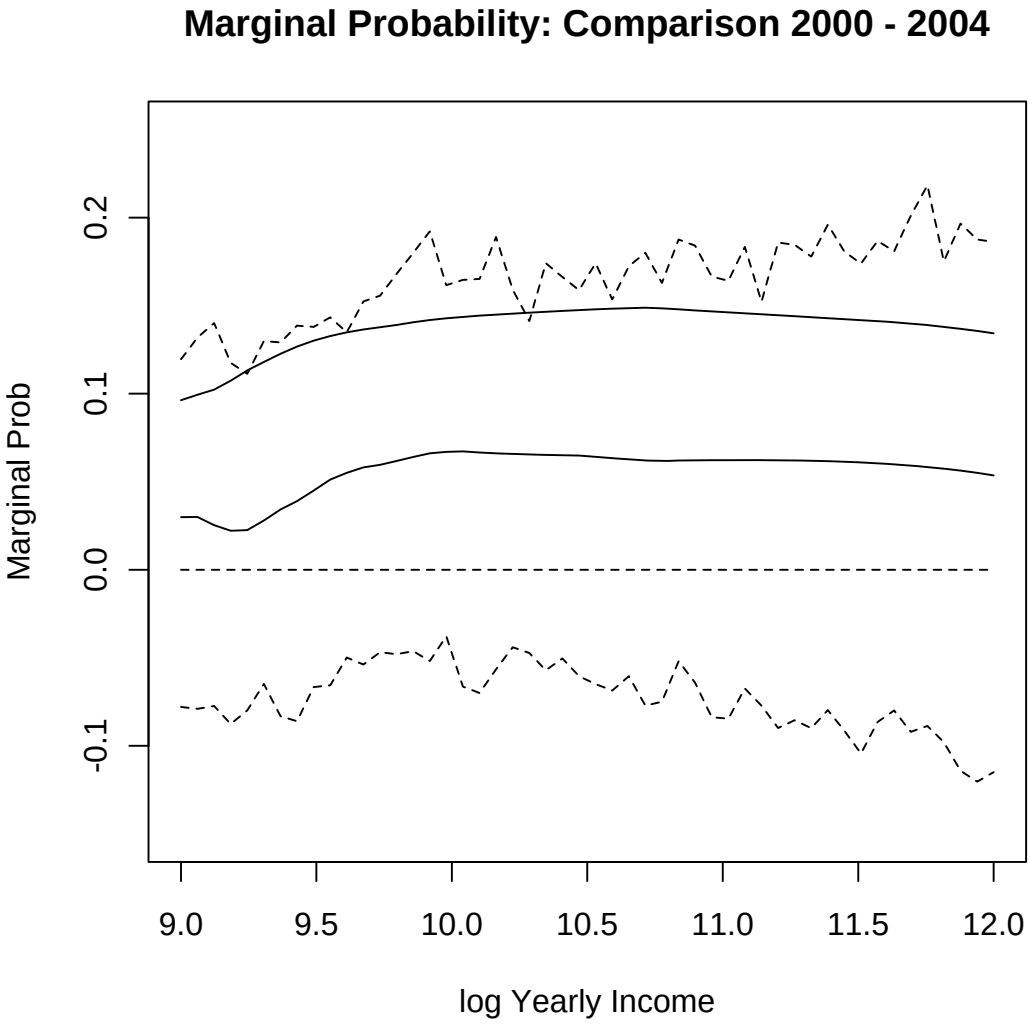
Marginal Probability: Comparison 2000 - 2004
Marginal Prob
0.2
0.1
0.0
-0.1
9.0
9.5
10.0
10.5
11.0
11.5
12.0
log Yearly Income

Figure 3