Inference in Panel Data Models under Attrition Caused by Unobservables

Debopam Bhattacharya*

Department of Economics

Dartmouth College.

First draft: April 15, 2004.
This draft: July 23, 2007.

Abstract

This paper concerns identification and estimation of a finite-dimensional parameter in a panel data-model under nonignorable sample attrition. Attrition can depend on second period variables which are unobserved for the attritors but an independent refreshment sample from the marginal distribution of the second period values is available. This paper shows that under a quasi-separability assumption, the model implies a set of conditional moment restrictions where the moments contain the attrition function as an unknown parameter. This formulation leads to (i) a simple proof of identification under strictly weaker conditions than those in the existing literature and, more importantly, (ii) a sieve-based root-n consistent estimate of the finite-dimensional parameter of interest. These methods are applicable to both linear and nonlinear panel data models with endogenous attrition and analogous methods are applicable to situations of endogenously missing data in a single cross-section. The theory is illustrated with a simulation exercise, using Current Population Survey data where a panel structure is introduced by the rotation group feature of the sampling process.

*JEL Code: C14, C23. Keywords: Attrition, conditional moments, identification, estimation. Address for correspondence: 327, Rockefeller Hall, Dartmouth College, Hanover, NH 03755. e-mail: debopam.bhattacharya@dartmouth.edu.

1 Introduction

Panel data models with "fixed effects" typically imply moment conditions of the form

$$0 = E\phi\left(y_1, y_2, x_1, x_2, \beta_0 | x_1, x_2\right), \text{ a.e. } x_1, x_2 \tag{1}$$

where $\phi(.)$ is known, the subscripts $1, 2$ correspond to the two time periods, y are the dependent variable, the x's are $d \times 1$ time-varying covariates and β_0 is the parameter of interest. Such moment conditions arise when the individual specific effects are eliminated either by differencing the data or by a clever conditioning method. For example, with strictly exogenous regressors, a linear individual effects model has

$$\phi\left(y_1, y_2, x_1, x_2, \beta\right) = y_2 - y_1 - \left(x_2 - x_1\right)' \beta,$$

the fixed-effects logit model has

$$\phi\left(y_1, y_2, x_1, x_2, \beta\right) = 1\left(y_1 + y_2 = 1\right)\left(y_1 - \frac{1}{1 + e^{(x_2 - x_1)'\beta}}\right),$$

fixed effects censored regression (Honore' (1992)) with censoring below at 0 has

$$\begin{aligned} \phi\left(y_1, y_2, x_1, x_2, \beta\right) &= v_{12}\left(\beta\right) - v_{21}\left(\beta\right) \\ v_{st}\left(\beta\right) &= \max\left\{y_s, \left(x_s - x_t\right)' \beta\right\} - \max\left\{0, \left(x_s - x_t\right)' \beta\right\}. \end{aligned}$$

When we have a random sample of individuals, either with no attrition or with sample attrition (i.e. x_2 and y_2 are not observed for a subset of the sample) that is "completely ignorable"[1], solving the sample analog of (1) yields consistent estimates of β_0. Moreover, when either the function $\phi(.)$ or its expectation is smooth in β (c.f. Pakes and Pollard (1989)), the estimators typically converge to normal distributions at the root-n rate. However, when sample attrition depends on the y's, solving the sample analog of (1) with only the survivors does not yield consistent estimates of the parameter of interest. If attrition depends only on the first period values of the variables (i.e. it is independent of the unobservable y_2 conditional on the observables y_1, x_1- traditionally called "ignorable attrition"), then one can

[1] Formally, if S is a dummy for whether an individual does not drop out, then a necessary and sufficient condition for complete ignorability of attrition in this model is that

$$E\left\{S\phi\left(y_1, y_2, x_1, x_2, \beta_0\right) | x_1, x_2\right\} = 0$$

which is implied by the condition $S \perp y_1, y_2 | x_1, x_2$.

estimate the probability of surviving conditional on (y_1, x_1) and either reweigh the data by inverse of these predicted probabilities or impute the missing observations to get a consistent estimate. If however, survival depends on y_2 after conditioning on (y_1, x_1), there is no way of identifying β_0 without using either external information or relatively strong untestable assumptions on the structure of the attrition process. The idea that attrition could depend on outcome variables of the second period is most easily motivated in a treatment effect context (see, e.g. Hausman and Wise (1979)) where people whose treatment effects are small are less enthusiastic about responding to the survey in the post-treatment period. A second example would be where one wants to estimate the effect of covariates (e.g. employer-provided health insurance) on job-mobility using panel data but suspects that individuals who change jobs and move are most likely to drop out of the sample.

Existing econometric methods which attempt to correct panel data estimators for attrition can be divided into two broad categories. The first is where one makes stronger assumptions on the attrition process but does not require additional data and the second is where some of these assumptions are relaxed but additional data are used. The first category includes Hausman and Wise (1981), Wooldridge (1999, 2002) and Das (2004). The second category includes Ridder (1990, 1992) and Nevo (2002, 2003) which rely on models of attrition that are assumed to be fully parametrically specified (e.g. assumptions A1-A3 in Nevo, 2003). The present paper belongs to the second strand of the literature in that it uses additional data in the form of refreshment samples but at the same time it relaxes strong parametric assumptions on the attrition process. When the main parameters of interest come from a linear model, Fitzgerald, Gottschalk and Moffitt (1998) discuss alternative approaches (including semiparametric ones) to estimation under attrition on observables and unobservables. Verbeek and Nijman (1992) consider testing attrition on unobservables under normality and linearity of the main model while Nicoletti (2006) considers testing under fully parametric settings but allowing for dynamic panel data models.[2]

The present paper extends the existing literature by combining the following features. It (i) allows for the estimands to come from nonlinear (such as probit) models, (ii) uses a flexible

[2] The econometric and applied literature on attrition in panel data is larger than the papers cited above. The Spring 1988 volume of the Journal of Human Resources, for instance, includes several other papers and a more complete citation of the literature. The above papers were cited only to illustrate the broad strands in the literature and to put the present paper in context.

specification of attrition, (iii) derives *both* identification and estimation results and (iv) does not use distributional assumptions for making selection-type corrections. It also adds to the existing body of work on semiparametric estimation based on combined samples (c.f. Moffitt and Ridder (forthcoming) for a survey of such methods). The two key requirements for the methods of the present paper to work are the availability of refreshment samples and a quasi-separability assumption (see (2) below) on the attrition process, which are explained below.

Hirano, Imbens, Ridder and Rubin (2001, henceforth HIRR) have recently addressed attrition on unobservables and have shown that under a quasi-separability restriction, the attrition function can be semiparametrically identified using refreshment data. However, they did not analyze the properties of the resulting estimator of the attrition function.[3] The main difficulty in doing this is that the infinite dimensional parameter (the attrition function) is not directly estimable in their approach. It is only implicitly defined through a set of integral equations which cannot be analytically solved to yield a closed-form expression for the attrition function. Therefore, the standard procedures of the semiparametrics literature (e.g. Newey and McFadden (1994), pages 2194-2215), which are based on a kernel or series-based estimate of the nonparametric component, cannot be used here to derive the asymptotic properties of the estimate of β_0.

The key observation in the present paper is that the integral equations implied by the model and discussed in HIRR are equivalent to a set of conditional moment restrictions where the moment functions contain the unknown attrition function and β_0 as unknown parameters. The moment interpretation leads to (i) establishing identification under *weaker* conditions than HIRR via a proof that is much simpler and significantly more elegant than the proof in HIRR, (ii) a sieve-based method of estimating both the attrition function and β_0 and derivation of asymptotic properties of these estimates by modifying the approach of Ai and Chen (2003) (henceforth, AC) to allow for the presence of different conditioning variables in the different moment conditions. I show that the key smoothness conditions for a $n^{1/4}$ rate of convergence for the attrition function and the $\sqrt{n}$-rate for β in this problem can be established by "local" versions of the moment-based identification proof above. However, unlike AC, I do not discuss the issue of efficiency here and postpone that to future research.[4]

[3] Their prior working paper version briefly discussed estimation under a fully parametric set-up when all variables were discrete.

[4] In personal communication, Chen has informed me of the existence of an unpublished note by Ai and

1.1 Sampling Set-up

The sampling set-up is as follows. We have two datasets- respectively called a primary and a refreshment sample. The primary sample is a random sample of size n of individuals drawn from the population in period 1 and followed into period 2. In period 2, only $n_1 < n$ of the original individuals respond. Let y_t denote the dependent variables, x_t denote the time-varying explanatory variables and v denote the time-invariant explanatory variables. Thus, we know the values of (y_{i1}, x_{i1}, v_i) for all n individuals and the values of (y_{2i}, x_{2i}) only for individuals $1, ...n_1$. Assume that we have a refreshment sample of size n_2, which is an independent random sample (i.e. drawn from the same *population* but not necessarily including the same individuals as the primary sample) from the marginal distribution of (y_2, x_2, v), viz. the population distribution of the second period values and are denoted by $\left(y_{j2}^*, x_{j2}^*, v_j^*\right)$, $j = 1, , , n_2$ $(n, n_1, n_2$ go to infinity at the same rate). Attrition is allowed to depend on all elements of (x_1, y_1, y_2, x_2, v) but the survival function has a quasi-separable form

$$\begin{aligned} & \Pr\left(S = 1 | z_1, z_2, v\right) \\ \equiv\ & g\left(\kappa\left(z_1, z_2, v\right)\right), \text{ say} \\ =\ & g\left(k_0\left(v\right) + k_1\left(z_1, v\right) + k_2\left(z_2, v\right)\right), \end{aligned} \tag{2}$$

where S is a dummy for whether an individual observed in the original panel survives in year 2, $(y_j, x_j) = z_j$, $j = 1, 2$. $g(.)$ is a known c.d.f. while the functions $k_0(.), k_1(., .), k_2(., .)$ are unknown and satisfy a location normalization. Such a structure can arise from a separable specification of the latent survival equation

$$S^* = k_0(v) + k_1(z_1, v) + k_2(z_2, v) - u \quad \text{with } S = 1(S^* > 0)$$

and conditional on (v, z_1, z_2), u follows a distribution with c.d.f. $g(.)$. Note that this structure does *not* imply that there is no interaction between z_1 and z_2 in determining $\Pr(S = 1|z_1, z_2, v)$; the separability holds only in the underlying latent equation.

Observe that (2) and (1) imply

$$E\left(\frac{\phi\left(x_1, y_1, y_2, x_2, \beta_0\right) S}{g\left(\kappa\left(z_1, z_2, v\right)\right)} | x_1, x_2\right) = E\left\{\phi\left(x_1, y_1, y_2, x_2, \beta\right) | x_1, x_2\right\} = 0, \text{ a.e. } x_1, x_2.$$

Chen that derives the efficient estimator in the situation with different conditioning variables in different moments.

Thus, the population problem becomes

$$E\left(\frac{\phi\left(x_1,y_1,y_2,x_2,\beta_0\right)S}{g\left(\kappa\left(z_1,z_2,v\right)\right)}|x_1,x_2\right)=0 \text{ a.e. } x_1,x_2 \tag{3}$$

which, under $\Pr\left(S=1|x_1,x_2\right)>0$ for all x_1,x_2,[5],[6] is equivalent to

$$E\left(\frac{\phi\left(x_1,y_1,y_2,x_2,\beta_0\right)}{g\left(\kappa\left(z_1,z_2,v\right)\right)}|S=1,x_1,x_2\right)=0 \text{ for all } x_1,x_2 \tag{4}$$

subject to

$$\int\frac{f\left(z_1,z_2,v|S=1\right)\Pr\left(S=1\right)}{g\left(\kappa\left(z_1,z_2,v\right)\right)}dz_2 = f_1\left(z_1,v\right) \tag{5}$$

$$\int\frac{f\left(z_1,z_2,v|S=1\right)\Pr\left(S=1\right)}{g\left(\kappa\left(z_1,z_2,v\right)\right)}dz_1 = f_2\left(z_2,v\right). \tag{6}$$

where $f\left(w|S=1\right)$ denotes the density of w conditional on $S=1$. The restrictions (5) and (6) equate the observed marginal densities $f_1\left(.,.\right)$ and $f_2\left(.,.\right)$ of $\left(z_1,v\right)$ and $\left(z_2,v\right)$ to the ones predicted by the model; the marginal in the RHS of (5) is observed in the original sample and that in the RHS of (6) is observed in the refreshment data. Using results from the theory of functional optimization, HIRR show that the restrictions (5) and (6) are satisfied uniquely (up to a location normalization) by the true functions $k_0\left(.\right),k_1\left(.,.\right),k_2\left(.,.\right)$ but do not propose an estimator. The main difficulty in doing so, as mentioned in the introduction, is that $\kappa\left(.\right)$ is only implicitly defined here through the integral equations in (5) and (6). A closed-form expression for $\kappa\left(.\right)$ is not in general obtainable from (5) and (6). Thus the standard procedures of the semiparametrics literature, which rely on a closed-form expression for the estimate of κ, are not directly applicable here. In section 2, I show that (i) (5) and (6) can be rewritten as conditional moment conditions containing β and $\kappa\left(.\right)$ as unknown parameters, (ii) the peculiar forms of these moment conditions permit the identification of

[5] This condition is not overtly strong- if $\Pr\left(S=1|x_1,x_2\right)=0$ for some $\left(x_1,x_2\right)$, then (3) has no information on either β or κ for those values of $\left(x_1,x_2\right)$ and one can simply discard them, i.e. our relevant moment condition is that for $A=\left\{\left(x_1,x_2\right):\Pr\left(S=1|x_1,x_2\right)>0\right\}$,

$$E\left(\frac{\phi\left(x_1,y_1,y_2,x_2,\beta\right)S}{g\left(\kappa\left(z_1,z_2,v\right)\right)}|x_1,x_2\right)=0 \text{ a.e. } x_1,x_2\in A.$$

[6] Since $\Pr\left(S=1|x_1,x_2\right)=\frac{f\left(x_1,x_2|S=1\right)}{f\left(x_1,x_2\right)}\Pr\left(S=1\right)$, a sufficient condition for $\Pr\left(S=1|x_1,x_2\right)>0$ for all x_1,x_2 is that the support of $\left(x_1,x_2\right)$ coincides with the support of $\left(x_1,x_2\right)$ conditional on $S=1$- which is also assumed in HIRR.

$k_0(.), k_1(.,.), k_2(.,.)$ from the primary and refreshment data without requiring any result from functional optimization theory and (iii) a modification of AC leads to a sieve-based $\sqrt{n}$-consistent estimate of β_0.

1.2 Refreshment Samples

Refreshment samples are prevalent in both the USA and the rest of the world-the German Socioeconomic Panel, the Russian Socioeconomic Transition Study, the Malaysian Family Life Survey etc. being prominent examples. A commonly analyzed US dataset with refreshment samples is the Current Population Survey (CPS) owing to its rotation group structure (see below for more discussion of the CPS). Another example is the Medical Expenditure Panel Survey (MEPS) which started in 1996 and employs an overlapping panel structure. It is not necessary that the refreshment sample comes from exactly the same data source as the primary sample. In principle, one can use any sample from the same population drawn in the later year. In particular, the census provides a large refreshment source of data if the census year corresponds to the latter year of the panel. If anything, the analysis in this paper should convince sampling agencies about the usefulness of refreshment samples!

1.3 Organization of the paper

The rest of the paper is organized as follows. Section 2 discusses the method of moment interpretation, provides a simple proof of identification and describes estimation by sieves. It also shows how the same methods apply to missing data in a single cross-section. Section 3 discusses consistency, the rate of convergence, asymptotic normality and consistent estimation of the covariance matrix together with a brief note on actual implementation in real datasets. Section 4 discusses two short extensions, section 5 contains results from the simulation exercise and section 6 concludes. Formal statements and proofs of the main propositions, some technical regularity conditions for the large sample results, especially asymptotic normality and some details about the simulation experiment are collected in the appendix.

2 Moment Interpretation, Identification and Estimation

The analysis starts by observing that the restrictions in (5) and (6) can be re-written as conditional moment conditions, involving the unknown functions $k_0(.)$, $k_1(.,.)$, $k_2(.,.)$. To see this, consider the following steps.

Let for $s = 0, 1$,

$$P(z_1, z_2, v, S = s) = f(z_1, z_2, v|S = s) \Pr(S = s)$$

where $f(z_1, z_2, v|S = s)$ denotes the joint density of (z_1, z_2, v) conditional on $S = s$. So, condition (6) is equivalent to

$$\begin{aligned} 0 &= \int \frac{P(z_1, z_2, v, S = 1)}{g(\kappa(z_1, z_2, v)) f_2(z_2, v)} dz_1 - 1 \\ &= \sum_{s=0,1} \int \frac{sP(z_1, z_2, v, S = s)}{g(\kappa(z_1, z_2, v)) f_2(z_2, v)} dz_1 - 1 \\ &= E\left\{\frac{S}{g(\kappa(z_1, z_2, v))}|z_2, v\right\} - 1. \end{aligned}$$

Thus the restrictions (5) and (6) reduce to

$$E\left\{\frac{S}{g(\kappa(z_1, z_2, v))} - 1|z_1, v\right\} = 0 \text{ for all } z_1, v \tag{7}$$

$$E\left\{\frac{S}{g(\kappa(z_1, z_2, v))} - 1|z_2, v\right\} = 0 \text{ for all } z_2, v. \tag{8}$$

These conditional moments can be transformed to unconditional moments which are estimable from the primary and refreshment samples, yielding a method of identifying and estimating the attrition function. For instance, condition (8) can be transformed to

$$E\left\{\frac{Sh(z_2, v)}{g(\kappa(z_1, z_2, v))}\right\} = E(h(z_2, v))$$

for any function $h(.,.)$. The left hand side can be estimated by $\frac{1}{n}\sum_{i=1}^{n} \frac{S_i h(z_{2i}, v_i)}{g(\kappa(z_{1i}, z_{2i}, v_i))}$, which is numerically equal to $\frac{1}{n}\sum_{i=1}^{n_1} \frac{h(z_{2i}, v_i)}{g(\kappa(z_{1i}, z_{2i}, v_i))}$ since S_i equals 1 for $i = 1, ...n_1$ and is zero otherwise. This last quantity can be computed from the primary sample alone. The right hand side can be estimated by $\frac{1}{n_2}\sum_{i=1}^{n_2} h(z_{2i}^*, v_i^*)$ using the refreshment sample. Doing this for a range of functions $h(.,.)$ (and similarly for (7)) will lead to an estimate of $\kappa(.)$ under conditions discussed below.

Remark 1 *Note that it is possible to start with the two moment conditions (7) and (8) without deriving them from the conditions (5) and (6). But ex ante it was far from obvious*[7] *that these are precisely the forms of the moment conditions, with* $g(\kappa)$ *in the denominator and conditioning on* (z_1, v) *and* (z_2, v) *one at a time, which can be used to identify and estimate the unknown attrition function from the incomplete data available. Observe that the more obvious moment condition*

$$E\{S - g(\kappa(z_1, z_2, v)) | z_1, z_2, v\} = 0 \text{ for all } z_1, z_2, v$$

is useless here since the joint distribution of (z_1, z_2, v) *is only observed for* $S = 1$.

Remark 2 *Further, the equivalence of (7), (8) and (5), (6) implies that the information content of these conditions are identical, i.e. we are not losing efficiency by concentrating on the moment formulation.*

2.1 Identification

In this subsection, I provide the main statement of identification. The proof of identification (in appendix) works under somewhat weaker conditions than HIRR and yet is significantly simpler and (in my view) much more elegant than their proof. In particular, the proof, unlike HIRR, does not require any complicated result from the theory of functional optimization and instead makes clever use of the conditional moment conditions derived above.[8]

First note from above that the population problem reduces to solving

$$\begin{aligned} E\left(\frac{\phi(x_1, y_1, y_2, x_2, \beta)}{g(\kappa(z_1, z_2, v))} | S = 1, x_1, x_2\right) &= 0 \text{ for all } x_1, x_2 \\ E\left\{\frac{S}{g(\kappa(z_1, z_2, v))} - 1 | z_1, v\right\} &= 0 \text{ for all } z_1, v \\ E\left\{\frac{S}{g(\kappa(z_1, z_2, v))} - 1 | z_2, v\right\} &= 0 \text{ for all } z_2, v. \end{aligned} \tag{9}$$

I first show that the last two moment conditions are satisfied only by the true attrition function under the quasi-separability assumptions and some regularity conditions. This in turn will imply identification of β, given the triangular nature of the problem, under the

[7] Perhaps, they become "obvious" after one has derived them!

[8] It is this elegant proof which is the main innovation in this part of the present paper, since most common cdf's are differentiable anyway.

assumption that β were identified in the original model (1), absent attrition. Below, let κ denote a generic function and κ_0 the true function. Identification amounts to showing that if any function κ satisfies the moment conditions (7) and (8), then $\kappa = \kappa_0$.

Proposition 1 *If (i) Conditional on each value v in the support of V, the support $\mathcal{Z}_1(v) \times \mathcal{Z}_2(v)$ of Z_1, Z_2 is not a lower-dimensional subspace of $\mathbb{R}^{2\times \dim(z)}$, (ii)*

$$\begin{aligned} E\left\{\frac{S}{g(k_0(v)+k_1(z_1,v)+k_2(z_2,v))}-1|z_1,v\right\} &= 0 \text{ for all } z_1,v \\ E\left\{\frac{S}{g(k_0(v)+k_1(z_1,v)+k_2(z_2,v))}-1|z_2,v\right\} &= 0 \text{ for all } z_2,v. \end{aligned} \tag{10}$$

(iii) $g(.)$ is strictly increasing, $\lim_{a\to-\infty} g(a) = 0 = 1 - \lim_{a\to\infty} g(a)$, (iv) for each v, there exists $\bar{z}_1(v) \in \mathcal{Z}_1(v)$ and $\bar{z}_2(v) \in \mathcal{Z}_2(v)$ such that $k_1(\bar{z}_1(v), v) = 0 = k_2(\bar{z}_2(v), v)$, then $\kappa = \kappa_0$ w.p. 1.

Remark 3 *Note that it is not required that $g(.)$ is differentiable and so the identification result is stronger than HIRR who assume all of (i)-(iv) in addition to differentiability of $g(.)$.*

Proof. See appendix. ■

2.2 Estimation

The starting point is the set of moment conditions (9), which imply that the true $\alpha_0 = (\beta_0, \kappa_0)$ uniquely minimizes (sets to 0) the positive semi-definite quadratic form[9]

$$Q(\alpha) = E_{x_1,x_2}\left\{m_0^2(\alpha,x_1,x_2)|S=1\right\} + E_{z_1,v}\left\{m_1^2(\alpha,z_1,v)\right\} + E_{z_2,v}\left\{m_2^2(\alpha,z_2,v)\right\} \tag{11}$$

where

$$\begin{aligned} m_0(\alpha,x_1,x_2) &= E\left\{\frac{\phi(y_1,y_2,x_1,x_2,\beta)}{g(\kappa(z_1,z_2,v))}|S=1,x_1,x_2\right\}, \\ m_1(\alpha,z_1,v) &= E\left\{\frac{S}{g(\kappa(z_1,z_2,v))}-1|z_1,v\right\}, \\ m_2(\alpha,z_2,v) &= E\left\{\frac{S}{g(\kappa(z_1,z_2,v))}-1|z_2,v\right\}. \end{aligned}$$

The estimation strategy is to minimize the sample analog of (11) over α. Since these sample analogs will be based on both the primary and refreshment samples, it is useful to write them out explicitly.

[9] Note that we are abstracting from efficiency considerations here, which is postponed to future research.

A consistent estimate of the first term in (11) is given by,

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\hat{E}\left\{\frac{\phi\left(y_1,y_2,x_1,x_2,\beta\right)}{g\left(\kappa\left(z_1,z_2,v\right)\right)}|S=1,x_{1j},x_{2j}\right\}\right)^2.$$

To estimate $\hat{E}\left\{\frac{\phi(y_1,y_2,x_1,x_2,\beta)}{g(\kappa(z_1,z_2,v))}|S=1,x_{1j},x_{2j}\right\}$ the idea is to use predicted values from a regression of $\frac{\phi(y_1,y_2,x_1,x_2,\beta)}{g(\kappa(z_1,z_2,v))}$ on a set of basis functions of x_1,x_2 where all the observations come from the subsample with $S=1$. Let $p_{0j}\left(x_1,x_2\right)$, $p_{1j}\left(z_1,v\right)$, $p_{2j}\left(z_2,v\right)$ be known "basis" functions whose number (k_n) grows slowly enough with the sample size. Recall that (z_1,z_2,v) denotes a typical observation in the primary sample and (z_2^*,v^*) denotes an observation in the auxiliary sample. Let

$$\begin{aligned}
p_0^{k_n}\left(x_{1i},x_{2i}\right) &= \left(p_{0j}\left(x_{1i},x_{2i}\right)\right)_{j=1,,,k_n}, P_0=\left(p_0^{k_n}\left(x_{1i},x_{2i}\right)\right)'_{i=1,2\ldots n_1}\\
p_1^{k_n}\left(z_{1i},v_i\right) &= \left(p_{1j}\left(z_{1i},v_i\right)\right)_{j=1,,,k_n}, P_1=\left(p_1^{k_n}\left(z_{1i},v_i\right)\right)'_{i=1,2\ldots n}\\
p_2^{k_n}\left(z_{2l},v_l\right) &= \left(p_{2j}\left(z_{2l},v_l\right)\right)_{j=1,,,k_n}, P_2^*=\left(p_2^{k_n}\left(z_{2l}^*,v_l^*\right)\right)'_{l=1,2\ldots n_2}.
\end{aligned}$$

The sample counterparts of m_0,m_1,m_2 are given by

$$\begin{aligned}
\hat{m}_0\left(\alpha,x_{1j},x_{2j}\right) &= \sum_{l=1}^{n_1}\frac{\phi\left(x_{1l},y_{1l},y_{2l},x_{2l},\beta\right)}{g\left(\kappa\left(z_{1l},z_{2l},v_l\right)\right)}p_0^{k_n}\left(x_{1l},x_{2l}\right)'\left(P_0'P_0\right)^{-1}p_0^{k_n}\left(x_{1j},x_{2j}\right)\\
\hat{m}_1\left(\alpha,z_{1j},v_j\right) &= \sum_{l=1}^{n}\left(\frac{S_l}{g\left(\kappa\left(z_{1l},z_{2l},v_l\right)\right)}-1\right)p_1^{k_n}\left(z_{1l},v_l\right)'\left(P_1'P_1\right)^{-1}p_1^{k_n}\left(z_{1j},v_j\right)\\
\hat{m}_2\left(\alpha,z_{2j}^*,v_j^*\right) &= \left\{\frac{1}{n}\sum_{l=1}^{n_1}\frac{S_l p_2^{k_n}\left(z_{2l},v_l\right)}{g\left(\kappa\left(z_{1l},z_{2l},v_l\right)\right)}-\frac{1}{n_2}\sum_{l=1}^{n_2}p_2^{k_n}\left(z_{2l}^*,v_l^*\right)\right\}'\times\left(\frac{P_2^{*\prime}P_2^*}{n_2}\right)^{-1}p_2^{k_n}\left(z_{2j}^*,v_j^*\right)
\end{aligned} \quad (12)[10]$$

Then the objective function that is minimized over α is the sample analog of (11):

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\hat{m}_0\left(\alpha,x_{1j},x_{2j}\right)^2+\frac{1}{n}\sum_{j=1}^{n}\hat{m}_1\left(\alpha,z_{1j},v_j\right)^2+\frac{1}{n_2}\sum_{j=1}^{n_2}\hat{m}_2\left(\alpha,z_{2j}^*,v_j^*\right)^2. \quad (13)$$

Since α contains the infinite dimensional $\kappa\left(.\right)$ it is convenient to carry out the minimization over a sieve space which "covers" the parameter space as the sample size grows to infinity. Thus, the estimate is obtained by minimizing (13) over $\Theta\times\mathbb{K}_n$, where $\mathbb{K}_n$ is an appropriately

[10] In the expression for $\hat{m}_2\left(\alpha,z,v\right)$, the two terms within $\{.\}$ are respectively the estimates of the population expectations of $p_2^{k_n}\left(z_2,v\right)$ obtained respectively from the (weighted) primary and the auxiliary samples and $\left(\frac{P_2^{*\prime}P_2^*}{n_2}\right)$ is a consistent estimate of the population expectation of the cross-product matrix of the $p_2^{k_n}\left(z_2,v\right)$'s. The difference with the expression for $\hat{m}_1\left(\alpha,z_1,v\right)$ arises because $\left(z_{2l},v_l\right)_{l=1,\ldots n_1}$ is not a random sample from the population distribution of (z_2,v).

defined sieve space. While one can use any standard set of basis functions such as splines, in my applications, I use power series.

Note also that the moment condition (12) is based on 2 different samples. One can write it in the standard form (to make it similar to Ai and Chen (2003)) as follows. Re-index observations by $k = 1, ...n + n_2$ with $n_2/n = \delta$, define the (deterministic) variable D_k equal to 1 if the kth observation comes from the primary sample and zero if it comes from the refreshment sample. Also, rewrite

$$P_2^{*\prime} P_2^* = \sum_{k=1}^{n+n_2} (1 - D_k) p_2^{k_n} (z_{2k}, v_k) p_2^{k_n} (z_{2k}, v_k)'$$

Then we have

$$\begin{aligned} \hat{m}_2 (\alpha, z_2, v) &= \sum_{k=1}^{n+n_2} \left\{ \delta D_k \frac{S_k p_2^{k_n} (z_{2k}, v_k)}{g (\kappa (z_{1k}, z_{2k}, v_k))} - (1 - D_k) p_2^{k_n} (z_{2k}, v_k) \right\}' \times (P_2^{*\prime} P_2^*)^{-1} p_2^{k_n} (z_2, v) \\ \frac{1}{n_2} \sum_{j=1}^{n_2} \hat{m}_2 \left(\alpha, z_{2j}^*, v_j^*\right)^2 &= \frac{1}{n + n_2} \sum_{k=1}^{n+n_2} \frac{1+\delta}{\delta} (1 - D_k) \hat{m}_2 (\alpha, z_{2k}, v_k)^2 = \frac{1}{n + n_2} \sum_{k=1}^{n+n_2} \bar{m}_2 (\alpha, z_{2k}, v_k)^2 \end{aligned}$$

where we have rewritten $\sqrt{\frac{1+\delta}{\delta} (1 - D_k)} \hat{m}_2 (\alpha, z_{2k}, v_k) = \bar{m}_2 (\alpha, z_{2k}, v_k)$. Similarly, the first 2 moment functions are

$$\begin{aligned} \hat{m}_0 (\alpha, x_1, x_2) &= \sum_{k=1}^{n+n_2} S_k D_k \frac{\phi (x_{1k}, y_{1k}, y_{2k}, x_{2k}, \beta)}{g (\kappa (z_{1k}, z_{2k}, v_k))} p_0^{k_n} (x_{1k}, x_{2k})' (P_0' P_0)^{-1} p_0^{k_n} (x_1, x_2) \\ \hat{m}_1 (\alpha, z_1, v) &= \sum_{k=1}^{n+n_2} D_k \left(\frac{S_k}{g (\kappa (z_{1k}, z_{2k}, v_k))} - 1 \right) p_1^{k_n} (z_{1k}, v_k)' (P_1' P_1)^{-1} p_1^{k_n} (z_1, v) \end{aligned}$$

This makes the problem have exactly the Ai and Chen structure with one i.i.d. sample of size $n + n_2$.

2.3 Missing Data in Single Cross-Section

The problem of missing outcome data in a single cross-section is a very similar problem and can be handled using the ideas developed above. The set-up is where one observes the joint occurrence of (y, x) for a subset of observations in a master dataset. For the other observations, y is missing. A refreshment sample in this case would be a random sample drawn from the same population where no y is missing. An example will be a single cross-section of the CPS as the masterdata with y being weekly earnings. Social security administration data on earnings could then be the refreshment sample. Suppose the models defining the parameter of interest and non-missingness are respectively

$$\begin{aligned} E\left[\phi\left(y,x,\beta_0\right)|x\right] &= 0 \text{ a.e. } x \\ \Pr\left(S=1|y,x\right) &= m\left(y,x\right). \end{aligned}$$

One observes the distribution of $(y, x|S = 1)$ and the joint distribution of (S, x) from the primary sample and the marginal of y from the refreshment data. Then under a quasi-separability condition

$$m\left(y,x\right)=g\left(k_0+k_1\left(x\right)+k_2\left(y\right)\right),$$

one can point-identify the conditional probability of missing outcomes and consequently derive a $\sqrt{n}$-consistent and asymptotically normal estimate of β_0. The key moment conditions analogous to (9) are

$$\begin{aligned} E\left[\frac{\phi\left(y,x,\beta_0\right)}{g\left(k_0+k_1\left(x\right)+k_2\left(y\right)\right)}|S=1,x\right] &= 0 \text{ a.e. } x \\ E\left(\frac{S}{g\left(k_0+k_1\left(x\right)+k_2\left(y\right)\right)}-1|x\right) &= 0 \text{ a.e. } x \\ E\left(\frac{S}{g\left(k_0+k_1\left(x\right)+k_2\left(y\right)\right)}-1|y\right) &= 0 \text{ a.e. } y. \end{aligned}$$

Note that a more realistic scenario is where the individual cross-sections are subject to nonignorable nonresponse and there is also attrition across the panel. It would be an interesting future project to develop data combination based methods to handle such problems. But that is not within the scope of the present paper.

3 Asymptotic Properties

In this section, I discuss consistency, rate of convergence and asymptotic distribution of the estimates. The approach taken here is to verify that the sufficient conditions of AC hold in the present problem and invoke their theorems to establish the results. I use this section to establish the key sufficient conditions and postpone discussion of regularity conditions, formal statements and proofs to the appendix. Even the substantive conditions (vis-a-vis regularity conditions) in AC's original paper are somewhat abstract, especially those related to the rate of convergence and asymptotic normality. The purpose of this section is to specialize the general and somewhat abstract assumptions of AC to the present problem. So this section is not technically "self-contained" in that it repeatedly refers to specific assumptions and results in AC and so is probably best read in conjunction with sections 3 and 4 of the AC paper.

The last subsection is a note on implementing the estimation of the main parameters and estimation of the corresponding asymptotic variances. It aims to provide necessary guidelines to practitioners who want to use these methods on real data.

3.1 Consistency

The proof of consistency will be analogous to AC lemma 3.1 which, in turn, appeals to the proof in Newey and Powell (2003). Specification of the parameter space, $\mathbb{K}$ and construction of the sieve, $\mathbb{K}_n$ will be similar to them. Since their approach requires that the conditioning variables have compact support with a density bounded away from 0 on this compact support (e.g. assumption 3.1 (ii) and (iii) on page 1803), I require that the supports of all the y, v and x variables are compact and have densities that are bounded away from 0.[11] The definition of the parameter space and the consistency norm are as follows (the specifications here are somewhat similar to AC, example 2.2). Suppose that the support of z_1, z_2 is $\mathcal{Z} \subset \mathbb{R}^{d_z}$ and that of v is $\mathcal{V} \subset \mathbb{R}^{d_v}$, $\mathcal{Z}$, $\mathcal{V}$ compact with densities that are bounded away from 0. The parameter space for the infinite dimensional parameters is a Holder ball $\mathbb{K}$, containing those (k_0, k_1, k_2) which satisfy

$$\begin{aligned}
\sup_{v\in\mathcal{V}} |k_0(v)| + \sup_{a_1+a_2+\dots a_{\dim v}\leq[\gamma_0]} \sup_{v\neq v'} \frac{|\nabla^a k_0(v) - \nabla^a k_0(v')|}{\|v-v'\|_E^{\gamma_0-[\gamma_0]}} &\leq c_0 < \infty, \\
\sup_{z_1\in\mathcal{Z}, v\in\mathcal{V}} |k_1(z_1, v)| + \sup_{a_1+a_2+\dots a_{\dim z_1+\dim v}\leq[\gamma]} \sup_{(z_1,v)\neq(z_1',v')} \frac{|\nabla^a k_1(z_1, v) - \nabla^a k_1(z_1', v')|}{\|(z_1, v) - (z_1', v')\|_E^{\gamma-[\gamma]}} &\leq c_1 < \infty, \\
\sup_{z_2\in\mathcal{Z}, v\in\mathcal{V}} |k_2(z_2, v)| + \sup_{a_1+a_2+\dots a_{\dim z_2+\dim v}\leq[\gamma]} \sup_{(z_2,v)\neq(z_2',v')} \frac{|\nabla^a k_2(z_2, v) - \nabla^a k_2(z_2', v')|}{\|(z_2, v) - (z_2', v')\|_E^{\gamma-[\gamma]}} &\leq c_2 < \infty \qquad (14)
\end{aligned}$$

where $\frac{\dim v}{2} < \gamma_0$, $\frac{2\dim z+\dim v}{2} < \gamma$, $[.]$ denotes the greatest integer function, $\|.\|_E$ denotes the Euclidean metric and for a function $k(.)$ of a $\dim v$ dimensional vector v, $\nabla^a k_0(v)$ denotes

[11] It is possible to achieve consistency without requiring the compact support assumptions as in Newey and Powell (2003). In this approach, one allows the unknown function to be nonparametric in the "middle" of its support but parametric in the "tails". However, I am not aware of a corresponding theory for deriving the rates of convergence for sieve estimators without bounded support assumptions on the *conditioning* variables.

an $(a_1 + a_2 + ...a_{\dim v})$th order partial derivative of the function $k_0(.)$.

For consistency, I shall use the norm $\|.\|_s$, defined as

$$\|\alpha\|_s = \sup_{v\in\mathcal{V}} |k_0(v)| + \sup_{z_1\in\mathcal{Z},v\in\mathcal{V}} |k_1(z_1,v)| + \sup_{z_2\in\mathcal{Z},v\in\mathcal{V}} |k_2(z_2,v)| + \sqrt{\beta'\beta}.$$

The functions in κ are approximated by the power series (with number of terms increasing slowly enough with the sample size)

$$\bar{k}_0(v) = \sum_{j=0}^{J_n} \delta_{0j} v^{[j]}, \;\; \bar{k}_1(z_1,v) = \sum_{l=1}^{J_n}\sum_{j=0}^{J_n} \delta_{1jl} v^{[j]} z_1^{[l]}, \;\; \bar{k}_2(z_2,v) = \sum_{l=1}^{J_n}\sum_{j=0}^{J_n} \delta_{2jl} v^{[j]} z_2^{[l]},$$

where the coefficients satisfy (14) above and the notation $v^{[j]}$ denotes products of elements of the v-vector raised to exponents that sum to j. Location normalization is automatically imposed since $\bar{k}_1(0,v) = 0$, $\bar{k}_2(0,v) = 0$ for all v.

Thus the estimation problem amounts to minimizing (13) subject to (14) and $\beta'\beta \leq B_\beta$. I shall assume that no interactions exist within the different components of v, of z_1 and of z_2. So number of "unknowns" to solve equals $1 + d_\beta + J_n \times d_v + 2J_n^2 \times d_{z_1} \times (1 + d_v) = d_\beta + K_{1n}$. Number of moment conditions we have from (12) will be denoted by K_n.

The formal assumptions, proof of a necessary Lipschitz property and statement for consistency are in the appendix section 7.2. Under these assumptions, one can invoke theorem 4.1 of Newey and Powell (2003) to show that $\|\hat{\alpha} - \alpha_0\|_s \xrightarrow{P} 0$.

3.2 Rate of convergence

This subsection will use the methods of section 3 in AC to define a norm $\|.\|$ such that $\|\hat{\alpha} - \alpha_0\| = o_p\left(n^{-1/4}\right)$. This rate turns out to be sufficient to guarantee (under additional regularity conditions) asymptotic normality of the estimate of β. Below, lemma 1 will establish that the key smoothness condition (15) required to obtain the $n^{1/4}$ rate holds for this problem. This condition is roughly that the objective function can be expressed as a quadratic in the true parameter locally around the true value.

The relevant norm is defined as

$$\begin{aligned} & \|\alpha_1 - \alpha_2\|^2 \\ = \; & E\left\{\left\{\frac{dm_0(\alpha_0, x_1, x_2)}{d\alpha}(\alpha_1 - \alpha_2)\right\}^2 |S = 1\right\} \\ & + E\left\{\frac{dm_1(\alpha_0, z_1, v)}{d\alpha}(\alpha_1 - \alpha_2)\right\}^2 + E\left\{\frac{dm_2(\alpha_0, z_2, v)}{d\alpha}(\alpha_1 - \alpha_2)\right\}^2, \end{aligned}$$

where $\frac{dm(\alpha_0)}{d\alpha}$ denotes a pathwise derivative. Writing out, using u to denote (z_1, z_2, v), we have

$$\begin{aligned}
&\frac{dm_0\left(\alpha_0, x_1, x_2\right)}{d\alpha}\left(\alpha_1-\alpha_2\right) \\
= \;& E\left\{\frac{\nabla_\beta \phi\left(z_1, z_2, \beta_0\right)'\left(\beta_1-\beta_2\right)}{g\left(\kappa_0(u)\right)}-\frac{g'\left(\kappa_0(u)\right) \phi\left(u, \beta_0\right)}{g^2\left(\kappa_0(u)\right)}\left(\kappa_1(u)-\kappa_2(u)\right) | x_1, x_2, S=1\right\}
\end{aligned}$$

$$\begin{aligned}
\frac{dm_1\left(\alpha_0, z_1, v\right)}{d\alpha}\left(\alpha_1-\alpha_2\right) &= -E\left\{\frac{g'\left(\kappa_0(u)\right)}{g\left(\kappa_0(u)\right)}\left(\kappa_1(u)-\kappa_2(u)\right) | z_1, v\right\} \\
\frac{dm_2\left(\alpha_0, z_2, v\right)}{d\alpha}\left(\alpha_1-\alpha_2\right) &= -E\left\{\frac{g'\left(\kappa_0(u)\right)}{g\left(\kappa_0(u)\right)}\left(\kappa_1(u)-\kappa_2(u)\right) | z_2, v\right\}.
\end{aligned}$$

I will verify conditions 3.6 (iii) and (iv) and conditions 3.9 (i) and (ii) of AC, since they pertain to the specification of the moment functions. The other conditions relate to the properties of the sieves which do not depend on the specific moment functions we have here.[12]

First, I verify condition 3.9 (ii) of AC, viz. for some constants $c_1, c_2 > 0$, we have that for all $\alpha \in \mathbb{K}_n$ satisfying $\left\|\alpha-\alpha_0\right\|_s = o(1)$,

$$c_1 E\{Q(\alpha)\} \leq \left\|\alpha-\alpha_0\right\|^2 \leq c_2 E\{Q(\alpha)\}. \tag{15}$$

This will be established using the two facts that (A)$\|\alpha-\alpha_0\| > 0$ if $\alpha \neq \alpha_0$ and (B) higher order derivatives of the objective function are bounded in an appropriate sense. Then, (B) will imply that $E\{Q(\alpha)\} = \left\|\alpha-\alpha_0\right\|^2 + o\left(\left\|\alpha-\alpha_0\right\|_s^2\right)$ whence (A) will imply the result.

(A) is established below, using the following lemma. This lemma will be used again in showing the requisite smoothness properties for establishing the $\sqrt{n}$-rate for β in the next section. The proof of this lemma can be viewed as a "local" analog of the proof of identification in section 2 of the paper, except that differentiability of $g(.)$ is assumed. The idea of the lemma is as follows. Let $w_0(.), w_1(.,.)$ and $w_2(.,.)$ be arbitrary functions. Suppose that for each fixed v, the support of (z_1, z_2) is given by $\mathcal{Z}_1(v) \times \mathcal{Z}_2(v)$. Let

$$\tilde{H}\left(z_1, z_2, v\right) = \frac{g'\left(\kappa_0\left(z_1, z_2, v\right)\right)}{g\left(\kappa_0\left(z_1, z_2, v\right)\right)}\left\{w_0(v) + w_1\left(z_1, v\right) + w_2\left(z_2, v\right)\right\}.$$

We want to show that if for each fixed v, $E\left(\tilde{H}(z_1, z_2, v) | z_1, v\right) = 0$ for all $z_1 \in \mathcal{Z}_1(v)$ and $E\left(\tilde{H}(z_1, z_2, v) | z_2, v\right) = 0$ for all $z_2 \in \mathcal{Z}_2(v)$, then $w_1(z_1, v) = 0 = w_2(z_2, v)$ a.e. on $\mathcal{Z}_1(v) \times \mathcal{Z}_2(v)$. The formal statement is as follows.

[12] Except that the degree of smoothness required for bounding the bracketing numbers and therefore the precision of the approximation of the true functions by the sieves depends on the dimensions of the variables

Lemma 1 *If for each fixed v, (i) the support $\mathcal{Z}_1(v) \times \mathcal{Z}_2(v)$ of (z_1, z_2) is not a lower dimensional subspace of $\mathbb{R}^{2\times\dim(z)}$, (ii) $E\left(\tilde{H}(z_1, z_2, v)|z_1, v\right) = 0$ for all $z_1 \in \mathcal{Z}_1(v)$ and $E\left(\tilde{H}(z_1, z_2, v)|z_2, v\right) = 0$ for all $z_2 \in \mathcal{Z}_2(v)$, (iii) $g(.)$ is differentiable with $g'(\kappa_0(z_1, z_2, v)) > 0$ on $\mathcal{Z}_1(v) \times \mathcal{Z}_2(v)$, (iv) there exists $\bar{z}_1(v) \in \mathcal{Z}_1(v)$ and $\bar{z}_2(v) \in \mathcal{Z}_2(v)$ such that $w_1(\bar{z}_1(v), v) = 0 = w_2(\bar{z}_2(v), v)$, then $w_1(z_1, v) = 0 = w_2(z_2, v)$ a.e. on $\mathcal{Z}_1(v) \times \mathcal{Z}_2(v)$. Moreover, $w_0(v)$ is 0 a.e. v.*

Proof. See appendix ■

Remark 4 *The separation between z_1 and z_2 is important here. To see that, consider the following example. Let Z_1, Z_2, V be independent normals with 0 mean. Let $w(z_1, z_2, v) = z_1 z_2$. Then*

$$\begin{aligned} E(w(z_1, z_2, v)|z_1, v) &= z_1 E(z_2|v) = z_1 E(z_2) = 0 \\ E(w(z_1, z_2, v)|z_2, v) &= z_2 E(z_1|v) = z_2 E(z_1) = 0 \end{aligned}$$

but $z_1 z_2$ is not 0 with probability 1.

Remark 5 *Note that the proof of this lemma is somewhat similar to the proof of proposition 1, above. In fact, one can view the key conditions for the $n^{1/4}$ rate of convergence (i.e. (A) above) and the $\sqrt{n}$-rate for β (i.e. (17) below) as local versions of the original identification condition.*

(A) now follows from lemma 1 since

$$\begin{aligned} & \|\alpha - \alpha_0\|^2 \\ \geq\ & E\left\{\frac{dm_1(\kappa_0, z_1, v)}{d\kappa}(\kappa - \kappa_0)\right\}^2 + E\left\{\frac{dm_2(\kappa_0, z_2, v)}{d\kappa}(\kappa - \kappa_0)\right\}^2 \\ =\ & E_{z_1,v}\left\{\left(E\left\{\frac{g'(\kappa_0(z_1, z_2, v))}{g(\kappa_0(z_1, z_2, v))}(\kappa(z_1, z_2, v) - \kappa_0(z_1, z_2, v))|z_1, v\right\}\right)^2\right\} \\ & + E_{z_2,v}\left\{\left(E\left\{\frac{g'(\kappa_0(z_1, z_2, v))}{g(\kappa_0(z_1, z_2, v))}(\kappa(z_1, z_2, v) - \kappa_0(z_1, z_2, v))|z_2, v\right\}\right)^2\right\} \\ >\ & 0 \text{ if } \kappa \neq \kappa_0 \end{aligned}$$

where the last inequality is a direct consequence of the lemma.

One can verify (B) above term by term using steps analogous to AC, example 2.2. I go through the argument for $m_0(\alpha(, x_1, x_2)$; the other terms are similar but simpler. By a first-order Taylor expansion,

$$
\begin{aligned}
& m_0(\alpha, x_1, x_2) \\
= \; & E\left\{\frac{\phi(w,\beta)}{g(\kappa(z_1,z_2,v))}|x_1,x_2,S=1\right\} \\
= \; & E\left\{\frac{\nabla_\beta \phi(w,\bar{\beta})'}{g(\bar{\kappa}(w))}(\beta-\beta_0)-\frac{g'(\bar{\kappa}(w))\,\phi(w,\bar{\beta})}{g^2(\bar{\kappa}(w))}(\kappa(w)-\kappa_0(w))\,|x_1,x_2,S=1\right\}
\end{aligned}
$$

for some intermediate values $\bar{\beta}$ and $\bar{\kappa}$. Therefore,

$$
E\left[m_0(\alpha,x_1,x_2)^2|S=1\right]=(\beta-\beta_0)'E\left(D_w(\bar{\alpha},x_1,x_2)'D_w(\bar{\alpha},x_1,x_2)\,|S=1\right)(\beta-\beta_0)
$$

where

$$
D_w(\bar{\alpha},x_1,x_2)=E_w\left\{\frac{\nabla_\beta \phi(u,\bar{\beta})}{g(\bar{\kappa}(u))}-\frac{g'(\bar{\kappa}(u))\,\phi(w,\bar{\beta})}{g^2(\bar{\kappa}(u))}\mathbf{w}(u)\,|x_1,x_2,S=1\right\}
$$

and $\mathbf{w}(u)$ satisfies $(\beta-\beta_0)'\mathbf{w}(u)=\kappa(u)-\kappa_0(u)$. Also,

$$
\begin{aligned}
& E\left\{\left(\frac{dm_0(\alpha_0,x_1,x_2)}{d\alpha}(\alpha-\alpha_0)\right)'\left(\frac{dm_0(\alpha_0,x_1,x_2)}{d\alpha}(\alpha-\alpha_0)\right)|S=1\right\} \\
= \; & (\beta-\beta_0)'E\left(D_w(\alpha_0,x_1,x_2)'D_w(\alpha_0,x_1,x_2)\,|S=1\right)(\beta-\beta_0).
\end{aligned}
$$

Using a first-order Taylor series expansion of $D_w(\bar{\alpha},x_1,x_2)$ around $D_w(\alpha_0,x_1,x_2)$ and using a set of uniform boundedness assumptions on the relevant first derivatives in that expansion, it will follow that

$$
\begin{aligned}
& E\left[m_0(\alpha,x_1,x_2)^2|S=1\right] \\
= \; & (\beta-\beta_0)'E\left(D_w(\bar{\alpha},x_1,x_2)'D_w(\bar{\alpha},x_1,x_2)\,|S=1\right)(\beta-\beta_0) \\
= \; & (\beta-\beta_0)'E\left(D_w(\alpha_0,x_1,x_2)'D_w(\alpha_0,x_1,x_2)\,|S=1\right)(\beta-\beta_0)+o\left(\|\alpha-\alpha_0\|_s^2\right) \\
= \; & \left\{\text{1st term of } \|\alpha-\alpha_0\|^2\right\}+o\left(\|\alpha-\alpha_0\|_s^2\right).
\end{aligned}
$$

Using similar arguments for $m_1(.)$ and $m_2(.)$, it follows that one can approximate $Q(\alpha)$ locally around α_0 by $\|\alpha-\alpha_0\|^2$ which is the essential step in deriving the rate of convergence and the asymptotic normality below.

The formal assumptions and statement for the rate of convergence are in the appendix. These assumptions guarantee that hypotheses of theorem 3.1 of AC hold here. Invoking that theorem, it follows that

$$
\|\hat{\alpha}-\alpha_0\|=o_p\left(n^{-1/4}\right).
$$

3.3 Asymptotic normality

Asymptotic normality of the estimate of β follows from the fact that it can be written as an appropriately defined inner product between the full parameter α and the so-called Riesz representor. This inner product essentially turns the estimate of β into an average of nonparametric estimates, asymptotically speaking, whence the normality follows. Clearly, the influence function for the estimate of β will involve this Riesz representor. Section 4 of AC outlines this approach which is due to Shen (1997).

In this subsection, I (i) show that the key smoothness condition for the Riesz representation (see (17) below) holds for the current problem, (ii) derive the form of the Riesz representor as a projection and (iii) derive the influence functions for the estimate of β. The additional regularity conditions, formal statements and proofs appear in the appendix.

Let $\bar{\mathcal{V}} = \mathbb{R}^{d_\beta} \times \bar{\mathcal{W}}$ denote the closure (w.r.t. $\|.\|$) of the linear span of the space $\mathcal{A}-\{\alpha_0\}$ where $\mathcal{A} = \Theta \times \mathbb{K}$ and $\bar{\mathcal{W}} \equiv \bar{\mathbb{K}} - \{\kappa_0\}$. Then $\left(\bar{\mathcal{V}}, \|.\|\right)$ is a Hilbert space with the inner product corresponding to the norm $\|.\|$, defined above. For each component β_j, $j = 1, 2, ...d_\beta$, let $w_j^* \in \bar{\mathcal{W}}$ minimize (over $w_j \in \bar{\mathcal{W}}$)

$$\begin{aligned}
&E_{x_1,x_2}\left\{\left\{E\left(\frac{\frac{\partial}{\partial\beta_j}\phi\left(u,\beta_0\right)}{g\left(\kappa_0\left(u\right)\right)} - \frac{g'\left(\kappa_0\left(u\right)\right)\phi\left(u,\beta_0\right)}{g^2\left(\kappa_0\left(u\right)\right)}w_j\left(u\right)|x_1,x_2,S=1\right)\right\}^2|S=1\right\} \\
&+E_{z_1,v}\left\{\left(E\left\{\frac{g'\left(\kappa_0\left(u\right)\right)}{g\left(\kappa_0\left(u\right)\right)}w_j\left(u\right)|z_1,v\right\}\right)^2\right\} \\
&+E_{z_2,v}\left\{\left(E\left\{\frac{g'\left(\kappa_0\left(u\right)\right)}{g\left(\kappa_0\left(u\right)\right)}w_j\left(u\right)|z_2,v\right\}\right)^2\right\}. \qquad (16)
\end{aligned}$$

Then, for $f\left(\alpha\right) = \lambda'\beta$, a sufficient smoothness condition for $\sqrt{n}$-normality of $\lambda'\left(\hat{\beta} - \beta_0\right)$ is that

$$\sup_{0\neq\alpha-\alpha_0\in\bar{\mathcal{V}}} \frac{\left|f\left(\alpha\right) - f\left(\alpha_0\right)\right|^2}{\left\|\alpha-\alpha_0\right\|^2} = \lambda'\Delta^{*-1}\lambda < \infty \qquad (17)$$

where

$$\begin{aligned}
&\Delta^* \\
= \; & E_{x_1,x_2}\left\{\left(\begin{array}{c} E\left(\frac{\nabla\phi(u,\beta_0)}{g(\kappa_0(u))} - \frac{\phi(u,\beta_0)g'(\kappa_0(u))}{g^2(\kappa_0(u))} w^*(u)\,|x_1,x_2,S=1\right) \\ \times E\left(\frac{\nabla\phi(u,\beta_0)}{g(\kappa_0(u))} - \frac{\phi(u,\beta_0)g'(\kappa_0(u))}{g^2(\kappa_0(u))} w^*(u)\,|x_1,x_2,S=1\right)' \end{array}\right)|S=1\right\} \\
& + E_{z_1,v}\left\{E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))} w^*(u)\,|z_1,v\right\} E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))} w^*(u)\,|z_1,v\right\}'\right\} \\
& + E_{z_2,v}\left\{\left(E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))} w^*(u)\,|z_2,v\right\}\right) E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))} w^*(u)\,|z_2,v\right\}'\right\}
\end{aligned}$$

where $\nabla\phi(u,\beta) = \left(\frac{\partial}{\partial\beta_j}\phi(u,\beta)\right)_{j=1,2,\ldots d_\beta}$ and $w^*(u) = \left(w_j^*(u)\right)_{j=1,2,\ldots d_\beta}$.

Now, I shall show (17). In what follows, I will suppress the arguments of the functions to avoid notational clutter, and use a subscript 0 to indicate that the functions are evaluated at the true values of the parameters. First note that by Cauchy-Schwartz,

$$|f(\alpha) - f(\alpha_0)|^2 = |\lambda'(\beta-\beta_0)|^2 \leq \lambda'\lambda\,\|\beta-\beta_0\|^2.$$

Next,

$$\begin{aligned}
&\frac{\|\beta-\beta_0\|^2}{\|\alpha-\alpha_0\|^2} \\
= \; & \frac{\|\beta-\beta_0\|^2}{\left\{\begin{array}{c} E_{x_1,x_2}\left\{\left(E\left\{\frac{\nabla_\beta\phi(\beta_0)'}{g_0}(\beta-\beta_0) - \frac{g_0'\phi_0}{g_0^2}(\kappa-\kappa_0)\,|x_1,x_2,S=1\right\}\right)^2 |S=1\right\} \\ + E_{z_1,v}\left\{E\left\{\frac{g_0'}{g_0}(\kappa-\kappa_0)\,|z_1,v\right\}\right\}^2 + E_{z_2,v}\left\{E\left\{\frac{g_0'}{g_0}(\kappa-\kappa_0)\,|z_2,v\right\}\right\}^2 \end{array}\right\}}
\end{aligned}$$

Note that the last two terms in the denominator do not depend on β and are strictly positive for $\kappa \neq \kappa_0$ (from Lemma 1). Therefore a sufficient condition for smoothness is that

$$\begin{aligned}
\infty \; > \; & \sup_{\beta\neq\beta_0} \frac{\|\beta-\beta_0\|^2}{E_{x_1,x_2}\left[\left\{E\left\{\frac{\nabla_\beta\phi(\beta_0)}{g_0}|x_1,x_2,S=1\right\}'(\beta-\beta_0)\right\}^2 |S=1\right]} \\
= \; & \sup_{\beta\neq\beta_0} \frac{\|\beta-\beta_0\|^2}{(\beta-\beta_0)' E_{x_1,x_2}\left\{\frac{E\left\{\nabla_\beta\phi(\beta_0)|x_1,x_2\right\}E\left\{\nabla_\beta\phi(\beta_0)|x_1,x_2\right\}'}{(\Pr(S=1|x_1,x_2))^2}|S=1\right\}(\beta-\beta_0)} \\
= \; & \Pr(S=1)\sup_{\beta\neq\beta_0} \frac{\|\beta-\beta_0\|^2}{(\beta-\beta_0)' E_{x_1,x_2}\left\{\frac{E\left\{\nabla_\beta\phi(\beta_0)|x_1,x_2\right\}E\left\{\nabla_\beta\phi(\beta_0)|x_1,x_2\right\}'}{\Pr(S=1|x_1,x_2)}\right\}(\beta-\beta_0)}
\end{aligned}$$ [13]

Using the well-known result that $\inf_{x\neq 0}\frac{x'Ax}{x'x}$ =smallest eigenvalue of A, and the fact that $\Pr(S=1|x_1,x_2)>0$, a.e. (x_1,x_2), it is sufficient for smoothness that

$$E\{\nabla_\beta\phi(\beta_0)|x_1,x_2\}E\{\nabla_\beta\phi(\beta_0)|x_1,x_2\}'$$

is full rank a.e. (x_1,x_2).[14]

The Riesz representor for this problem (see e.g. Shen (1997)) is given by

$$v^*=\left(\Delta^{*-1}\lambda,-w^*\times\Delta^{*-1}\lambda\right)$$

and satisfies

$$\lambda'(\beta-\beta_0)=\langle v^*,\alpha-\alpha_0\rangle.$$

Following the steps in Shen (1997) and AC (proof of corollary C.3 in Appendix C and theorem 4.1), one gets

$$\begin{aligned}&\sqrt{n}\left(\hat\beta-\beta_0\right)\\=\;&\sqrt{n}\Delta^{*-1}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}s_{0i}+\frac{1}{n}\sum_{i=1}^{n}s_{1i}+\frac{1}{n_2}\sum_{i=1}^{n_2}s_{21i}+\frac{1}{n}\sum_{i=1}^{n}s_{22i}\right)+o_p(1)\\=\;&\Delta^{*-1}\left(\sqrt{\frac{1}{\Pr(S=1)}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}s_{0i}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{1i}+\frac{\gamma}{\sqrt{n_2}}\sum_{i=1}^{n_2}s_{21i}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{22i}\right)\\&+o_p(1)\end{aligned}$$

[13] The 3rd line follows from the 2nd since for any vector of functions $l(x)$ where $x=(x_1,x_2)$, the fact that $f(x|S=1)=\frac{\Pr(S=1|x)f(x)}{\Pr(S=1)}$ implies that

$$E_x\left\{\left(\frac{1}{\Pr(S=1|x)}\right)^2 l(x)\,l(x)'|S=1\right\}=\frac{1}{\Pr(S=1)}E_x\left\{\frac{l(x)\,l(x)'}{\Pr(S=1|x)}\right\}$$

[14] For both the linear and logit model, a necessary and sufficient condition for this is that $(X_2-X_1)(X_2-X_1)'$ is full rank, which is the standard identification condition.

where

$$
\begin{aligned}
s_{0i} &= E\left(\frac{\nabla\phi(u,\beta_0)}{g(\kappa_0(u))} - \frac{g'(\kappa_0(u))\,\phi(u,\beta_0)}{g^2(\kappa_0(u))}w^*(u)\,|S=1,x_{1i},x_{2i}\right) \times \frac{\phi(y_{1i},y_{2i},x_{1i},x_{2i},\beta_0)}{g(\kappa_0(z_{1i},z_{2i},v_i))} \\
s_{1i} &= -E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))}w^*(u)\,|z_{1i},v_i\right\} \times \left(\frac{S_i}{g(\kappa_0(z_{1i},z_{2i},v_i))} - 1\right) \\
s_{21i} &= -E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))}w^*(u)\,|z_{2i}^*,v_i^*\right\} \\
s_{22i} &= E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))}w^*(u)\,|z_{2i},v_i\right\} \times \frac{S_i}{g(\kappa_0(z_{1i},z_{2i},v_i))} \qquad (18) \\
\gamma &= \lim_{n_2,n\to\infty}\sqrt{\frac{n}{n_2}},\ \Pr(S=1) = \lim_{n_2,n\to\infty}\sqrt{\frac{n_1}{n}}.
\end{aligned}
$$

While the expressions for s_{0i}, s_{1i} are exactly analogous to the RHS in AC's corollary C.3 (iii), the forms of s_{21i} and s_{22i} are worked out in the appendix. Thus, the asymptotic variance of $\hat{\beta}$ is $\Delta^{*-1}V\Delta^{*-1}$ where V is the asymptotic variance of

$$
\sqrt{\frac{1}{\Pr(S=1)}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}s_{0i} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{1i} + \frac{\gamma}{\sqrt{n_2}}\sum_{i=1}^{n_2}s_{2i} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{22i}.
$$

Comparing to AC's notation (see their equation 16), their $\Sigma(X)$ corresponds to the identity matrix here, their $E\left(D_{w^*}(X)'D_{w^*}(X)\right)$ corresponds to Δ^* here and their $E\left(D_{w^*}(X)'\Sigma_0(X)D_{w^*}(X)\right)$ corresponds to V here.

The additional regularity conditions for the asymptotic normality result are outlined in the appendix, together with a proof of the fact that under these conditions, the assumptions 4.1-4.6 in AC hold. Theorem 4.1 of AC then implies that for $\Omega = \Delta^{*-1}V\Delta^{*-1}$,

$$
\sqrt{n}\left(\hat{\beta} - \beta_0\right) \to N(0,\Omega).
$$

3.4 Estimation of covariance matrix

First consider estimation of Δ^*. Define

$$
\begin{aligned}
H_{0j}(x_{1i},x_{2i},w_j) &= \sum_{k=1}^{n_1}\left\{\begin{array}{c}\left(\frac{\frac{\partial}{\partial\beta_j}\phi(u_k,\hat{\beta})}{g(\hat{\kappa}(u_k))} - \frac{g'(\hat{\kappa}(u_k))\phi(u_k,\hat{\beta})}{g^2(\hat{\kappa}(u_k))}w_j(u_k)\right) \\ \times p_0^{k_n}(x_{1k},x_{2k})'(P_0'P_0)^{-1}p_0^{k_n}(x_{1i},x_{2i})\end{array}\right\} \\
H_{1j}(z_{1i},v_i,w_j) &= \sum_{k=1}^{n}\left(-\frac{S_kg'(\hat{\kappa}(u_k))}{g^2(\hat{\kappa}(u_k))}w_j(u_k)\right)p_1^{k_n}(z_{1k},v_k)'(P_1'P_1)^{-1}p_1^{k_n}(z_{1i},v_i) \\
H_{2j}(z_{2i}^*,v_i^*,w_j) &= \frac{1}{n}\sum_{k=1}^{n_1}\left(-\frac{S_kg'(\hat{\kappa}(u_k))}{g^2(\hat{\kappa}(u_k))}w_j(u_k)\right)p_2^{k_n}(z_{2k},v_k)'\left(\frac{P_2^{*\prime}P_2^*}{n_2}\right)^{-1}p_2^{k_n}(z_{2i}^*,v_i^*).
\end{aligned}
$$

Estimate w_j^* by $\hat{w}_j^*$ which solves

$$\min_{w_j \in \mathbb{K}_n} H(w_j) = \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{H_{0j}(x_{1i}, x_{2i}, w_j)\right\}^2 + \frac{1}{n}\sum_{i=1}^{n}\left\{H_{1j}(z_{1i}, v_i, w_j)\right\}^2 + \frac{1}{n_2}\sum_{i=1}^{n_2}\left\{H_{2j}(z_{2i}^*, v_i^*, w_j)\right\}^2. \tag{19}$$

Then Δ^* can be estimated by

$$\frac{1}{n_1}\sum_{i=1}^{n_1} H_0(x_{1i}, x_{2i}, \hat{w}^*)' H_0(x_{1i}, x_{2i}, \hat{w}^*) + \frac{1}{n}\sum_{i=1}^{n} H_1(z_{1i}, v_i, \hat{w}^*)' H_1(z_{1i}, v_i, \hat{w}^*)$$
$$+\frac{1}{n_2}\sum_{i=1}^{n_2} H_2(z_{2i}^*, v_i^*, \hat{w}^*)' H_2(z_{2i}^*, v_i^*, \hat{w}^*),$$

where for $j = 1, 2, ...d_\beta$, $H_0(x_{1i}, x_{2i}, \hat{w}^*) = \{H_{0j}(x_{1i}, x_{2i}, \hat{w}^*)\}$ etc.

Now, consider the estimation of V. Recall the terms that go into the definition of V. I shall outline the estimation for three of the terms in (18). The rest are analogous. Consider the last two terms and let $E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))} w^*(u) | z_2, v\right\} \equiv G(z_2, v)$. The variance of this sum is

$$\begin{aligned}
& Var\left\{E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))} w^*(u) | z_{2i}, v_i\right\} \times \left(\frac{S_i}{g(\kappa_0(z_{1i}, z_{2i}, v_i))} - 1\right)\right\} \\
= \; & E\left\{G(z_2, v) G(z_2, v)' \left(\frac{S_i}{g(\kappa_0(z_{1i}, z_{2i}, v_i))} - 1\right)^2\right\} \\
= \; & E_{z_2,v}\left\{G(z_2, v) G(z_2, v)' E\left(\left(\frac{S_i}{g(\kappa_0(z_{1i}, z_{2i}, v_i))} - 1\right)^2 | z_2, v\right)\right\} \\
= \; & E_{z_2,v}\left\{G(z_2, v) G(z_2, v)' E\left(\frac{1 - g(\kappa_0(z_{1i}, z_{2i}, v_i))}{g(\kappa_0(z_{1i}, z_{2i}, v_i))} | z_2, v\right)\right\} \\
\equiv \; & E_{z_2,v}\left\{G(z_2, v) G(z_2, v)' M(z_2, v)\right\}
\end{aligned}$$

which is consistently estimated by

$$\begin{aligned}
& \frac{1}{n_2}\sum_{j=1}^{n_2} \hat{G}\left(z_{2j}^*, v_j^*\right) \hat{G}\left(z_{2j}^*, v_j^*\right)' \hat{M}\left(z_{2j}^*, v_j^*\right) \\
\hat{G}\left(z_{2j}^*, v_j^*\right) \; = \; & H_2\left(z_{2j}^*, v_j^*, \hat{w}^*\right) \\
\hat{M}\left(z_{2j}^*, v_j^*\right) \; = \; & \frac{1}{n}\sum_{k=1}^{n_1}\left(\frac{1 - g(\hat{\kappa}(z_{1k}, z_{2k}, v_k)) S_k}{g(\hat{\kappa}(z_{1k}, z_{2k}, v_k))}\right) p_2^{k_n}(z_{2k}, v_k)' \left(\frac{P_2^{*\prime} P_2^*}{n_2}\right)^{-1} p_2^{k_n}\left(z_{2j}^*, v_j^*\right)
\end{aligned}$$

The variance of the first term (which has to be conditioned on $S = 1$) is estimated by

$$\frac{1}{n_1}\sum_{i=1}^{n_1} H_0(x_{1i}, x_{2i}, \hat{w}^*) H_0'(x_{1i}, x_{2i}, \hat{w}^*) \frac{\phi^2\left(y_{1i}, y_{2i}, x_{1i}, x_{2i}, \hat{\beta}\right)}{g^2(\hat{\kappa}(z_{1i}, z_{2i}, v_i))}.$$

Finally, the covariance between the first and the third term, by a similar condition argument, is given by

$$\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} H_0\left(x_{1i}, x_{2i}, \hat{w}^*\right) H_2'\left(z_{2j}^*, v_j^*, \hat{w}^*\right) \frac{\phi\left(y_{1i}, y_{2i}, x_{1i}, x_{2i}, \hat{\beta}\right)}{g^2\left(\hat{\kappa}\left(z_{1i}, z_{2i}, v_i\right)\right)}.$$

Consistency of this estimate follows from an envelope condition on $\left(\frac{\phi(u_i,\beta)}{g(\kappa(u_i))}\right)^2, \left(\frac{S_i}{g(\kappa(u_i))}-1\right)^2$ over the parameter space and Holder continuity of the derivatives $\left(\frac{\partial}{\partial\beta_j}\phi\left(u_i,\beta\right) - \frac{g'(\kappa(u_i))\phi(u_i,\beta)}{g(\kappa(u_i))} w\left(u_i\right)\right)$ and $\left(\frac{S_i g'(\kappa(u_i))}{g^2(\kappa(u_i))}\right)$ in neighborhoods of α_0, which can be achieved by bounding the second derivatives uniformly over the neighborhoods. (The proof is analogous to theorem 5.1 in AC).

Please see the next subsection for discussion of implementation of these methods.

3.5 Notes on Implementation

Expressions like $\hat{m}_0$, $\hat{m}_1$ and $\hat{m}_2$ which enter the objective function to be minimized for obtaining the main estimates and terms that enter the asymptotic variance formula may look somewhat complicated at first sight. But the actual implementation of these formulae are quite straightforward. Note that all these terms involve expressions like

$$\hat{f}_i = \sum_{k=1}^{n} f_k \left[p_1^{k_n}\left(z_{1k}, v_k\right)'\left(P_1'P_1\right)^{-1}\right] p_1^{k_n}\left(z_{1i}, v_i\right)$$

for different f_k's . These expressions can be calculated by an OLS regression of f_k's on the "regressors" $p_1^{k_n}\left(z_{1k}, v_k\right)$ and calculating the predicted values at the regressor $p_1^{k_n}\left(z_{1i}, v_i\right)$. For example, if z_{1k}, v_k are scalars and $K_n = 2$, one would regress f on $z_1, v, z_1^2, v^2, z_1 v$ and calculate the predicted values at the ith data point to get $\hat{f}_i$.

Implementation of the estimator, i.e. minimization of (13) and the estimation of its asymptotic variance (which involves the minimization step (19)) are computationally nontrivial but not prohibitively difficult. Both minimands are *smooth* functions of their arguments and so can be optimized using standard routines, e.g. those included in "Numerical Recipes" such as conjugate gradient methods. In the simulation below, Nelder-Mead's downward simplex method (which is usually applied to nonsmooth problems) has worked the best.

The choice of how many terms to include in the power series is somewhat arbitrary (just as in bandwidth choice in kernel based estimation) since the asymptotic requirements specify only the order (such as $n^{1/3}$). Larger number of terms increases the computational burden nontrivially. A rule of thumb that I have followed is to start with terms up to second degree and stop when either computation takes way too long and/or results hardly change by increasing the order. In the simulations, I could calculate the RMSE and based my choice of the order based on minimizing the RMSE within the limits of computational feasibility. As can be seen there, orders up to the computationally feasible range produce good answers.

4 Extensions

In this section, I consider 2 extensions to the problem, (i) inference on the attrition function alone without estimation of β and (ii) v the time-invariant regressors enter the original moment function $\phi(.)$.

4.1 Attrition Function

Notice that in the analysis above, β and κ were estimated jointly. But estimating the attrition function κ separately is an interesting and useful problem in itself because it helps one estimate *any* panel data model subsequently by inverse probability weighting. This can be done without altering the above analysis too much. Notice that proposition 1 already discusses identification of κ. κ can be estimated by minimizing (over a sieve space for κ) the sample analog of

$$E_{z_1,v}\left\{E\left\{\frac{S}{g\left(\kappa\left(z_1,z_2,v\right)\right)}-1|z_1,v\right\}^2\right\}+E_{z_2,v}\left\{E\left\{\frac{S}{g\left(\kappa\left(z_1,z_2,v\right)\right)}-1|z_2,v\right\}^2\right\}.$$

Consistency and the rate of convergence of this estimate can be obtained by dropping the m_0 terms from $Q(.)$ and dropping β from α and retaining the m_1 and m_2 terms in the analysis of section 3.1 and 3.2. One would get the rate

$$\begin{aligned}
&\left\|\hat{\kappa}-\kappa_0\right\|^2\\
=\ & E\left\{E\left\{\frac{g'\left(\kappa_0\left(u\right)\right)}{g\left(\kappa_0\left(u\right)\right)}\left(\hat{\kappa}\left(u\right)-\kappa_0\left(u\right)\right)|z_1,v\right\}\right\}^2+E\left\{E\left\{\frac{g'\left(\kappa_0\left(u\right)\right)}{g\left(\kappa_0\left(u\right)\right)}\left(\hat{\kappa}\left(u\right)-\kappa_0\left(u\right)\right)|z_2,v\right\}\right\}^2\\
=\ & o_p\left(n^{-1/2}\right).
\end{aligned}$$

4.2 Time-Invariant Regressors

The methods developed above extend to situations where time-invariant regressors v are included in the original moment function $\phi(.)$ e.g. in random effect models where β will include the coefficients of such variables. Notice that the attrition function includes v as an argument and so no substantive adjustments are necessary in the reweighted moment condition

$$E\left(\frac{S\phi(x_1, y_1, y_2, x_2, v, \beta)}{g(\kappa(z_1, z_2, v))}|x_1, x_2\right) = 0 \text{ for all } x_1, x_2,$$

which can be used in conjunction with (7) and (8) to estimate κ and the corresponding β.

5 Simulation Experiment with CPS Data

5.1 Panel structure of the CPS

The Current Population Survey sample-rotation scheme works as follows (for further details, please see the CPS website at http://www.census.gov/prod/2002pubs/tp63rv.pdf). A housing unit is interviewed for 4 consecutive months, is not in the sample for the next 8 months, is interviewed again the next four months and then retired from the sample. In addition, the outgoing units (ORG's) are replaced by housing units drawn from the same geographical area and this fresh sample is called the "incoming rotation group" (IRG). In any CPS sample, the rotation group status of the household is denoted by the variable "month-in-sample" or MIS. Thus every household has an MIS number from 1 through 8.

For the purpose of this paper, I concentrate on the earners file (which has data only on the outgoing rotation groups, i.e. MIS=4 and MIS=8) from the CPS for 1999 and 2000. This file has information on union status which I use in the analysis. The panel is constructed by matching the individuals in 1999 with MIS=4 with those in 2000 with MIS=8, using both the household and individual ID's as well as sex and race (see Madrian and Lefgren (2000), for an account of the imperfect matching based only on ID's). The ideal refreshment sample is the incoming rotations group, i.e. set of households with MIS=1, in the month following the month for which the outgoing rotations group (ORG) had MIS=8. However, since union status is only reported for individuals with MIS=4 or 8, one can use as the refreshment sample the individuals with MIS=4 in 2000. This assumes that there is no attrition from

MIS=1 till MIS=4 for this incoming group. Sample attrition between 1999 and 2000 is about 25%.

5.2 Simulation

The simulation experiment is run for those units of the 1999-2000 panel for whom there is no attrition in the true data. I treat this sample as the "population". The main equation of interest is a wage equation

$$\log(wage)_{it} = \alpha_i + \beta_1 * union_{it} + \beta_2 * age_{it} + \beta_3 * age_{it}^2 + \varepsilon_{it}.$$

The simulation is conducted as follows.

1. I estimate the parameters β for the above equation from the "population".

2. For this population, I use these estimates to generate "artificial" wage data for both periods after including a fixed effect α which equals log of age (normalized by its mean) in the first period plus a standard normal random variable ζ. This makes α mean 0 but correlated with the covariate. Also generated are standard normal error terms $\varepsilon_1, \varepsilon_2$, independent of covariates; the joint distribution of the covariates in the simulation is left identical to that in the "population". This forces the moment condition

$$E\left[\Delta\varepsilon_{it} | w_{i1}, w_{i2}\right] = 0, \tag{20}$$

with

$$\begin{aligned} \Delta\varepsilon_{it} &= \varepsilon_{it} - \varepsilon_{i,t-1} \\ w_{it} &= \left(\text{union}_{it}, \text{age}_{it}, \text{age}_{it}^2\right). \end{aligned}$$

on the data-generating process.

3. I draw a sample from this "population" and artificially introduce attrition according to a known attrition function plus a random noise. Survival (=1-Attrition) from the sample is modelled as

$$S_i = 1\left(h\left(\text{lwage}_{i2}\right) - h\left(\text{lwage}_{i1}\right) + c \times \text{lwage}_{i1} \times \text{lwage}_{i2} - u_i > 0\right) \tag{21}$$

where lwage is the natural log of weekly wage in dollars, u_i is generated from a standard normal distribution and

$$h\left(u\right) = \ln\left(1 + |u - 3|\right) \times sgn\left(u - 3\right).$$

The function $h(.)$ is deliberately chosen to be identical to the one in Newey and Powell (2003) and is smooth enough to satisfy the requirements of the consistency and asymptotic normality proofs. The constant c is used to model deviations from the quasi-separability assumption (2). I report simulation results for 3 values of c- 0 (no mis-specification), 0.5 (moderate mis-specification) and $c = 1$ (significant mis-specification).

4. I draw a random sample from the second period observations and take this as my refreshment sample. I then estimate the attrition function and the β's based on the artificial y and the true x's. Each replication corresponds to one draw of a primary and refreshment sample from the "population". In order to get an estimate of the rate of convergence, I perform this analysis for three different sample sizes drawn from the original CPS sample and compare the mean-squared error (RMSE) and the mean absolute deviation (MAD) as the average squared and absolute differences respectively between my estimates from the replications to the "true" values.

5.3 Implementation and Results

I create the population by keeping only men between the ages of 15 and 65 who report non-zero wages and for whom there is no attrition in the CPS data. This "population" consists of 20500 individuals, each observed in both 1999 and 2000. I randomly selected half of those to make the samples not too large in comparison to the sizes of other samples that are commonly used. The summary statistics of the relevant variables for these individuals are given in Table 1. Each replication consists of the following steps.

1. Take a 50% random sample of the population as the primary sample.

2. An independent 50% random sample of these households is taken as the refreshment sample and the values of their variables corresponding to year 2000 are retained.

3. Introduce attrition according to (21) on the primary sample observations and retain the individuals for whom $S = 1$. The remaining observations of the primary sample are discarded.

4. Estimate (1) using the sample with only the survivors and again after correcting for attrition.

5. Steps 1-4 are repeated for 12.5% and 5% samples to check how fast the performance falls with decreasing sample size.

Approximating functions (for $\Pr(S = 1|z_1, z_2, v)$) are of the form $\Phi\left(\bar{k}_1(z_1) + \bar{k}_2(z_2)\right)$

where

$$\bar{k}_1(z_1) \equiv \bar{k}_1(z_1) = \sum_{j=0}^{k} \delta_{1j} \text{lwage}_1^j$$
$$\bar{k}_2(z_2) \equiv \bar{k}_2(z_2) = \sum_{j=1}^{k} \delta_{2j} \text{lwage}_2^j$$

Thus we have $K_{1n} + 3 = 2k + 4$ parameters ($K_{1n} = 2k + 1$ δ's and 3 β's) to solve. The asymptotic theory above suggests that a choice of $k = n^{1/7}$ should work for the present problem in that it satisfies the conditions S3-S6 of the consistency and asymptotic normality propositions in the appendix. The precise choice of k was guided by both computational ease and size of RMSE. For the 50% sample, $k = 1, 2, 3$ and $k = 5$ led to larger MSE compared to $k = 4$ but for $k = 6$, the computation was significantly more time-consuming and prohibitively so if one has to repeat this many times as in a simulation.

Thus, for $n = 5125$ (=50% of 10250) we get a value of $k = (5125)^{1/7} \simeq 3$. Thus, I have a total of 10 parameters (3 β's and 7 δ's) to solve. I use the first four powers of log-wage to get a total of 12 unconditional moments (see appendix for the exact moments). For the 12.5% sample (n=1280) and the 5% sample (n=256), we get $k = 2$ and thus a total of 8 parameters. For these cases, I use only the first 3 powers of lwage.

The compactness restrictions are imposed by bounding the coefficients β, δ. The results shown in the tables correspond to uniform bounds of -4 and 4 on all coefficients. Choice of different bounds had very little impact upon the estimates of the β's but produced somewhat different estimates of the δ's, which is to be expected.

Optimization was done via the Nelder-Mead algorithm using the IMSL routine "UMPOL" in Fortran 77 on a Dell (2.4 GHz) machine. The initial values were drawn from a uniform distribution on (-0.5,0.5), the initial simplex was taken to have each side equal to 1. Each replication took about 2 minutes in real time for the 50% sample and about 40 seconds for the 12.5% sample.

Tables 2-5 show the results of the simulation for 100 replications. Table 2 reports the estimates for β from the "population". Table 3, 4 and 5 correspond to c=0, c=0.5 and c=1.0, respectively. Recall that when c=0, the quasi-separability assumption holds exactly and larger values of c imply moving away from that assumption. Therefore, one expects my coefficient estimates to deteriorate as c increases and also as the sample size falls. Within each of these tables I report the estimates that are corrected for attrition using the artificial refreshment data, viz. a random sample of year 2000 observations from the original rotations group, as well as the uncorrected estimates for each of three different sample sizes. I report coefficient estimates for the β's averaged over the replications, their mean absolute deviation and root mean-squared error. One would expect that the root mean squared error for the β's to increase roughly 2 times as one goes from a sample of size 5125 to 1280 if the root-n rate is correct. This is roughly validated by the RMSE values in table 3A and 3B.

Under no mis-specification, i.e. when c=0, the corrected estimates perform much better than the uncorrected ones and this improvement becomes more pronounced as the sample size grows. This can be seen by comparing the RMSE's across panels C, B and A in table 3. Under moderate mis-specification (table 4), this feature continues to hold although the RMSE's for the "corrected" estimates are, as expected, larger than those in table 3. Comparing RMSE numbers in table 4 to those in table 5 (largest mis-specification with c=1.0), one can see that the (mean) point estimates corrected for attrition are much closer to the truth but the RMSE seems to be of similar order of magnitude to the moderately mis-specified case. In panel C of table 5 (the worst case- i.e. largest mis-specification and smallest sample size), the uncorrected coefficient for union membership appears to have a smaller RMSE than the corrected one.

6 Conclusion

This paper analyzes a two-period panel data-model with attrition. Sample attrition is allowed to depend on second period values which are unobserved for the attritors. The set-up is the one considered in HIRR, viz. that a refreshment sample from the second period is available and the attrition function is quasi separable into period one and period two variables. The main insight of the present paper is that the restrictions implied by the model are equivalent to a set of conditional moment conditions involving the unknown finite dimensional parameter of interest as well as the infinite-dimensional function. Under weaker assumptions than HIRR, the paper provides a simple and elegant proof of identification of the model parameters using the primary and refreshment datasets. The proof, unlike HIRR, does not require any complicated result from the theory of functional optimization and instead makes clever use of the peculiar forms of the conditional moment conditions derived above. Further, the moment interpretation leads to a sieve-based estimate of the model parameters. Adapting the framework of Ai and Chen (2003) to accommodate different conditioning sets in the different moment conditions, the paper provides a theory of consistency and asymptotic normality of the finite dimensional parameter estimates. The key smoothness condition required by AC for the root-n rate is established here through what may be viewed as a local analog of the moment-based identification proof. These methods are applicable to both linear and nonlinear panel data models and analogous methods are applicable to situations of nonrandomly missing data in a single cross-section.

The paper provides brief practical guidelines for implementation of these methods on real datasets and an empirical simulation exercise using CPS data shows that the estimates work well in finite samples.

Future research would aim to investigate efficiency by using appropriate weighting matrices at the estimation stage. What is indeed the efficiency bound and whether it is possible to attain this variance by either using the continuously updated estimator or by a two-step procedure need to be addressed.

Table 1: “Population” Characteristics

--

1999, N=20500

Variable	Mean	SD	Min	Max
Lnwage	10.94	0.770	4.66	12.57
Union	0.172	0.355	0	1
Age	38.84	11.87	15	65
Agesq	1650.04	934.56	225	4225

2000, N=20500

Variable	Mean	SD	Min	Max
Lnwage	11.00	0.782	1.09	12.57
Union	0.170	0.375	0	1
Age	39.72	11.79	16	65
Agesq	1716.76	946.84	256	4225

Table 2: “Population” Regression Coefficients

Union	0.0857
Age	0.1499
Agesq	-0.00162

Table 3: c=0.0

A. Size of Primary Sample=5125, Size of Auxiliary Sample=5125

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.083	0.0134	0.0109
Age	0.1447	0.0063	0.0052
Agesq	-0.0015	0.000079	0.000067
	Coefficients not corrected for attrition		
Union	0.050	0.0247	0.0228
Age	0.1184	0.0297	0.0293
Agesq	-0.0012	0.00024	0.00023

B. Size of Primary Sample=1280; Size of Auxiliary Sample=1280

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.0776	0.0288	0.0263
Age	0.1467	0.0108	0.0098
Agesq	-0.0016	0.00015	0.00013
	Coefficients not corrected for attrition		
Union	0.567	0.0759	0.0695
Age	0.1219	0.0246	0.0232
Agesq	-0.0013	0.0002	0.0002

C. Size of Primary Sample=256; Size of Auxiliary Sample=256

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.0885	0.0642	0.0512
Age	0.1488	0.0144	0.0115
Agesq	-0.0016	0.00018	0.00014
	Coefficients not corrected for attrition		
Union	0.066	0.076	0.069
Age	0.1246	0.263	0.023
Agesq	-0.0013	0.00032	0.00028

Table 4: c=0.5

A. Size of Primary Sample=5125, Size of Auxiliary Sample=5125

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.078	0.0201	0.0168
Age	0.1455	0.0129	0.0118
Agesq	-0.00156	0.00016	0.00014
	Coefficients not corrected for attrition		
Union	0.050	0.0165	0.0135
Age	0.1189	0.0386	0.0383
Agesq	-0.00128	0.0004	0.0004

B. Size of Primary Sample=1280, Size of Auxiliary Sample=1280

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.0871	0.0416	0.0339
Age	0.1462	0.0218	0.0096
Agesq	-0.00157	0.00028	0.000119
	Coefficients not corrected for attrition		
Union	0.0651	0.0255	0.0213
Age	0.1224	0.0319	0.0308
Agesq	-0.001323	0.00035	0.00033

C. Size of Primary Sample=256; Size of Auxiliary Sample=256

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.0883	0.0577	0.0551
Age	0.1488	0.0318	0.0106
Agesq	-0.0016	0.00035	0.00012
	Coefficients not corrected for attrition		
Union	0.0649	0.0568	0.0530
Age	0.1251	0.0360	0.0281
Agesq	-0.0013	0.00039	0.00031

Table 5: c=1.0

A. Size of Primary Sample=5125, Size of Auxiliary Sample=5125

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.076	0.0223	0.0171
Age	0.1441	0.0124	0.0065
Agesq	-0.0015	0.00015	0.00008
	Coefficients not corrected for attrition		
Union	0.0584	0.0393	0.0372
Age	0.1181	0.0321	0.0317
Agesq	-0.0013	0.00036	0.00035

B. Size of Primary Sample=1280, Size of Auxiliary Sample=1280

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.0844	0.0412	0.0352
Age	0.1489	0.0218	0.0092
Agesq	-0.0016	0.00028	0.00012
	Coefficients not corrected for attrition		
Union	0.0643	0.0443	0.0375
Age	0.1244	0.0194	0.0175
Agesq	-0.0013	0.00019	0.00016

C. Size of Primary Sample=256; Size of Auxiliary Sample=256

	Coeff	RMSE	MAD
	Coefficients corrected for attrition		
Union	0.074	0.0597	0.0395
Age	0.1458	0.0338	0.0105
Agesq	-0.00157	0.00037	0.00013
	Coefficients not corrected for attrition		
Union	0.0558	0.0526	0.0344
Age	0.1213	0.0328	0.0203
Agesq	-0.0013	0.000316	0.000187

References

1. Ai, C and Chen, X. (2003): Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. Econometrica, Vol 71, No 6, pp. 1795-1843.

2. Das, M. (2004): Simple estimators for nonparametric panel data models with sample attrition, Journal of Econometrics, Volume 120, Issue 1, May 2004, pp. 159-180.

3. Fitzgerald, J., Gottschalk, P. & Moffitt, R. (1998): An analysis of sample attrition in panel data, Journal of Human Resources. vol 33, number 2, pp 251-299.

4. Hausman, J. and D. Wise (1979): Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment, Econometrica Volume: 47, Issue: 2, pp. 455-474.

5. Hirano, K, Imbens, G, Ridder, G and Rubin, R. (2001): Combining panel data sets with attrition and refreshment samples, Econometrica, Vol. 69, No. 6, pp. 1645-1659.

6. Honore', B (1992): Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects, Econometrica, 60, pp. 533-565.

7. Kesavan, K. (1989): Topics in functional analysis and applications, Wiley Eastern Limited, New Delhi, India.

8. Madrian, B. and Lefgren, L. (2000): An Approach to Longitudinally Matching Current Population Survey (CPS) Respondents. Journal of Economic and Social Measurement, pp 31-62.

9. Moffitt, R. and Ridder, G. (forthcoming): Data combination methods in Econometrics.

10. Nevo, A. (2003): Using weights to adjust for sample selection when auxiliary information is available, Journal of Business and Economic Statistics, Vol. 21 (1), pp. 43-52.

11. Nevo, A. (2002): Sample selection and information-theoretic alternatives to GMM, Journal of Econometrics, Volume 107, Issues 1-2, March 2002, pp. 149-157.

12. Newey, W. and McFadden, D. (1994): Large sample estimation and hypothesis testing, Handbook of Econometrics, vol IV, pp 2113-2245, Elsevier Science B.V. Amsterdam.

13. Newey, W. and Powell, J. (2003): Instrumental variables estimation of Nonparametric models, Econometrica, 71, pp 1557-1569.

14. Nicoletti, C. (2006): Nonresponse in dynamic panel data models, Journal of Econometrics, Volume 132, Issue 2, June 2006, Pages 461-489.

15. Pakes, A. and David Pollard (1989): Simulation and the Asymptotics of Optimization Estimators, Econometrica, vol. 57, no. 5, pp. 1027-1057.

16. Ridder G. (1990): Attrition in Multi-wave Panel Data, Panel data and labor market studies. 1990, pp. 45-67.

17. Ridder, G. (1992): An Empirical Evaluation of Some Models for Non-random Attrition in Panel Data, Structural Change and Economic Dynamics, vol. 3, no. 2, pp. 337-55

18. Shen, X. (1997): On methods of sieves and penalization, Annals of Statistics, 25, pp 2555-2591.

19. Verbeek, M & Nijman, T. (1992): Testing for selectivity bias in dynamic panel data models, International Economic Review, vol 33, pp. 681-703.

20. Wooldridge, J. (2002): Inverse probability weighted M-estimators for sample selection, attrition and stratification, Portugese Economic Journal, Vol 1, pp 117-139.

21. Wooldridge, J. (1999): Asymptotic properties of Weighted M-Estimators for variable probability sampling, Econometrica, Vol. 67, pp 1385-1406.

7 Appendix

7.1 Identification

Proof of Proposition 1

Proof. Let the subscript 0 denote true parameters- e.g. $k_{00}(.)$ is the true function while $k_0(.)$ is a generic candidate function, $\kappa_0 = k_{00}(.)+k_{10}(.,.)+k_{20}(.,.)$ and $\kappa = k_0(.)+k_1(.,.)+k_2(.,.)$. Below, I suppress the arguments of $\kappa(.)$ and $\kappa_0(.)$ but note that both $\kappa_0(.)$ and $\kappa(.)$ are functions of (z_1, z_2, v). Notice further that $E\{S|z_1, z_2, v\} = g(\kappa_0)$ and so

$$\begin{aligned} E\left\{\frac{S}{g(\kappa)} - 1|z_1, v\right\} &= E\left(E\left\{\frac{S}{g(\kappa)} - 1|z_1, z_2, v\right\}|z_1, v\right) \\ &= E\left(\frac{1}{g(\kappa)}E\{S|z_1, z_2, v\} - 1|z_1, v\right) \\ &= E\left\{\frac{g(\kappa_0)}{g(\kappa)} - 1|z_1, v\right\}. \end{aligned}$$

Therefore, (10) is equivalent to

$$\begin{aligned} E\left\{\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}|z_1, v\right\} &= 0 \text{ for all } z_1, v \\ E\left\{\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}|z_2, v\right\} &= 0 \text{ for all } z_2, v. \end{aligned} \tag{22}$$

This implies, by the law of iterated expectations, that for $w_0(v) = k_{00}(v) - k_0(v)$ and $w_j(z_j, v) = k_{0j}(z_j, v) - k_j(z_j, v)$, $j = 1, 2$,

$$\begin{aligned} & E\left\{\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}\{\kappa_0 - \kappa\}|v\right\} \\ = \; & E\left\{\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}\{w_0(v) + w_1(z_1, v) + w_2(z_2, v)\}|v\right\} \\ = \; & E\left\{E\left(\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}\{w_0(v) + w_1(z_1, v)\}|z_1, v\right)|v\right\} \\ & +E\left\{E\left(\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}w_2(z_2, v)|z_2, v\right)|v\right\} \\ = \; & E\left\{\{w_0(v) + w_1(z_1, v)\}E\left(\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}|z_1, v\right)|v\right\} \\ & +E\left\{w_2(z_2, v)E\left(\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}|z_2, v\right)|v\right\} \\ = \; & 0, \end{aligned}$$

where the last equality follows from (22). Since g is strictly increasing, $\{g(\kappa_0) - g(\kappa)\} \times \{\kappa_0 - \kappa\}$ is strictly positive w.p.1 if $\kappa \neq \kappa_0$. This is because if $\kappa_0 > \kappa$, then $g(\kappa_0) - g(\kappa) >$

0 and if $\kappa > \kappa_0$, then $g(\kappa_0) - g(\kappa) < 0$. In either case, $g(\kappa_0) - g(\kappa)$ has the same sign as $\{\kappa_0 - \kappa\}$. Next, $g(.)$ is a c.d.f. and therefore nonnegative; so the random variable $\frac{g(\kappa_0)-g(\kappa)}{g(\kappa)}\{\kappa_0 - \kappa\}$ is non-negative with probability 1. Then for the condition

$$\begin{aligned} & E\left\{\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}\{\kappa_0 - \kappa\}|v\right\} \\ \equiv \quad & E\left\{\frac{g(\kappa_0) - g(\kappa)}{g(\kappa)}\{w_0(v) + w_1(z_1, v) + w_2(z_2, v)\}|v\right\} = 0 \end{aligned}$$

to hold we must have that for each fixed v,

$$w_0(v) + w_1(z_1, v) + w_2(z_2, v) = 0$$

for all $z_1, z_2 \in \mathcal{Z}_1(v) \times \mathcal{Z}_2(v)$. By (i), we must have that (w.p.1) for each v, $w_1(z_1, v)$ does not depend on z_1 and $w_2(z_2, v)$ does not depend on z_2. Then by (iv), we have that for each v, $w_1(z_1, v) = w_1(\bar{z}_1(v), v) = 0$ for all z_1 and $w_2(z_2, v) = w_2(\bar{z}_2(v), v) = 0$ for all z_2, implying the conclusion. ■

7.2 Consistency

Assumptions

P1 $\mathbb{K} = \{k_0(.), k_1(.,.), k_2(.,.)\}$, satisfying (14).

P2 The parameter space for β is $\Theta = \left\{\beta \in R^{d_\beta} : \beta'\beta \leq B_\beta\right\}$.

P3 $\Pr(S = 1|x_1, x_2) > 0$, a.e. x_1, x_2

C All variables x, v, y in all periods have compact support with density bounded away from zero uniformly on it. The matrices $P_j'P_j$ for $j = 0, 1, 2$ have eigen values bounded away from zero with probability 1.

ID The true value $\kappa_0 = (k_{00}(.), k_{10}(.,.), k_{20}(.,.))$ and β_0 uniquely minimize (11) (Proposition 1 outlines sufficient conditions).

LIP $\phi(u, \beta)$ is Lipschitz in β with a square integrable envelope

$$\left|\phi(u, \beta) - \phi\left(u, \tilde{\beta}\right)\right| \leq M(u)\left\|\beta - \tilde{\beta}\right\|, \; E\left(M^2(u)|S = 1, x_1, x_2\right) < \infty.$$

M1 $\sup_u \sup_{\kappa \in \mathbb{K}} \left| \frac{g'(\kappa(u))}{g^2(\kappa(u)} \right| < M_0$.

M2 $E\left\{\phi^2(u,\beta)|S=1\right\} < \infty$, $\sup_{\kappa \in \mathbb{K}} E\left(\frac{M^2(u)}{g^2(\kappa(u))}|S=1,x_1,x_2\right) < \infty$.

S1 For each n, the space $\mathbb{K}_n \subset \mathbb{K}$ is compact under the norm $\|.\|_s$.

S2 For any $\kappa \in \mathbb{K}$, there exists $\tilde{\kappa} \in \mathbb{K}_n$ such that $\|\tilde{\kappa} - \kappa\|_s \xrightarrow{P} 0$.

S3 $K_n \geq K_{1n} + d_\beta$, $K_{1n} \to \infty$ and $K_n/n \to 0$.

Proposition 2 *Under the assumptions P1,P2,P3,LIP,ID,M1,M2 the above construction of the sieve space satisfying S1,S2,S3,*

$$\|\hat{\alpha} - \alpha_0\|_s = o_p(1).$$

Proof. It is sufficient to check that all conditions in AC, Lemma 3.1 are satisfied. First note that it is possible that there exist $\beta \neq \beta_0$ and $\kappa \neq \kappa_0$ such that

$$E\left\{\frac{S\phi(y_1,y_2,x_1,x_2,\beta)}{g(\kappa(z_1,z_2,v))}|x_1,x_2\right\} = 0 \text{ for all } x_1,x_2.$$

But $E\left\{m(\alpha)' m(\alpha)\right\}$ will be a strictly positive number if $\beta \neq \beta_0$ and $\kappa \neq \kappa_0$ and equal 0 for $\beta = \beta_0$ and $\kappa = \kappa_0$ since the conditions

$$\begin{aligned} E\left\{\frac{S}{g(\kappa(z_1,z_2,v))} - 1|z_1,v\right\} &= 0 \text{ for all } z_1,v \\ E\left\{\frac{S}{g(\kappa(z_1,z_2,v))} - 1|z_2,v\right\} &= 0 \text{ for all } z_2,v. \end{aligned}$$

will hold if and only if $\kappa = \kappa_0$, by identification.

Next I show Holder continuity, analogous to condition 3.6 (ii) of AC. To see this, note that by definition of the parameter space, the functions $\kappa(.)$ are bounded away from $-\infty$, whence $\frac{g'(.)}{g^2(.)}$ can be assumed to be strictly bounded on the parameter space. For example, if $g(u) = \frac{e^u}{1+e^u}$, then

$$\left|\frac{g'(\kappa)}{g^2(\kappa)}\right| = \left|-e^{-\kappa}\right| << \infty.$$

So, the mapping $\kappa \to \frac{1}{g(\kappa(.))}$ is Frechet differentiable w.r.t. the sup norm on $\kappa(.)$ and so by the mean-value theorem for functionals (e.g. Kesavan, Theorem A3.3),

$$\left|\frac{1}{g(\kappa(u))} - \frac{1}{g(\tilde{\kappa}(u))}\right| \leq \left(\sup_u \sup_{\kappa \in \mathbb{K}} \left|\frac{g'(\kappa(u))}{g^2(\kappa(u))}\right|\right) \|\kappa - \tilde{\kappa}\|_\infty.$$

By assumption, $m(u,\beta)$ is Lipschitz in β. Letting $\alpha = (\kappa,\beta)$ and $\tilde{\alpha} = \left(\tilde{\kappa},\tilde{\beta}\right)$, we have that

$$\begin{aligned}
\left|\frac{S}{g(\kappa(u))} - \frac{S}{g(\tilde{\kappa}(u))}\right| &\leq \left(\sup_u \sup_{\kappa\in\mathbb{K}}\left|\frac{g'(\kappa(u))}{g^2(\kappa(u))}\right|\right)\|\kappa-\tilde{\kappa}\|_\infty \text{ and} \\
\left|\frac{S\phi(u,\beta)}{g(\kappa(u))} - \frac{S\phi\left(u,\tilde{\beta}\right)}{g(\tilde{\kappa}(u))}\right| &\leq \left|\frac{\phi(u,\beta)}{g(\kappa(u))} - \frac{\phi(u,\beta)}{g(\tilde{\kappa}(u))}\right| + \left|\frac{\phi(u,\beta)}{g(\tilde{\kappa}(u))} - \frac{\phi\left(u,\tilde{\beta}\right)}{g(\tilde{\kappa}(u))}\right| \\
&\leq |\phi(u,\beta)|\left(\sup_u \sup_{\kappa\in\mathbb{K}}\left|\frac{g'(\kappa(u))}{g^2(\kappa(u))}\right|\right)\|\kappa-\tilde{\kappa}\|_\infty + \frac{M(u)}{g(\tilde{\kappa}(u))}\left\|\beta-\tilde{\beta}\right\| \\
&\leq \max\left\{|\phi(u,\beta)|\left(\sup_u \sup_{\kappa\in\mathbb{K}}\left|\frac{g'(\kappa(u))}{g^2(\kappa(u))}\right|\right), \frac{M(u)}{g(\tilde{\kappa}(u))}\right\}\|\alpha-\tilde{\alpha}\|_s
\end{aligned}$$

whence 3.6 (ii) of AC follow via assumptions M1 and M2. The other conditions are standard and follow from well-known properties of standard sieves. The rest of the proof is analogous to Newey and Powell (2003). ■

7.3 Rate of Convergence

Proof of Lemma 1

Proof. The proof works by showing that under the hypotheses of lemma 1, the (conditional) expectation of a certain non-negative random variable becomes 0, implying that the random variable must therefore be 0 with probability 1. Consider

$$\begin{aligned}
& E\left[(w_0(v)+w_1(z_1,v)+w_2(z_2,v))^2\frac{g'(\kappa_0(z_1,z_2,v))}{g(\kappa_0(z_1,z_2,v))}|v\right] \\
= \;& E\left[(w_0(v)+w_1(z_1,v))(w_0(v)+w_1(z_1,v)+w_2(z_2,v))\frac{g'(\kappa_0(z_1,z_2,v))}{g(\kappa_0(z_1,z_2,v))}|v\right] \\
& +E\left[w_2(z_2,v)\times(w_0(v)+w_1(z_1,v)+w_2(z_2,v))\frac{g'(\kappa_0(z_1,z_2,v))}{g(\kappa_0(z_1,z_2,v))}|v\right] \\
= \;& E\left[\{w_0(v)+w_1(z_1,v)\}E\left\{(w_0(v)+w_1(z_1,v)+w_2(z_2,v))\frac{g'(\kappa_0(z_1,z_2,v))}{g(\kappa_0(z_1,z_2,v))}|z_1,v\right\}|v\right] \\
& +E\left[w_2(z_2,v)E\left\{(w_0(v)+w_1(z_1,v)+w_2(z_2,v))\frac{g'(\kappa_0(z_1,z_2,v))}{g(\kappa_0(z_1,z_2,v))}|z_2,v\right\}|v\right] \\
= \;& E\left[(w_0(v)+w_1(z_1,v))E\left\{\tilde{H}(z_1,z_2,v)|z_1,v\right\}|v\right] + E\left[w_2(z_2,v)E\left\{\tilde{H}(z_1,z_2,v)|z_2,v\right\}|v\right] \\
= \;& 0
\end{aligned}$$

by (ii). Note that the random variable $\{w_0(v)+w_1(z_1,v)+w_2(z_2,v)\}^2 \times \frac{g'(\kappa_0(z_1,z_2,v))}{g(\kappa_0(z_1,z_2,v))}$ is nonnegative w.p. 1 by (iii). Conclude that for each fixed v,

$$w_0(v)+w_1(z_1,v)+w_2(z_2,v)=0$$

for all $z_1 \in \mathcal{Z}_1(v)$, $z_2 \in \mathcal{Z}_2(v)$. By (i), the above display can hold if and only if for all $z_1 \in \mathcal{Z}_1(v)$, $z_2 \in \mathcal{Z}_2(v)$, $w_1(z_1, v)$ and $w_2(z_2, v)$ do not depend on z_1 and z_2 respectively and

$$w_1(z_1, v) + w_2(z_2, v) = -w_0(v).$$

By (iv), the conclusion follows. ■

Technical assumptions and statement of rate of convergence

The following assumptions specialize the technical assumptions in AC (for establishing the rate of convergence as in theorem 3.1 of AC) to the problem of the present paper. The notation $\|.\|_s$ denotes the sup norm and ∇, ∇^2 are short-hand for gradient and Hessian, respectively.

Assumptions

SM0

(0) $\phi(.)$ is twice continuously differentiable everywhere

(i) $\sup_{\beta\in\Theta,\kappa\in\mathbb{K}_n,\|\alpha-\alpha_0\|_s=o(1)} E\left(\frac{|\nabla^2\phi(u,\beta)|}{g(\kappa(u))}|x_1, x_2, S=1\right) < \infty$, a.e. (x_1, x_2)

(ii) $\sup_{\beta\in\Theta,\kappa\in\mathbb{K}_n,\|\alpha-\alpha_0\|_s=o(1)} E\left(\left|\nabla_\beta\phi(u,\beta)\frac{g'(\kappa(u))}{g^2(\kappa(u))}\right||x_1, x_2, S=1\right) < \infty$, a.e. (x_1, x_2)

(iii) $\sup_{\beta\in\Theta,\kappa\in\mathbb{K}_n,\|\alpha-\alpha_0\|_s=o(1)} E\left\{\left|\phi(u,\beta)\frac{2(g'(\kappa(u)))^2-g(\kappa(u))g''(\kappa(u))}{g^3(\kappa(u))}\right||x_1, x_2, S=1\right\} < \infty$, a.e. (x_1, x_2).

SM1

(i) $\sup_{\kappa\in\mathbb{K}_n,\|\alpha-\alpha_0\|_s=o(1)} E\left(\left|\frac{g'(\kappa(u))}{g^2(\kappa(u))}\right||z_1, v\right) < \infty$, a.e. (z_1, v)

(ii) $\sup_{\kappa\in\mathbb{K}_n,\|\alpha-\alpha_0\|_s=o(1)} E\left\{\left|\frac{2(g'(\kappa(u)))^2-g(\bar{\kappa}(u))g''(\kappa(u))}{g^3(\kappa(u))}\right||z_1, v\right\} < \infty$, a.e. (z_1, v).

SM2

(i) $\sup_{\kappa\in\mathbb{K}_n,\|\alpha-\alpha_0\|_s=o(1)} E\left(\left|\frac{g'(\kappa(u))}{g^2(\kappa(u))}\right||z_2, v\right) < \infty$, a.e. (z_2, v)

(ii) $\sup_{\kappa\in\mathbb{K}_n,\|\alpha-\alpha_0\|_s=o(1)} E\left\{\left|\frac{2(g'(\kappa(u)))^2-g(\bar{\kappa}(u))g''(\kappa(u))}{g^3(\kappa(u))}\right||z_2, v\right\} < \infty$, a.e. (z_2, v).

Note that when $g(.)$ is the logistic function, second derivatives of $\frac{1}{g(\kappa)}$ are basically $e^{-\kappa(.)}$, whence these boundedness assumptions are sensible, given the definition of the parameter space. Similarly, if $\phi(.)$ is the linear regression function, then finite second moments of the x's and y's conditional on the x's suffices.

For assumption 3.6 (iii) of AC, I require that

Hol1 $\sup_{\sup_{\beta\in\Theta,\kappa\in\mathbb{K}_n}} \left|\frac{S\phi(u,\beta)}{g(\kappa(u))}\right| < c_2(u)$ with $E\left(c_2^2(u)\right) < \infty$.

Since Θ is compact, $\phi(u,.)$ is continuous and $\mathbb{K}_n \subset \mathbb{K}$, this condition trivially holds.

Hol2 For each value of $\beta \in \Theta, \kappa \in \mathbb{K}_n$ each of the functions $m_0(\alpha, x_1, x_2), m_1(\alpha, z_1, v), m_2(\alpha, z_2, v)$ lie in a Holder ball of diameter c, e.g.

$$\sup_{x_1,x_2} |m_0(\alpha, x_1, x_2)| + \max_{a_1+a_2+\ldots a_{\dim x_1+\dim x_2} \le [\gamma]} \max_{\substack{(x_1,x_2) \\ \neq (x_1',x_2')}} \frac{|\nabla^a m_0(\alpha, x_1, x_2) - \nabla^a m_0(\alpha, x_1', x_2')|}{\|(x_1, x_2) - (x_1', x_2')\|_E^{\gamma-[\gamma]}} \le c < \infty.$$

where γ denotes the number of derivatives considered for defining the Holder ball and $\gamma > \dim x$ for m_0, $\gamma > \frac{1}{2}(\dim z + \dim v)$ for m_1 and $\gamma > \frac{1}{2}(\dim z + \dim v)$ for m_2.

This condition is basically saying that the conditional mean $m_0(\alpha, x_1, x_2)$ is a smooth and bounded function of the conditioning variables. In particular, partial derivatives up to order at least half the dimension of the conditioning variables should be uniformly bounded w.p. 1.

The following conditions are analogous to assumptions 3.2 (iii), 3.5(iii), 3.7(ii) in AC

S4 For $\gamma > \frac{1}{2}\{d_v + d_{z_1}\}$ any function of (z_1, v) which is smooth up to order γ, can be approximated by power series up to degree k_n in (z_1, v) respectively, with maximum error of the order of $O\left(k_n^{-\gamma/\left(d_v+d_{z_1}\right)}\right) = o\left(n^{-1/4}\right)$. Analogously for functions of (x_1, x_2) and (z_2, v).

S5 All the functions $k_0(.), k_1(.,.)$ and $k_2(.,.)$ belonging to the parameter space are smooth enough that they can be approximated by power series in their arguments up to order K_{1n}, with the approximation error of the order of $O\left(K_{1n}^{-\mu}\right)$ with $K_{1n}^{-\mu} = o\left(n^{-1/4}\right)$.

S6 $K_n^2 K_{1n} \ln(n) = o\left(n^{1/2}\right)$.

Proposition 3 *Under all conditions of propositions 1 and 2, SM0, SM1, SM2, Hol1 and Hol2 plus conditions S4,S5,S6, we have*

$$\|\hat{\alpha} - \alpha_0\| = o_p\left(n^{-1/4}\right).$$

Proof. Follow proof of AC, theorem 3.1. ∎

7.4 Asymptotic Normality

The assumptions and formal proposition are as follows.

PD

(i) Conditional on $S = 1$ a.e. (x_1, x_2), $E\{\nabla_\beta \phi(z_1, z_2, \beta_0) | x_1, x_2\} \times E\{\nabla_\beta \phi(z_1, z_2, \beta_0) | x_1, x_2\}'$ is full rank ,

(ii) V, defined above, is positive definite,

(iii) $\beta_0 \in int\{\Theta\}$,

In addition, I assume that the following regularity conditions hold: for all $(\beta, \kappa) \equiv \alpha \in \Theta \times \mathcal{K}_n$ which satisfy $\|\alpha - \alpha_0\|_s = o(1)$ and $\|\alpha - \alpha_0\| = o\left(n^{-1/4}\right)$,

HESS1 $\sup_\beta \left\|\nabla^2 \phi(u, \beta)\right\| \leq c_3(u)$ with $E\left(c_3^2(u)\right) < \infty$.

HESS2 $\sup_{\beta,\kappa} \left\|\frac{g'(\kappa(u))\nabla\phi(u,\beta)}{g(\kappa(u))}\right\| \leq c_4(u)$ with $E\left(c_4^2(u)\right) < \infty$.

HESS3 $\sup_{\beta,\kappa} \left|\phi(u, \beta) \frac{2(g'(\kappa(u)))^2 - g(\kappa(u))g''(\kappa(u))}{g^2(\kappa(u))}\right| \leq c_5(u)$ with $E\left(c_5^2(u)\right) < \infty$.

Proposition 4 *Under assumptions of proposition 2 and the assumptions PD(i)-PD(iii), HESS1, HESS2, HESS3, we have that*

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \to N(0, \Omega)$$

where

$$\Omega = \Delta^{*-1} V \Delta^{*-1}$$

Proof. After verifying the regularity conditions (see immediately below), the proof is analogous to AC theorem 4.1. ■

<u>Regularity conditions for asymptotic normality</u>

First, I verify that the regularity conditions imposed in the propositions imply that conditions 4.1-4.6 of AC hold. For details on why these conditions are necessary, the reader should consult the AC manuscript. I use the notation of the AC paper verbatim and specialize the AC assumptions to the present problem. The upper ~notation will indicate an intermediate value as will be used while evoking the mean value theorem.

Denote the Riesz representor above as $v^* = \left(v_\beta^*, v_\kappa^*\right) = \left(v_\beta^*, -w^* \times v_\beta^*\right)$, $v_n^* = [v_\beta^*, -\Pi_n w^*(t) \times v_\beta^*] \in \Theta \times \mathbb{K}_n - \alpha_0$ such that $\|v_n^* - v^*\| = O\left(n^{-1/4}\right)$ (see AC assumption 4.2) and write

$$\begin{aligned}
\frac{d\rho_0(\alpha,t)}{d\alpha}[v_n^*] &= \nabla\phi(t,\beta)\,v_\beta^* - \frac{g'(\kappa(t))\,\phi(t,\beta)}{g(\kappa(t))}\Pi_n w^*(t)\times v_\beta^* \\
\frac{d\rho_1(\alpha,t)}{d\alpha}[v_n^*] &= \frac{g'(\kappa(t))}{g(\kappa(t))}\Pi_n w^*(t)\times v_\beta^* \\
\frac{d\rho_2(\alpha,t)}{d\alpha}[v_n^*] &= \frac{g'(\kappa(t))}{g(\kappa(t))}\Pi_n w^*(t)\times v_\beta^*.
\end{aligned}$$

Then, envelope condition analogous to 4.3 (i) in AC follows from the definition of the parameter space. Next,

$$\begin{aligned}
& \left|\frac{d\rho_0(\beta_1,\kappa_1,t)}{d\alpha}[v_n^*] - \frac{d\rho_0(\beta_2,\kappa_2,t)}{d\alpha}[v_n^*]\right| \\
\leq\ & \left|\nabla\phi(t,\beta_1)\,v_\beta^* - \nabla\phi(t,\beta_2)\,v_\beta^*\right| \\
& + \left|\frac{g'(\kappa_1(t))\,\phi(t,\beta_1)}{g(\kappa_1(t))}\Pi_n w^*(t)\times v_\beta^* - \frac{g'(\kappa_2(t))\,\phi(t,\beta_2)}{g(\kappa_2(t))}\Pi_n w^*(t)\times v_\beta^*\right| \\
=\ & \left\|\nabla\phi(t,\beta_1) - \nabla\phi(t,\beta_2)\right\|\left\|v_\beta^*\right\| \\
& + \left|\frac{g'(\kappa_1(t))\,\phi(t,\beta_1)}{g(\kappa_1(t))} - \frac{g'(\kappa_2(t))\,\phi(t,\beta_2)}{g(\kappa_2(t))}\right|\left\|\Pi_n w^*(t)\times v_\beta^*\right\|
\end{aligned}$$

Given the bounds on the coefficients in the sieve space and the definition of the parameter space, condition 4.3 in AC can be satisfied by bounding the second derivatives by square integrable envelopes over a $\|.\|_s = o(1)$ neighborhood of the truth.

Next,

$$\begin{aligned}
\frac{dm_0(\alpha,x_1,x_2)}{d\alpha}[v_n^*] &= E\left(\frac{\nabla\phi(t,\beta)\,v_\beta^*}{g(\kappa(t))} - \frac{g'(\kappa(t))\,\phi(t,\beta)}{g^2(\kappa(t))}\Pi_n w^*(t)\times v_\beta^* | S=1,x_1,x_2\right) \\
\frac{dm_1(\alpha,z_1,v)}{d\alpha}[v_n^*] &= E\left(\frac{g'(\kappa(t))}{g(\kappa(t))}\Pi_n w^*(t)\times v_\beta^* | z_1,v\right) \\
\frac{dm_2(\alpha,z_2,v)}{d\alpha}[v_n^*] &= E\left(\frac{g'(\kappa(t))}{g(\kappa(t))}\Pi_n w^*(t)\times v_\beta^* | z_2,v\right).
\end{aligned}$$

Therefore, the conditions HESS1, HESS2 and HESS3, imply that for $\alpha \in \mathcal{A}_n$, $\|\alpha - \alpha_0\| = o(1)$ and $\|\alpha - \alpha_0\|_s = o\left(n^{-1/4}\right)$, we have

$$\begin{aligned}
& \frac{dm_0(\alpha, x_1, x_2)}{d\alpha}[v_n^*] - \frac{dm_0(\tilde{\alpha}, x_1, x_2)}{d\alpha}[v_n^*] \\
= \; & E\left(\frac{\nabla\phi(t,\beta)\, v_\beta^*}{g(\kappa(t))} - \frac{g'(\kappa(t))\,\phi(t,\beta)}{g^2(\kappa(t))}\Pi_n w^*(t) \times v_\beta^* | S = 1, x_1, x_2\right) \\
& - E\left(\frac{\nabla\phi(t,\beta)\, v_\beta^*}{g(\tilde{\kappa}(t))} - \frac{g'(\tilde{\kappa}(t))\,\phi\left(t,\tilde{\beta}\right)}{g^2(\tilde{\kappa}(t))}\Pi_n w^*(t) \times v_\beta^* | S = 1, x_1, x_2\right) \\
= \; & E\left(\frac{\nabla^2\phi\left(t,\bar{\beta}\right)}{g(\kappa(t))} | S = 1, x_1, x_2\right)\left(\beta - \tilde{\beta}\right) v_\beta^* \\
& - E\left(\left\{\begin{array}{c} \frac{g'(\bar{\kappa}(t))}{g(\bar{\kappa}(t))}\nabla\phi\left(t,\bar{\beta}\right)\left(\beta - \tilde{\beta}\right) \\ + (\kappa(t) - \tilde{\kappa}(t))\,\phi\left(t,\bar{\beta}\right)\frac{2(g'(\bar{\kappa}(t)))^2 - g(\bar{\kappa}(t))g''(\bar{\kappa}(t))}{g^2(\bar{\kappa}(t))} \end{array}\right\}\Pi_n w^*(t) | S = 1, x_1, x_2\right) v_\beta^*
\end{aligned}$$

Therefore, under assumptions SM0-SM2, we have

$$\begin{aligned}
E\left(\left\|\frac{dm_0(\alpha, x_1, x_2)}{d\alpha}[v_n^*] - \frac{dm_0(\tilde{\alpha}, x_1, x_2)}{d\alpha}[v_n^*]\right\|^2\right) &= o\left(n^{-1/2}\right) \\
E\left(\left\|\frac{dm_1(\alpha, z_1, v)}{d\alpha}[v_n^*] - \frac{dm_1(\tilde{\alpha}, z_1, v)}{d\alpha}[v_n^*]\right\|^2\right) &= o\left(n^{-1/2}\right) \\
E\left(\left\|\frac{dm_2(\tilde{\alpha}, z_2, v)}{d\alpha}[v_n^*] - \frac{dm_2(\tilde{\alpha}, z_2, v)}{d\alpha}[v_n^*]\right\|^2\right) &= o\left(n^{-1/2}\right)
\end{aligned}$$

which is assumption 4.4 of AC. Assumption 4.5 follows similarly from SM0-SM2. Finally, assumption 4.6 of AC follows from conditions HESS1, HESS2 and HESS3, given the definition of the parameter space and the bounded coefficients that constitute the sieve space.

Expressions for s_{21i} and s_{22i}

In analogy with AC's corollary C.3 (iii), assuming all variables are scalar for ease of notation and letting

$$G(z_2, v) = E\left\{\frac{g'(\kappa_0(u))}{g(\kappa_0(u))} w^*(u) | z_2, v\right\},$$

the contribution of the third term to the ultimate influence function is

$$
\begin{aligned}
& -\Delta^{*-1}\frac{1}{n_2}\sum_{j=1}^{n_2}\left(\begin{array}{c}\left\{\frac{1}{n}\sum_{l=1}^{n_1}\frac{S_l p_2^{k_n}(z_{2l},v_l)}{g(\kappa(z_{1l},z_{2l},v_l))}-\frac{1}{n_2}\sum_{l=1}^{n_2}p_2^{k_n}\left(z_{2l}^*,v_l^*\right)\right\} \\ G\left\{z_{2j}^*,v_j^*\right\}\left(\frac{P_2^{*\prime}P_2^*}{n_2}\right)^{-}p_2^{k_n}\left(z_{2j}^*,v_j^*\right)\end{array}\right) \\
= & -\Delta^{*-1}\frac{1}{n}\sum_{l=1}^{n_1}\frac{S_l}{g\left(\kappa\left(z_{1l},z_{2l},v_l\right)\right)}p_2^{k_n}\left(z_{2l},v_l\right)\frac{1}{n_2}\sum_{j=1}^{n_2}\left(\frac{P_2^{*\prime}P_2^*}{n_2}\right)^{-}p_2^{k_n}\left(z_{2j}^*,v_j^*\right)G\left\{z_{2j}^*,v_j^*\right\} \\
& +\Delta^{*-1}\frac{1}{n_2}\sum_{l=1}^{n_2}p_2^{k_n}\left(z_{2l}^*,v_l^*\right)\left\{\frac{1}{n_2}\sum_{j=1}^{n_2}\left(\frac{P_2^{*\prime}P_2^*}{n_2}\right)^{-}p_2^{k_n}\left(z_{2j}^*,v_j^*\right)G\left\{z_{2j}^*,v_j^*\right\}\right\} \\
= & -\Delta^{*-1}\frac{1}{n}\sum_{l=1}^{n_1}\frac{S_l}{g\left(\kappa\left(z_{1l},z_{2l},v_l\right)\right)}\hat{G}\left\{z_{2l},v_l\right\}+\Delta^{*-1}\frac{1}{n_2}\sum_{l=1}^{n_2}\hat{G}\left\{z_{2l},v_l\right\} \\
\simeq & -\Delta^{*-1}\frac{1}{n}\sum_{l=1}^{n_1}\frac{S_l}{g\left(\kappa\left(z_{1l},z_{2l},v_l\right)\right)}G\left\{z_{2l},v_l\right\}+\Delta^{*-1}\frac{1}{n_2}\sum_{l=1}^{n_2}G\left\{z_{2l},v_l\right\}+o_p\left(n^{-1/2}\right)
\end{aligned}
$$

where the arguments implying the last line is analogous to AC proof of C.3 (iii) on page 1833.

7.5 Simulation

Unconditional moments used in simulations

Letting

$$u_i(\beta)=lwage_{2i}-lwage_{1i}-\beta_1*(\text{union}_{i2}\text{-union}_{i1})-\beta_2*(\text{age}_{i2}-\text{age}_{i1})+\beta_3*\left(\text{age}_{i2}^2-\text{age}_{i1}^2\right)$$

and

$$
\begin{aligned}
\bar{k}_1(z_1) & \equiv \bar{k}_1(z_1)=\sum_{j=0}^{k_n}\delta_{1j}\text{lwage}_1^j \\
\bar{k}_2(z_2) & \equiv \bar{k}_2(z_2)=\sum_{j=1}^{k_n}\delta_{2j}\text{lwage}_2^j,
\end{aligned}
$$

we have the following sets of moment conditions corresponding to the original model

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(\frac{S_i u_i(\beta)}{\Phi\left(\bar{k}_1(z_{1i})+\ \bar{k}_2(z_{2i})\right)}(\text{union}_{i2}\text{-union}_{i1})\right) & \simeq 0 \\
\frac{1}{n}\sum_{i=1}^{n}\left(\frac{S_i u_i(\beta)}{\Phi\left(\bar{k}_1(z_{1i})+\ \bar{k}_2(z_{2i})\right)}(\text{age}_{i2}-\text{age}_{i1})\right) & \simeq 0 \\
\frac{1}{n}\sum_{i=1}^{n}\left(\frac{S_i u_i(\beta)}{\Phi\left(\bar{k}_1(z_{1i})+\ \bar{k}_2(z_{2i})\right)}\left(\text{age}_{i2}^2-\text{age}_{i1}^2\right)\right) & \simeq 0
\end{aligned}
$$

and the following moments for the attrition function corresponding to (8) (which uses both the primary and refreshment sample) together with the analogous ones for (7) (which uses only the primary sample)

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{S_i}{\Phi\left(\bar{k}_1\left(z_{1i}\right)+\ \bar{k}_2\left(z_{2i}\right)\right)}-1 &\simeq 0 \\
\frac{1}{n}\sum_{i=1}^{n}\frac{S_i\text{lwage}_{i1}^{j}}{\Phi\left(\bar{k}_1\left(z_{1i}\right)+\ \bar{k}_2\left(z_{2i}\right)\right)}-\frac{1}{n}\sum_{k=1}^{n}\text{lwage}_{k}^{j} &\simeq 0 \text{ for } j=1,..4 \\
\frac{1}{n}\sum_{i=1}^{n}\frac{S_i\text{lwage}_{i2}^{j}}{\Phi\left(\bar{k}_1\left(z_{1i}\right)+\ \bar{k}_2\left(z_{2i}\right)\right)}-\frac{1}{n_2}\sum_{k=1}^{n_2}\text{lwage}_{k}^{*j} &\simeq 0 \text{ for } j=1,..4
\end{aligned}$$

where lwage_{k2}^{*} is the the log-wage of the kth individual in the refreshment sample.