Varying Coefficient Panel Data Model in the Presence of Endogenous Selectivity and Fixed Effects

Malikov, Emir and Kumbhakar, Subal C. and Sun, Yiguo

Department of Economics, State University of New York at Binghamton, NY, Department of Economics, State University of New York at Binghamton, NY, Department of Economics and Finance, University of Guelph, Guelph, ON

2013

Varying Coefficient Panel Data Model in the Presence of Endogenous Selectivity and Fixed Effects

Emir Malikov* Subal C. Kumbhakar† Yiguo Sun‡

This Draft: November 20, 2013

Abstract

This paper considers a flexible panel data sample selection model in which (i) the outcome equation is permitted to take a semiparametric, varying coefficient form to capture potential parameter heterogeneity in the relationship of interest, (ii) both the outcome and (parametric) selection equations contain unobserved fixed effects and (iii) selection is generalized to a polychotomous case. We propose a two-stage estimator. Given consistent parameter estimates from the selection equation obtained in the first stage, we estimate the semiparametric outcome equation using data for the observed individuals whose likelihood of being selected into the sample stays approximately the same over time. The selection bias term is then "asymptotically" removed from the equation along with fixed effects using kernel-based weights. The proposed estimator is consistent and asymptotically normal. We first investigate the finite sample properties of the estimator in a small Monte Carlo study and then apply it to study production technologies of U.S. retail credit unions from 2002 to 2006.

Keywords: Credit Union, Fixed Effects, Selection, Semiparametric, Smooth Coefficient, Switching Regression, Varying Coefficient

JEL Classification: C14, C33, C34, G21

*Department of Economics, State University of New York at Binghamton, NY; *Email:* emalikov@binghamton.edu.
†Department of Economics, State University of New York at Binghamton, NY; *Email:* kkar@binghamton.edu.
‡Department of Economics and Finance, University of Guelph, Guelph, ON. *Email:* yisun@uoguelph.ca.

1 Introduction

Semiparametric methods have become a part of a standard methodological toolkit of applied researchers in economics. These methods are attractive for their ability to circumvent limitations of conventional parametric models by allowing more flexible specifications and thus mitigating (at least partly) the risk of misspecification. While they admittedly require more prior assumptions and therefore are not as flexible as their (completely) nonparametric counterparts, semiparametric models have nevertheless gained popularity due to their capability to alleviate the so-called "curse of dimensionality" associated with nonparametric estimation.

This paper considers a particular class of semiparametric models in which parameters of a linear regression are permitted to be unspecified smooth functions of some variables (Hastie and Tibshirani, 1993; Cai et al., 2000; Li et al., 2002). Such "varying coefficient" (hereinafter VC) models[1] have recently become a subject of prolific research in the econometric literature that attempts to extend the method to new settings. For instance, Das (2005), Cai et al. (2006) and Cai and Xiong (2012) consider VC models in the presence of endogenous variables and propose applying instrumental variables approach to tackle the endogeneity problem. However, the overwhelming majority of these studies place the model either in the cross-sectional (as in the above cited papers) or in the time series settings (e.g., Cai, 2007). Analysis of VC models in a panel data setting is however relatively scarce, arguably due to difficulties associated with tackling unobserved effects. For instance, Cai and Li (2008) study a VC model in the dynamic panel setting that assumes any unobserved effects away. Sun et al. (2009) somewhat fill the void by proposing a VC panel data model estimator which allows treatment of both random and fixed effects.[2]

However, the semiparametric literature has broadly overlooked another important feature of the data that applied researchers often have to deal with: namely, the presence of selectivity. Such a problem is acute in studies of wage and labor supply decisions that go back to Heckman's (1974, 1979) seminal work and many other labor economics applications and not only. In this paper, we therefore take a semiparametric VC model a step further by considering it in the panel data setting and the presence of endogenous selection and fixed effects.[3] For a similar model in a cross-sectional setting, see Das et al. (2003), whose model allows both the outcome and selection equations to take completely nonparametric forms. Das (2004) extends the above model to a panel data case with (exogenous) random effects.

Thus, we consider a flexible panel data sample selection model in which (i) the outcome equation is permitted to take a semiparametric VC form to capture potential parameter heterogeneity in the relationship of interest, (ii) both the outcome and selection equations contain unobserved fixed effects and (iii) selection is generalized to a polychotomous case. In this paper, we restrict our analysis to models with parametric selection equations. Our model can be considered as a generalization of conventional parametric panel data sample selection models [see Baltagi (2013) for a comprehensive review]. Relatively few such parametric models allow for a fixed-effect type heterogeneity. For instance, in the case of strictly exogenous right-hand-side covariates, Wooldridge (1995) and

[1] Such models are also referred to as "smooth coefficient" or "functional coefficient" models.

[2] The studies of nonparametric panel data models that consider the presence of either random or fixed effects include, e.g., Das (2003); Henderson and Ullah (2005); Henderson et al. (2008). Alternatively, there are studies that focus on panel data applications of other classes of semiparametric models such as Li and Stengos (1996); Su and Ullah (2006); Lin and Carroll (2006).

[3] Here, we focus on a panel data application, given its increasing availability to researchers (as opposed to mere cross-sectional data). We do not consider the case of random effects because applied researchers often consider the assumption of exogenous heterogeneity unsupported by the data and difficult to justify. The violation of such an assumption would yield inconsistent estimates.

Rochina-Barrachina (1999) propose correlated effects estimators, whereas Kyriazidou (1997) develops an estimator that allows for completely unspecified fixed effects in both the selection and outcome equations. In this paper, we let fixed effects to be correlated with the right-hand-side covariates in an arbitrary way and remove them "nonparametrically", which makes Kyriazidou's (1997) estimator be the closest parametric counterpart to the semiparametric one that we propose in this paper. The difference between the two lies in the facts that we let the outcome equation take a more flexible VC form and that we generalize selection to a polychotomous case.

We propose estimating our model in two stages. We suggest consistently estimating the selection equation in the first stage via any of several parametric methods available in the literature such as Manski's (1987) and Horowitz's (1992) (smoothed) conditional maximum score or Chamberlain's (1980) conditional logit estimators. The obtained estimates can then be used to evaluate the conditional probability of an individual to be selected into the sample in each time period. In the second stage, we propose estimating the VC outcome equation using data for observed individuals (cross-sections) whose estimated likelihood of being selected into the sample stays approximately the same over time. For such individuals, the sample selection bias would be approximately time-invariant and thus can be treated as another component of fixed effects present in the outcome equation. Given that there are unlikely to be many (if any at all) cross-sections with exactly the same selection probabilities over time, we adopt the idea of Ahn and Powell (1993) and Kyriazidou (1997) and weigh these cross-sections based on "closeness" of their respective selection probabilities (and thus their selectivity biases) to being the same over time. The weighted semiparametric outcome equation can then be estimated in a manner similar to that proposed by Sun et al. (2009). The selection bias term is "asymptotically" removed from the equation along with fixed effects using kernel-based weights. The latter is advantageous over conventional first-differencing[4] because it mitigates the need to use backfitting and allows identification of an intercept coefficient function. We show that, under appropriate assumptions on the rate of convergence of the first-stage estimator of the selection equation, our proposed estimator is consistent and asymptotically normal.

We first investigate the finite sample performance of the proposed estimator in a small Monte Carlo simulation. The results are encouraging and show that, in the presence of endogenous selectivity, our estimator is less biased than a "naive" estimator which overlooks the selection issue. We also find that the estimation becomes more stable as the sample size increases.

We next apply our estimator to study production technologies of U.S. retail credit unions in the period from 2002 to 2006. There has recently been a substantial interest in investigation of credit unions' production technologies, given a dramatic transformation that the U.S. credit union industry has been undergoing over the past few decades.[5] Copious mergers and acquisitions have transformed the industry from one which had primarily consisted of small-scale local institutions catering to a handful of members to a now trillion dollar industry that constitutes a significant portion of the U.S. financial services markets, serving a hundred million customers in the country (authors' calculations based on National Credit Union Administration, 2011).

Studies that have investigated the performance of U.S. credit unions had to deal with the problem of having a large number of observations for which the reported values of credit unions' outputs are zeros. Researchers have handled this problem either by linearly aggregating all types of outputs into a single bundle (e.g., Fried et al., 1999; Wheelock and Wilson, 2011, 2013) or by replacing zero outputs with an arbitrarily chosen small positive number (Frame et al., 2003). The presence of zero-value observations is however likely to be informative and may indicate significant

[4]For instance, Kyriazidou (1997) proposes applying first-differencing in order to purge the sample selection term and fixed effect from the outcome equation.

[5]See Wheelock and Wilson (2011), Malikov et al. (2013) and references therein.

differences among credit unions in terms of the service menu they offer to members. Ignoring this *observed* heterogeneity in the provision of services amounts to making a strong and rather unrealistic assumption that all credit unions share the same "production" technology that is invariant to the menu of services they provide. This assumption of homogeneous technology across credit unions is likely to result in the loss of information and the misspecification of the econometric model, which is further aggravated if the choice of the differing service menus by credit unions is endogenous (Malikov et al., 2013). In this paper, we model this observed heterogeneity as an outcome of an endogenous choice (selection). Moreover, we also allow for *unobserved* heterogeneity among credit unions, something that has been broadly overlooked in most existing studies.

We find some significant distortions in cost elasticity estimates if one ignores selectivity. Similarly, we document dramatic differences in elasticity estimates between our VC sample selection model and its parametric counterpart. We find that the estimated relationship between scale economies and the smoothing variable (here, the asset size) from our VC model is quite different from that implicitly implied by a parametric model. These findings call for extra caution when researchers first estimate a parametric model of credit union production technologies (even after controlling for selectivity) and then analyze how the estimated technological metrics change with the size of credit unions.

The rest of the paper proceeds as follows. Section 2 outlines the model. We outline the estimation procedure in Section 3. Large sample statistical properties are provided in Section 4. Section 5 presents results of a small Monte Carlo simulation. In Section 6, we apply the model to study heterogeneous production technologies of the U.S. credit unions in the period from 2002 to 2006. Section 7 concludes.

2 Varying Coefficient Panel Data Model with Endogenous Selection and Fixed Effects

We consider a VC panel data model in the presence of endogenous selection and unobserved individual fixed effects. In what follows, we confine our analysis to a selection equation that takes a parametric (single index) form.

2.1 Binary Sample Selection

In the presence of binary sample selection, the model takes the following form

$$y_{it} = \begin{cases} \mathbf{x}_{it}'\boldsymbol{\beta}(\mathbf{z}_{it}) + \mu_i + u_{it} & \text{if} \quad d_{it} = 1 \\ - & \text{otherwise} \end{cases} \tag{2.1a}$$

$$d_{it}^* = \mathbf{w}_{it}'\boldsymbol{\gamma} + \xi_i + e_{it} \ , \qquad (i = 1, \ldots, N;\ t = 1, \ldots, T) \tag{2.1b}$$

where the column vectors of exogenous covariates $\mathbf{x}_{it}$, $\mathbf{z}_{it}$ and $\mathbf{w}_{it}$ are of dimensions p, q and l, respectively; $\boldsymbol{\beta}(\mathbf{z}_{it})$ is the conformable vector of unknown parameter functions of $\mathbf{z}_{it}$; $\boldsymbol{\gamma}$ is the conformable vector of unknown (constant) parameters. None of the variables in $\mathbf{x}_{it}$ can be obtained from $\mathbf{z}_{it}$ and vice versa, whereas $\mathbf{w}_{it}$ can have common elements with both $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$. The random disturbances (u_{it}, e_{it}) are independently and identically distributed (*i.i.d.*) over i with zero means and finite variances and are orthogonal to covariates $\mathbf{x}_{it}$, $\mathbf{z}_{it}$, $\mathbf{w}_{it}$ and unobserved effects μ_i and ξ_i. The distributions of the errors are however allowed to be correlated, namely $\mathbb{E}[u_{it}e_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{w}_{it}, \mu_i, \xi_i] \neq 0$. We treat the unobserved individual effects μ_i and ξ_i as representative of the fixed-effect type (un-

observed) heterogeneity by allowing these effects to be correlated with any of the right-hand-side covariates in an arbitrary way.[6]

The selection into sample is governed by the latent variable d^*_{it} in (2.1b), of which only dichotomous realizations are observed in the form of a categorical variable $d_{it} \equiv \mathbb{1}\{d^*_{it} \geq 0\}$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function. The "selection" variable d_{it} determines observability of the response variable y_{it} in the outcome equation (2.1a), i.e., y_{it} is observed only if $d_{it} = 1$.[7]

Note that if random errors u_{it} and e_{it} are distributed independently of one another (which implies that $\mathbb{E}[u_{it}e_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{w}_{it}, \mu_i, \xi_i] = 0$), then selection is exogenous and thus "ignorable". In the latter case, the main equation of interest (2.1a) can be estimated from the selected sample while ignoring (2.1b). Thus, model (2.1) collapses to a more standard case of a semiparametric varying coefficient panel data model with fixed effects considered by Sun et al. (2009).

When $p = 1$ and $\mathbf{x}_{it} \equiv 1$ for all i and t, model (2.1) reduces to a nonparametric panel data model with selectivity and fixed effects, an extension of Henderson et al.'s (2008) model to the case of endogenous sample selection which is yet to be considered in the literature.

An extreme special case of model (2.1) is the instance when $q = 1$ and $\mathbf{z}_{it} \equiv 1$ for all i and t which renders constant parameters in the outcome equation (2.1a). Then, the model becomes completely parametric. Few papers have considered such parametric sample selection models with fixed-effect type heterogeneity in *both* outcome and selection equations. In the case of exogenous covariates (as in this paper), the three approaches to tackle unobserved effects in these types of parametric models are those of Wooldridge (1995), Kyriazidou (1997) and Rochina-Barrachina (1999).[8] Among these three papers, Kyriazidou (1997) is, however, the only study that models individual effects in a completely "nonparametric" way by making no assumption about the form of correlation between unobserved effects and right-hand-side covariates (as we do in this paper). Both Wooldridge (1995) and Rochina-Barrachina (1999) parameterize the relation between individual effects and covariates, following Chamberlain's (1980) correlated effects approach. For a concise comparison of these three estimators, see Dustmann and Rochina-Barrachina (2007).

2.2 Polychotomous Switching

We next consider an extension of model (2.1) to the case of polychotomous selection, i.e.,

$$y_{r,it} = \begin{cases} \mathbf{x}'_{r,it}\boldsymbol{\beta}_r(\mathbf{z}_{r,it}) + \mu_{r,i} + u_{r,it} & \text{if} \quad d_{r,it} = 1 \\ - & \text{otherwise} \end{cases} \tag{2.2a}$$

$$d^*_{r,it} = \mathbf{w}'_{it}\boldsymbol{\gamma}_r + \xi_{r,i} + e_{r,it} \ , \qquad (i = 1, \ldots, N;\ t = 1, \ldots, T;\ r = 1, \ldots, R) \tag{2.2b}$$

where subscript $r \equiv \{1, \ldots, R\}$, with $R \geq 2$, denotes the regimes between which regression (2.2a) switches. The regime (or, regression) switching is governed by the latent variable $d^*_{r,it}$ in (2.2b). For each regime r, we define a categorical variable $d_{r,it} \in \{0, 1\}$ such that $d_{r,it} \equiv \mathbb{1}\{\text{the } r\text{th regime is selected}\}$. The response variable $y_{r,it}$ is observed only if $d_{r,it} = 1$. The remaining variables are defined as their counterparts (with no subscript r) from Section 2.1.

The latent variable $d^*_{r,it}$ can naturally be thought of as measuring the propensity to select the

[6] Our analysis also applies to the case when $\mu_i \equiv \xi_i$.

[7] Clearly, $(d_{it}, \mathbf{w}_{it})$ are always observed. Our analysis is however insensitive to the assumption of whether $(\mathbf{x}_{it}, \mathbf{z}_{it})$ are always observed or observed only if $d_{it} = 1$.

[8] The three papers mainly consider Type 2 Tobit model, whereas Wooldridge (1995) also explicitly discusses Type 3 Tobit. For extensions of Kyriazidou's (1997) estimator, see Honoré et al. (2000) and Kyriazidou (2001).

regime r. Individual i selects the regime r in time period t if and only if

$$d^*_{r,it} > d^*_{j,it} \qquad \forall \;\; j = 1, \ldots, R \;\; (j \neq r) \; . \tag{2.3}$$

While one can treat the regime switching as a system of $(R-1)$ dichotomous decisions, we follow an alternative approach by considering the polychotomous selection problem in McFadden's (1974) random utility framework. That is, the rth regime is said to be selected if and only if

$$d^*_{r,it} > \max_{j=1,\ldots,R;\; j\neq r} \{d^*_{j,it}\} \; . \tag{2.4}$$

Substituting from (2.2b) and using the definition of $d_{r,it}$, we get

$$d_{r,it} = 1 \quad \Leftrightarrow \quad \mathbf{w}'_{it}\boldsymbol{\gamma}_r + \xi_{r,i} + e_{r,it} > \max_{j=1,\ldots,R;\; j\neq r} \{\mathbf{w}'_{it}\boldsymbol{\gamma}_j + \xi_{j,i} + e_{j,it}\} \; . \tag{2.5}$$

For convenience, let

$$\epsilon_{r,it} \equiv \max_{j=1,\ldots,R;\; j\neq r} \{\mathbf{w}'_{it}\boldsymbol{\gamma}_j + \xi_{j,i} + e_{j,it}\} - e_{r,it} \; . \tag{2.6}$$

Then it follows from (2.6) that

$$d_{r,it} = 1 \quad \Leftrightarrow \quad \epsilon_{r,it} < \mathbf{w}'_{it}\boldsymbol{\gamma}_r + \xi_{r,i} \; . \tag{2.7}$$

We can now look at the model in (2.2) as a *binary* choice (sample selection) model, for each given regime r (Maddala, 1983). That is, we can essentially replace the selection equation (2.2b) for each $r = 1, \ldots, R$ with its equivalent

$$\widetilde{d}^*_{r,it} = \mathbf{w}'_{it}\boldsymbol{\gamma}_r + \xi_{r,i} - \epsilon_{r,it} \; , \tag{2.8}$$

where $\widetilde{d}^*_{r,it}$ is a transformed latent variable such that $d_{r,it} \equiv \mathbb{1}\{\widetilde{d}^*_{r,it} > 0\}$, where the condition inside the indicator function ensures that (2.7) is satisfied. Thus, the transformed model with polychotomous switching is no different than the binary sample selection model (2.1).

3 Estimation Methodology

This section describes the estimation of the models presented above. Given that the model with polychotomous selection can be transformed into a dichotomous sample selection model that takes the form of (2.1), in what follows, we therefore primarily focus on the analysis of the latter. Unless otherwise specified, we consider $T = 2$ (which we later relax to $T > 2$).

The estimation of equation of interest (2.1a) is complicated due to two factors: (i) the presence of unobserved individual effects μ_i that are correlated with right-hand-side covariates $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ and (ii) potential "endogeneity" of these covariates, which arises as a result of their dependence on the selection variable d_{it} and thus may lead to a so-called "selection bias". The solution to neither of these two problems is trivial.

A conventional approach to remove fixed effects from (2.1a) would be to apply first-differencing to the selected sample, i.e., the observations for which $d_{it} = d_{is} = 1$ $(t \neq s)$. Equation (2.1a) would then transform to

$$y_{it} - y_{is} = \mathbf{x}'_{it}\boldsymbol{\beta}(\mathbf{z}_{it}) - \mathbf{x}'_{is}\boldsymbol{\beta}(\mathbf{z}_{is}) + u_{it} - u_{is} \qquad \text{if} \quad d_{it} = d_{is} = 1 \;\; (t \neq s) \; . \tag{3.1}$$

While the above procedure successfully removes μ_i from the equation of interest, it however comes at a cost. The right-hand side of equation (3.1) now contains the *same* unknown functions $\boldsymbol{\beta}(\cdot)$

evaluated at different observations (time periods). The kernel-based estimation of such a model would require some form of backfitting algorithm, which is known to suffer from common problems as documented in the literature on additive nonparametric models. In fact, equation (3.1) would also contain additive nonparametric functions, if some elements in $\mathbf{x}_{it}$ are time-invariant. In particular, if the first element of $\mathbf{x}_{it}$ is unity (for the intercept) with the corresponding unknown parameter function $\beta_1(\mathbf{z}_{it})$, then first-differencing would render $\beta_1(\mathbf{z}_{it}) - \beta_1(\mathbf{z}_{is})$ $[t \neq s]$ on the right-hand side of (3.1). For more on difficulties associated with the estimation of the first-differenced varying coefficient panel data model with fixed effects, see Sun et al. (2009).

Most importantly, estimation of the first-differenced model (3.1) is likely to yield inconsistent estimates of $\boldsymbol{\beta}(\cdot)$ due to endogenous sample selection. One generally should not expect that $\mathbb{E}[u_{it}|d_{it} = d_{is} = 1 \ (t \neq s), \boldsymbol{\zeta}_i] = 0$, or that $\mathbb{E}[u_{it}|d_{it} = d_{is} = 1 \ (t \neq s), \boldsymbol{\zeta}_i] = \mathbb{E}[u_{is}|d_{it} = d_{is} = 1 \ (t \neq s), \boldsymbol{\zeta}_i]$, where $\boldsymbol{\zeta}_i \equiv (\mathbf{x}_{it}, \mathbf{x}_{is}, \mathbf{z}_{it}, \mathbf{z}_{is}, \mathbf{w}_{it}, \mathbf{w}_{is}, \mu_i, \xi_i)$. Note that the conditioning set inside this "sample selection effect" contains $\mathbf{x}_{it}$ and other covariates correlated with it. Therefore, if one does not control for selectivity, the error term $(u_{it} - u_{is})$ in (3.1) is likely to be correlated with the right-hand-side covariates, thus leading to inconsistent estimates of unknown coefficient functions $\boldsymbol{\beta}(\cdot)$.

To make the discussion of the sample selection effect more explicit, we rewrite the outcome equation of interest (2.1a) for the selected sample as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}(\mathbf{z}_{it}) + \mu_i + \lambda_{it} + v_{it} \qquad \text{if} \quad d_{it} = d_{is} = 1 \ (t \neq s) \ , \tag{3.2}$$

where $\lambda_{it} \equiv \mathbb{E}[u_{it}|d_{it} = d_{is} = 1 \ (t \neq s), \boldsymbol{\zeta}_i]$ is a sample selection bias term; and $v_{it} \equiv u_{it} - \lambda_{it}$ is a new random error which satisfies $\mathbb{E}[v_{it}|d_{it} = d_{is} = 1 \ (t \neq s), \boldsymbol{\zeta}_i] = 0$ by construction.

If we assume that random errors (u_{it}, e_{it}) are *i.i.d.* not only over i but also over t, then the sample selection bias term is

$$\lambda_{it} = \mathbb{E}[u_{it}|d_{it} = 1] = \mathbb{E}[u_{it}|d_{it}^* \geq 0] = \mathbb{E}[u_{it}|e_{it} \leq \mathbf{w}_{it}'\boldsymbol{\gamma} + \xi_i] = \Lambda(\mathbf{w}_{it}'\boldsymbol{\gamma} + \xi_i) \ , \tag{3.3}$$

where $\Lambda(\cdot)$ is some unknown function, the same across individuals i and time periods t, of (partly unobservable) $\mathbf{w}_{it}'\boldsymbol{\gamma} + \xi_i$. It is clear that generally $\lambda_{it} \neq \lambda_{is}$ $(t \neq s)$ unless $\mathbf{w}_{it}'\boldsymbol{\gamma} = \mathbf{w}_{is}'\boldsymbol{\gamma}$, for each $i = 1, \ldots, N$. That is, individuals, for whom $\mathbf{w}_{it}'\boldsymbol{\gamma} = \mathbf{w}_{is}'\boldsymbol{\gamma}$, will have an equal likelihood of being selected into the sample in both time periods t and s $(t \neq s)$.

However, we note that the equality $\lambda_{it} = \lambda_{is}$ $(t \neq s)$ for an individual i such that $\mathbf{w}_{it}'\boldsymbol{\gamma} = \mathbf{w}_{is}'\boldsymbol{\gamma}$ would also hold under a weaker assumption. For instance, building on the work of Kyriazidou (1997), we can substitute the "*i.i.d.* over i and t" assumption with the assumption of $(u_{it}, u_{is}, e_{it}, e_{is})$ and $(u_{is}, u_{it}, e_{is}, e_{it})$ being identically distributed conditional on $\boldsymbol{\zeta}_i$ (for $t \neq s$), i.e., $F(u_{it}, u_{is}, e_{it}, e_{is}|\boldsymbol{\zeta}_i) = F(u_{is}, u_{it}, e_{is}, e_{it}|\boldsymbol{\zeta}_i)$, where $F(\cdot)$ is some distribution function. This "conditional exchangeability" assumption allows marginal distributions of errors (u_{it}, e_{it}) and hence the function $\Lambda(\cdot)$ to vary over i. Under this assumption, for each individual i such that $\mathbf{w}_{it}'\boldsymbol{\gamma} = \mathbf{w}_{is}'\boldsymbol{\gamma}$ $(t \neq s)$, the sample selection bias term is

$$\begin{aligned}
\lambda_{it} &= \mathbb{E}[u_{it}|d_{it} = d_{is} = 1 \ (t \neq s), \boldsymbol{\zeta}_i] = \mathbb{E}[u_{it}|e_{it} \leq \mathbf{w}_{it}'\boldsymbol{\gamma} + \xi_i, e_{is} \leq \mathbf{w}_{is}'\boldsymbol{\gamma} + \xi_i \ (t \neq s), \boldsymbol{\zeta}_i] \\
&= \mathbb{E}[u_{is}|e_{is} \leq \mathbf{w}_{it}'\boldsymbol{\gamma} + \xi_i, e_{it} \leq \mathbf{w}_{is}'\boldsymbol{\gamma} + \xi_i \ (t \neq s), \boldsymbol{\zeta}_i] \\
&= \lambda_{is} \ .
\end{aligned}$$

Thus, under either of the two above assumptions, the sample selection bias term for an individual i such that $\mathbf{w}_{it}'\boldsymbol{\gamma} = \mathbf{w}_{is}'\boldsymbol{\gamma}$ $(t \neq s)$ (i.e., for an individual with the same likelihood of being selected into the sample in periods t and s) is the same for the two time periods, $\lambda_{it} = \lambda_{is} = \lambda_i$, and can thus be treated as another time-invariant individual effect similar to μ_i. From (3.2), we get

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}(\mathbf{z}_{it}) + (\mu_i + \lambda_i) + v_{it} \qquad \text{if} \quad d_{it} = d_{is} = 1; \ \mathbf{w}_{it}'\boldsymbol{\gamma} = \mathbf{w}_{is}'\boldsymbol{\gamma} \ (t \neq s) \ , \tag{3.4}$$

where $(\mu_i + \lambda_i)$ is the "new" individual effect correlated with the right-hand-side covariates but orthogonal to the error v_{it}. Model (3.4) is a varying coefficient panel data model with fixed effects considered by Sun et al. (2009), which can be consistently estimated using observations for cross-sections that are selected into the sample in the time periods t and s $(t \neq s)$ and have $\mathbf{w}'_{it}\boldsymbol{\gamma} = \mathbf{w}'_{is}\boldsymbol{\gamma}$.

As briefly noted above, the estimation of a model like the one in (3.4) is however not straightforward due to the presence of individual effects. We propose an approach inspired by the least squares dummy variable method to tackle fixed effects in parametric panel data models and extended to semiparametric varying coefficient models by Sun et al. (2009). The approach removes fixed effects $(\mu_i + \lambda_i)$ from equation (3.4) by "asymptotically" subtracting a kernel-smoothed version of the time average from each individual i. Given that in our case we estimate the model using (two-period) *pairs* of observations for each individual selected into the sample, the approach is equivalent in its principle to a kernel-smoothed first-differencing but with some advantages over the conventional first-differencing applied to the model before the estimation. In particular, the method does not wipe out an additive constant (if there is any) present in the varying intercept coefficient in $\boldsymbol{\beta}(\cdot)$ in case there is a time-invariant element in $\mathbf{x}_{it}$.[9] Thus, this approach allows identification of the coefficient function for (at most one) time-invariant covariate in $\mathbf{x}_{it}$ (Sun et al., 2009), which, for instance, is not feasible in a completely parametric counterpart of our model considered by Kyriazidou (1997).

Until now, we have presumed that $\boldsymbol{\gamma}$ was known, which is unrealistic. We can however replace $\boldsymbol{\gamma}$ with its consistent estimate $\widehat{\boldsymbol{\gamma}}$ obtained from (2.1b) in the first stage of the estimation (which we discuss in detail later). Another concern is that there may be few individuals in the panel for whom $\mathbf{w}'_{it}\widehat{\boldsymbol{\gamma}} = \mathbf{w}'_{is}\widehat{\boldsymbol{\gamma}}$ $(t \neq s)$, which would dramatically decrease the size of a "usable" selected sample in the second stage. In fact, it is likely that, in practice, one may find no such individuals in the data at all: the case when $\Delta\mathbf{w}'_i\widehat{\boldsymbol{\gamma}} = (\mathbf{w}_{it} - \mathbf{w}_{is})'\widehat{\boldsymbol{\gamma}} \neq 0$ for all i, where Δ denotes the first-difference operator.[10] However, note that if we assume that $\Lambda(\cdot)$ is a sufficiently smooth function, the cross-sections for which the values of the single index in the selection equation are "close" to being equal across the two time periods t and s $(t \neq s)$, i.e., $\mathbf{w}'_{it}\widehat{\boldsymbol{\gamma}} \cong \mathbf{w}'_{is}\widehat{\boldsymbol{\gamma}}$, should also have $\lambda_{it} \cong \lambda_{is} \cong \lambda_i$. Our argument would therefore hold approximately.

Thus, it is natural to weigh the selected cross-sections on the basis of the "closeness" of $\Delta\mathbf{w}'_i\widehat{\boldsymbol{\gamma}}$ to zero (also see Kyriazidou, 1997). Intuitively, the cross-sections, for which the selection likelihoods are close to being the same in both time periods t and s $(t \neq s)$, ought to be given heavier weights. The latter is in the spirit of Ahn and Powell (1993) who propose a somewhat similar approach to remove the "sample selection effect" in a cross-sectional setting.[11]

To introduce our estimator, we first rewrite the outcome equation for the selected sample (3.2) as follows

$$\phi_i y_{it} = \phi_i \mathbf{x}'_{it}\boldsymbol{\beta}(\mathbf{z}_{it}) + \phi_i\mu_i + \phi_i\lambda_{it} + \phi_i v_{it} \qquad (t \neq s)\ , \tag{3.5}$$

where, for convenience, we define $\phi_i \equiv \mathbb{1}\{d_{it} = d_{is} = 1\ (t \neq s)\}$. For identification purposes, we need a restriction on unobserved fixed effects in order to estimate $\boldsymbol{\beta}(\cdot)$ in (3.5). Along the lines of Su and Ullah (2006), we assume $\sum_{i=1}^{N} \phi_i\mu_i = 0$.

[9] Also, unlike in the case of traditional first-differencing when the number of usable observations is halved, the method we consider saves all observations.

[10] Moreover, theory suggests that, if $\Delta\mathbf{w}'_i\widehat{\boldsymbol{\gamma}}$ is a continuous variable, then $\Pr[\Delta\mathbf{w}'_i\widehat{\boldsymbol{\gamma}} = 0] = 0$.

[11] The fundamental difference between Ahn and Powell's (1993) approach and ours lies in the following. They propose differencing out sample selection bias by subtracting one cross-sectional unit from another cross-sectional unit "matched" on the basis of similarity in the two individual's likelihoods of being selected into the sample. In contrast, we eliminate the sample selection effect by "comparing" observations for the *same* cross-section across the two time periods. Also, the selection effects are wiped out "asymptotically" rather than by the pre-estimation first-differencing.

Model (3.5) takes the following matrix form

$$\mathbf{Y} = \text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(\mathbf{z})\} + \mathbf{D}\boldsymbol{\mu} + \boldsymbol{\lambda} + \mathbf{V} \ , \tag{3.6}$$

where $\mathbf{Y}$, $\mathbf{V}$ and $\boldsymbol{\lambda}$ are $2N \times 1$ vectors defined as $B = (\phi_1 b_{1t}, \phi_1 b_{1s}, \ldots, \phi_i b_{it}, \phi_i b_{is}, \ldots, \phi_N b_{Nt}, \phi_N b_{Ns})$ for $t \neq s$ with $B \in \{\mathbf{Y}, \mathbf{V}, \boldsymbol{\lambda}\}$ and $b_{it} \in \{y_{it}, v_{it}, \lambda_{it}\}$, respectively; $\text{mtx}\{\cdot\}$ is the operator that stacks $\phi_i \mathbf{x}_{it}' \boldsymbol{\beta}(\mathbf{z}_{it})$ into a $2N \times 1$ vector with the (i,t) subscripts matching those of $\mathbf{Y}$, $\mathbf{V}$ and $\boldsymbol{\lambda}$. Let cross-sections be enumerated so that $\phi_1 = 1$. Then, $\boldsymbol{\mu} = (\phi_2 \mu_2, \ldots, \phi_i \mu_i, \ldots, \phi_N \mu_N)$ is an $(N-1) \times 1$ vector of fixed effects for $i = 2, \ldots, N$ and $\mathbf{D} = [-\mathbf{i}_{N-1} \ \mathbf{I}_{N-1}]' \otimes \mathbf{i}_2$ is a $2N \times (N-1)$ design matrix, where $\mathbf{I}_m$ is the identity matrix of dimension m, $\mathbf{i}_m$ is an $m \times 1$ vector of ones, and $\otimes$ is the Kronecker product operator. Both $\boldsymbol{\mu}$ and $\mathbf{D}$ are defined so that the identifying restriction $\sum_{i=1}^{N} \phi_i \mu_i = 0$ is satisfied.

Motivated by Sun et al. (2009) and Kyriazidou (1997), we propose estimating unknown coefficient functions $\boldsymbol{\beta}(\cdot)$ from the following (local) kernel-weighted least squares problem

$$\min_{\boldsymbol{\beta}(z), \boldsymbol{\mu}} \quad (\mathbf{Y} - \text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(z)\} - \mathbf{D}\boldsymbol{\mu})' \widehat{\mathbf{K}}_h(z) (\mathbf{Y} - \text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(z)\} - \mathbf{D}\boldsymbol{\mu}) \ , \tag{3.7}$$

where $\widehat{\mathbf{K}}_h(z) = \text{diag}\left\{\widehat{\psi}_1 \mathbf{K}_h(\mathbf{z}_1, z), \ldots, \widehat{\psi}_i \mathbf{K}_h(\mathbf{z}_i, z), \ldots, \widehat{\psi}_N \mathbf{K}_h(\mathbf{z}_N, z)\right\}$ is a $2N \times 2N$ diagonal local weighting matrix comprised of 2×2 product kernel matrices $\mathbf{K}_h(\mathbf{z}_i, z) = \text{diag}\{\mathcal{K}_h(\mathbf{z}_{it}, z), \mathcal{K}_h(\mathbf{z}_{is}, z)\}$ and scalar kernel weights $\widehat{\psi}_i$ for $i = 1, \ldots, N$. Here, $\mathcal{K}_h(\mathbf{z}_{it}, z) = \mathcal{K}(\mathbf{H}^{-1}(\mathbf{z}_{it} - z))$ is the product kernel and $\widehat{\psi}_i = k(h_0^{-1} \Delta \mathbf{w}_i' \widehat{\boldsymbol{\gamma}})$ is a kernel weight, where $\mathbf{H} = \text{diag}\{h_1, \ldots, h_q\}$ is the diagonal bandwidth matrix of dimension q for $\mathbf{z}_{it}$ and h_0 is the bandwidth for $\Delta \mathbf{w}_i' \widehat{\boldsymbol{\gamma}}$. $\mathcal{K}(\cdot)$ and $k(\cdot)$ are defined in Assumption **K** below.

It is convenient to think of $\widehat{\mathbf{K}}_h(z)$ in (3.7) as a "generalized" local weighting matrix. It (i) weights the selected *observations* on the basis of their closeness to z and (ii) weights the selected *cross-sections* on the basis of the closeness of their likelihoods to be selected in the two periods. The latter permits to asymptotically remove sample selection effects λ_{it}. Essentially, model (3.7) is equivalent to a varying coefficient panel data model with fixed effects [of the form in (2.1a)], where $\Delta \mathbf{w}_i' \widehat{\boldsymbol{\gamma}}$ is an extra argument of unknown coefficient functions $\boldsymbol{\beta}(\cdot)$ that are to be evaluated at the zero value of $\Delta \mathbf{w}_i' \widehat{\boldsymbol{\gamma}}$ for all i.

The first-order condition of the optimization problem in (3.7) with respect to unknown fixed effects $\boldsymbol{\mu}$ is

$$\mathbf{D}' \widehat{\mathbf{K}}_h(z) (\mathbf{Y} - \text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(z)\} - \mathbf{D}\boldsymbol{\mu}) = \mathbf{0}_{(N-1) \times 1} \ , \tag{3.8}$$

which can be solved for $\widehat{\boldsymbol{\mu}}$, i.e.,

$$\widehat{\boldsymbol{\mu}} = (\mathbf{D}' \widehat{\mathbf{K}}_h(z) \mathbf{D})^{-1} \mathbf{D}' \widehat{\mathbf{K}}_h(z) (\mathbf{Y} - \text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(z)\}) \ . \tag{3.9}$$

Substituting (3.9) into (3.7) yields the concentrated (local) kernel-weighted least squares problem from which unknown fixed effects are removed, i.e.,

$$\min_{\boldsymbol{\beta}(z)} \quad (\mathbf{Y} - \text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(z)\})' \widehat{\boldsymbol{\Gamma}}_h(z)' \widehat{\mathbf{K}}_h(z) \widehat{\boldsymbol{\Gamma}}_h(z) (\mathbf{Y} - \text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(z)\}) \ , \tag{3.10}$$

where $\widehat{\boldsymbol{\Gamma}}_h(z) \equiv \mathbf{I}_{2N} - \mathbf{D}(\mathbf{D}' \widehat{\mathbf{K}}_h(z) \mathbf{D})^{-1} \mathbf{D}' \widehat{\mathbf{K}}_h(z)$. Note that $\widehat{\boldsymbol{\Gamma}}_h(z) \mathbf{D}\boldsymbol{\mu} = \mathbf{0}_{2N \times 1}$, which removes individual effects from the model. One can easily see the resemblance between $\widehat{\boldsymbol{\Gamma}}_h(z)$ and a standard within-transformation matrix used in parametric fixed effects panel data models, with the difference between the two amounting to the presence of the kernel weighting matrix $\widehat{\mathbf{K}}_h(z)$ in the former.

Indeed, it is easy to show that $\widehat{\mathbf{\Gamma}}_h(z)$ transforms the data by subtracting the kernel-weighted time average from each cross-section i for which $\phi_i = 1$.[12]

Before we proceed, two remarks are warranted. First, in the above discussion, we have focused on the case when $T = 2$, i.e., when there is only *one* pair of time periods for each individual i to consider. However, our analysis naturally extends to a general case of the panel with $T \geq 2$. One can estimate model (3.10) for $\mathcal{C}(T, 2)$ unique pairs of the time periods.[13] The estimates of unknown coefficient functions $\boldsymbol{\beta}(\cdot)$ can then, for instance, be combined using some minimum distance measure. It is preferable to combine the estimates using optimal weights, in order to obtain which, one needs to estimate the covariance matrix of the estimators for different pairs of the time periods. For the case of a parametric counterpart of our model (2.1), i.e., when $\boldsymbol{\beta}(\cdot)$ are constant, Charlier et al. (2001) show that the covariances between the estimators for different pairs of the time periods converge to zero. We conjecture that the same holds in the case of our model. One therefore may combine estimates of $\boldsymbol{\beta}(\cdot)$ using the inverses of corresponding variances as weights.[14]

Second, practitioners often encounter truncated, rather than selected, samples of data. That is, data often contain observations for which $d_{it} = 1$ for all i and t, which renders the estimation of $\boldsymbol{\gamma}$ from (2.1b) infeasible due to the lack of variation in d_{it}. However, note that $\lambda_{it} \cong \lambda_{is} \cong \lambda_i$ $(t \neq s)$ should also hold when $\mathbf{w}_{it} \cong \mathbf{w}_{is}$ (i.e., when $\Delta\mathbf{w}_i \cong \mathbf{0}_{l\times 1}$). It is therefore natural to completely omit the estimation of (2.1b) in the first stage and proceed directly to the estimation of $\boldsymbol{\beta}(\cdot)$ from (3.10). The only modification needed is in the weights $\widehat{\psi}_i$, which would be natural to redefine as $\widehat{\psi}_i = \mathcal{K}(\mathbf{H}_0^{-1}\Delta\mathbf{w}_i)$, where $\mathbf{H}_0 = \text{diag}\{h_{0,1}, \ldots, h_{0,l}\}$ is the diagonal bandwidth matrix of dimension l. The drawback of this approach is twofold. It results in a slower rate of convergence for the estimator implied by (3.10), and it requires that all covariates in $\mathbf{w}_{it}$ be excluded from $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ (see Kyriazidou, 1997).

3.1 First Stage

In order to estimate unknown coefficient parameters of $\boldsymbol{\beta}(\cdot)$ from (3.10), we first need to obtain consistent estimates of the parameter vector $\boldsymbol{\gamma}$ in the selection equation. Since one can only observe dichotomous realizations of the latent variable d_{it}^*, the selection equation (2.1b) presents itself as a limited dependent variable (discrete choice) panel data model, which can be consistently estimated in a number of ways.

If we assume that the random error e_{it} in (2.1b) is *i.i.d.* over i and t with the logistic distribution, then we can estimate $\boldsymbol{\gamma}$ via Chamberlain's (1980) conditional logit estimator that yields $\sqrt{N}$ consistent $\widehat{\boldsymbol{\gamma}}$. In this paper, the latter approach is our primary choice. Clearly, in the presence of individual effects ξ_i, the parameters of time-invariant elements of $\mathbf{w}_{it}$ in the selection equation (2.1b) are not identified. We therefore restrict all elements of $\mathbf{w}_{it}$ to be time-varying.

Alternative methods to obtain $\widehat{\boldsymbol{\gamma}}$ include Manski conditional maximum score estimator or its "smoothed" version (the smoothed conditional maximum score estimator) considered by Kyriazidou (1997, 2001). The latter is an extension of Horowitz's (1992) smoothed maximum score estimator to a panel-data setting in the presence of individual effects. The advantage of these two estimators is that they avoid any distributional assumptions about the error term e_{it} in the selection equation (2.1b). However, eliminating a possibility of distributional misspecification, which may result in inconsistent estimates of $\widehat{\boldsymbol{\gamma}}$ obtained from the conditional logit, comes at a cost. Manski's (1987)

[12] For more, see the discussion in Sun et al. (2009).

[13] Here, $\mathcal{C}(T, 2) = \frac{T!}{2!(T-2)!}$.

[14] Combining estimates of $\boldsymbol{\beta}(\cdot)$ using such a minimum distance procedure is, however, not imperative. In this paper, we combine estimates of coefficient parameters by simple averaging, as for instance, suggested by Kyriazidou (1997).

conditional maximum score estimator converges at a slower rate of $N^{-1/3}$ and, more importantly, its asymptotic distribution is non-normal, which would complicate the analysis of the limit distribution of our estimator $\widehat{\boldsymbol{\beta}}(\cdot)$ obtained from (3.10). The "smoothed" conditional maximum score estimator is however asymptotically normal and its convergence rate can be made arbitrarily close to $N^{-1/2}$. See Kyriazidou (1997, 2001) for a more thorough discussion of these alternatives to the conditional logit estimator.

Throughout this paper, we employ Chamberlain's (1980) conditional logit estimator in the first stage. We opt for it primarily due to its ability to also tackle the models of polychotomous choice, like the one considered in (2.2). In the presence of polychotomous switching, we assume that the error term $e_{r,it}$ in (2.2b) is *i.i.d.* over i and t with the type I extreme-value distribution, which yields a multinomial logistic $\epsilon_{r,it}$ as defined in (2.6). Similar to the case of binary sample selection, only coefficients of time-varying elements of $\mathbf{w}_{it}$ are identified in (2.8).

To our knowledge, no other parametric method other than the maximum likelihood allows the estimation of a discrete choice model with polychotomous choice in the panel-data setting [of the form in (2.2b)] that offers a completely "nonparametric" treatment of the fixed effects present in the equation. While Manski's (1987) maximum score estimator has been extended to the case of multinomial choice by Fox (2007) and then generalized to a "smoothed" version by Yan (2013), both estimators are yet to be extended to the case of panel data when unobserved heterogeneity is controlled for. Other methods either treat individual effects as random, under a rather strong and unrealistic assumption of their exogeneity, or model them as correlated effects (see Malikov et al., 2013 and references therein).

3.2 Second Stage

We propose estimating unknown coefficient parameters $\boldsymbol{\beta}(\cdot)$ in the outcome equation from (3.10) via local-linear fitting, which has numerous well-documented advantages over a more commonly used local-constant estimator that suffers from non-adaptation and boundary effects (Fan and Gijbels, 1996). Under the assumption that smooth functions are twice differentiable in the neighborhood of z, each element in $\boldsymbol{\beta}(\cdot)$ can then be approximated using the first-order Taylor expansion, i.e., $\beta_j(\mathbf{z}_{it}) \approx \beta_j(z)+\nabla\boldsymbol{\beta}_j(z)'(\mathbf{z}_{it}-z)$ for $j=1,\dots,p$, where $\nabla\boldsymbol{\beta}_j(z) \equiv (\partial\beta_j(z)/\partial z_{1,it},\dots,\partial\beta_j(z)/\partial z_{q,it})$ is a $q\times 1$ vector of first derivatives of $\beta_j(\cdot)$ with respect to $\mathbf{z}_{it}$, evaluated at z.

Define a $(q+1)\times 1$ vector $\boldsymbol{\theta}_j(z) \equiv (\beta_j(z), \nabla\boldsymbol{\beta}_j(z))$ of an unknown coefficient function and its first derivatives with respect to $\mathbf{z}_{it}$ for each $j=1,\dots,p$. Then, the unknown $p\times(q+1)$ parameter matrix is defined as $\boldsymbol{\Theta}(z) \equiv [\boldsymbol{\theta}_1(z) \ \dots \ \boldsymbol{\theta}_p(z)]'$, i.e.,

$$\boldsymbol{\Theta}(z) \equiv \begin{bmatrix}\boldsymbol{\theta}_1(z)'\\ \vdots \\ \boldsymbol{\theta}_p(z)'\end{bmatrix} = \begin{bmatrix}\beta_1(z) & \nabla\boldsymbol{\beta}_1(z)'\\ \vdots & \vdots \\ \beta_p(z) & \nabla\boldsymbol{\beta}_p(z)'\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta}(z) & \nabla\boldsymbol{\beta}(z)'\end{bmatrix} ,$$

where the first column of the above matrix is $\boldsymbol{\beta}(z)$ as defined in (2.1a), evaluated at z; $\nabla\boldsymbol{\beta}(z) \equiv [\nabla\boldsymbol{\beta}_1(z) \ \dots \ \nabla\boldsymbol{\beta}_p(z)]$ is a $q\times p$ matrix of first derivatives.

Next, define a $(q+1)\times 1$ vector of deviations from z, i.e., $\mathcal{Z}_{it}(z) \equiv (1, \mathbf{z}_{it}-z)$. For $\mathbf{z}_{it}$ close to z, we approximate the outcome equation (3.5), which we are interested in estimating, by

$$\phi_i y_{it} \approx \phi_i \mathbf{x}_{it}'\boldsymbol{\Theta}(z)\mathcal{Z}_{it}(z) + \phi_i\mu_i + \phi_i\lambda_{it} + \phi_i v_{it} \qquad (t\neq s) , \tag{3.11}$$

where we have replaced $\boldsymbol{\beta}(z)$ with $\boldsymbol{\Theta}(z)\mathcal{Z}_{it}(z)$.

We can obtain the local-linear estimates of functional coefficients $\boldsymbol{\beta}(z)$ from the concentrated kernel-weighted least squares problem that take the following form [from (3.10)]

$$\min_{\boldsymbol{\Theta}(z)} \quad (\mathbf{Y} - \mathcal{X}(z)\text{vec}\{\boldsymbol{\Theta}(z)\})'\widehat{\mathbf{W}}_h(z)(\mathbf{Y} - \mathcal{X}(z)\text{vec}\{\boldsymbol{\Theta}(z)\}) \ , \tag{3.12}$$

where $\widehat{\mathbf{W}}_h \equiv \widehat{\mathbf{\Gamma}}_h(z)'\widehat{\mathbf{K}}_h(z)\widehat{\mathbf{\Gamma}}_h(z)$. Here, for the ease of matrix manipulations, we stack (by columns) the unknown parameter matrix $\boldsymbol{\Theta}(z)$ into a $p(q+1)\times 1$ vector denoted by the operator vec$\{\cdot\}$. The $2N \times p(q+1)$ data matrix $\mathcal{X}(z)$ is

$$\mathcal{X}(z) = \begin{bmatrix}\phi_1\mathcal{Z}_{1t}(z)\otimes\mathbf{x}_{1t} & \phi_1\mathcal{Z}_{1s}(z)\otimes\mathbf{x}_{1s} & \ldots & \phi_N\mathcal{Z}_{Nt}(z)\otimes\mathbf{x}_{Nt} & \phi_N\mathcal{Z}_{Ns}(z)\otimes\mathbf{x}_{Ns}\end{bmatrix}' \ .$$

Lastly, solving the first-order condition of (3.12) in terms of unknown $\boldsymbol{\Theta}(z)$ yields the following weighted least-squares estimator

$$\text{vec}\{\widehat{\boldsymbol{\Theta}}(z)\} = \left(\mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\mathcal{X}(z)\right)^{-1}\left(\mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\mathbf{Y}\right) \ . \tag{3.13}$$

Taking Taylor expansion at an interior point z and replacing $\mathbf{Y}$ by (3.6) gives

$$\text{vec}\{\widehat{\boldsymbol{\Theta}}(z)\} = \text{vec}\{\boldsymbol{\Theta}(z)\} + E_N^{-1}(A_N/2 + B_N + C_N + D_N) \ , \tag{3.14}$$

where $A_N \equiv \mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\mathbf{\Pi}(z)$, $B_N \equiv \mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\mathbf{D}\boldsymbol{\mu}$, $C_N \equiv \mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\boldsymbol{\lambda}$, $D_N \equiv \mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\mathbf{V}$, $E_N \equiv \mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\mathcal{X}(z)$ and the $[t+2(i-1)]$th element of the column vector $\mathbf{\Pi}(z)$ is $\phi_i\mathbf{x}_{it}'\mathbf{r}(\widetilde{\mathbf{z}}_{it}, z)$ with $\mathbf{r}(\cdot,\cdot) = [r_1(\cdot,\cdot),\ldots,r_p(\cdot,\cdot)]$ and $r_l(\widetilde{\mathbf{z}}_{it}, z) = (\mathbf{z}_{it} - z)'\frac{\partial^2\beta_l(\widetilde{\mathbf{z}}_{it})}{\partial z\partial z'}(\mathbf{z}_{it} - z)$, and $\widetilde{\mathbf{z}}_{it}$ lies between $\mathbf{z}_{it}$ and z for each i and t.

3.3 Bandwidth Selection

Bandwidths for covariates $\mathbf{z}_{it}$ as well as for $\Delta\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}$, a single index from the selection equation, are central to the local-linear estimator described in the previous subsection. As noted above, one can essentially view (3.13) as the estimator of a varying coefficient panel data model with fixed effects, where $\Delta\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}$ is an extra argument of the unknown functions $\boldsymbol{\beta}(\cdot)$ [and therefore of $\boldsymbol{\Theta}(\cdot)$] that are to be evaluated at the zero value of $\Delta\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}$ for all i. It is therefore natural to consider selecting bandwidths for both $\mathbf{z}_{it}$ and $\Delta\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}$ simultaneously. While several optimal bandwidth selection methods are available in the nonparametric literature, in this paper, we consider the data-driven cross-validation method (Li and Racine, 2004). Due to the presence of *unobserved* individual effects, the standard leave-one-(observation)-out cross-validation is however unlikely to perform well in our case. We therefore expand on Sun et al.'s (2009) suggestion and instead consider a leave-one-*selected-individual*-out cross-validation in order to select optimal bandwidths $\mathbf{H}$ and h_0.[15] In particular, when estimating $\widehat{\boldsymbol{\Theta}}(\cdot)$ at the data point $\mathbf{z}_{it}$ for an individual i such that $\phi_i = 1$ [i.e., an individual from the (observable) selected sample] via the estimator in (3.13), we withhold $\phi_i(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i)$ from the data actually used in the estimation. We denote the resulting estimate of $\boldsymbol{\beta}(\mathbf{z}_{it})$ as $\widehat{\boldsymbol{\beta}}_{-i}(\mathbf{z}_{it})$. Thus, the modified cross-validation objective is

$$\min_{\mathbf{H},h_0} \quad \mathcal{CV}(\mathbf{H}, h_0) \equiv \left(\mathbf{Y} - \text{mtx}\{\mathbf{x}, \widehat{\boldsymbol{\beta}}_{-i}(\mathbf{z}_{it})\}\right)'\mathbf{Q}'\mathbf{Q}\left(\mathbf{Y} - \text{mtx}\{\mathbf{x}, \widehat{\boldsymbol{\beta}}_{-i}(\mathbf{z}_{it})\}\right) \ , \tag{3.15}$$

[15] All individuals that are not selected into the sample (and thus are unobserved) are already left out of the estimation by means of $\phi_i = 0$.

where $\mathbf{Q}$ is a $2N \times 2N$ transformation matrix such that $\mathbf{Q}\mathbf{D}_{ur} = \mathbf{0}_{2N\times N}$, where $\mathbf{D}_{ur} = \mathbf{I}_N \otimes \mathbf{i}_2$ is a standard $2N \times N$ design matrix for individual effects under no identifying restrictions. No restriction is needed because $\mathbf{Q}$ removes fixed effects entirely from the equation. Among few options available for the choice of $\mathbf{Q}$, we consider a within-transformation matrix $\mathbf{Q} \equiv \mathbf{I}_{2N} - \mathbf{I}_N \otimes (\mathbf{i}_2\mathbf{i}_2')/2$.

The above cross-validation is asymptotically equivalent to the minimization of the integrated squared error (Li and Racine, 2004). In fact, one can easily rewrite the cross-validation function $\mathcal{CV}(\mathbf{H}, h_0)$ in (3.15) as

$$\begin{aligned}\mathcal{CV}(\mathbf{H}, h_0) = &\Big(\text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(\cdot)\} - \text{mtx}\{\mathbf{x}, \widehat{\boldsymbol{\beta}}_{-i}(\cdot)\}\Big)' \mathbf{Q}'\mathbf{Q} \Big(\text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(\cdot)\} - \text{mtx}\{\mathbf{x}, \widehat{\boldsymbol{\beta}}_{-i}(\cdot)\}\Big) + \\ &2\Big(\text{mtx}\{\mathbf{x}, \boldsymbol{\beta}(\cdot)\} - \text{mtx}\{\mathbf{x}, \widehat{\boldsymbol{\beta}}_{-i}(\cdot)\}\Big)' \mathbf{Q}'\mathbf{Q}\mathbf{V} + \mathbf{V}'\mathbf{Q}'\mathbf{Q}\mathbf{V} \ , \end{aligned} \tag{3.16}$$

where the first term on the right-hand side of (3.16) is the weighted mean squared error (a good approximation of the integrated squared error) with $\mathbf{Q}'\mathbf{Q}$ being the weighting matrix; the second term has zero expected value, since we have assumed that all covariates are exogenous and v_{it} is mean zero conditional on being selected into the sample by construction; and the third term does not depend on $(\mathbf{H}, h_0)$. Thus, minimizing the cross-validation function $\mathcal{CV}(\mathbf{H}, h_0)$ amounts to minimizing the weighted means squared error of the model as estimated by the first term in (3.16).

4 Asymptotic Properties

This section presents limit results for our proposed estimator (3.13). Given that the nature of our estimator is such that it uses pairs of the time periods, as in the previous section, here we consider the case of $T = 2$, i.e., $t = 1,\ s = 2$.

We first list some regularity assumptions used to support the limit results for the case when parameter vector γ is *known*.

Assumption D (*Data Generating Mechanism*). $(\xi_i, \mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{w}_{it}, u_{it}, e_{it})$ are *i.i.d.* over i, and y_{it} is generated from equation (2.1a).

(i) If $x_{it,l} \equiv x_{i,l}$ for at most one $l \in \{1, \cdots, p\}$, i.e., $x_{i,l}$ does not depend on t, we assume $\mathbb{E}[\phi_i x_{i,l}] \neq 0$;

(ii) $\sum_{i=1}^N \phi_i \mu_i = 0$ for all i, and $\phi_1 = 1$;

(iii) v_{it} is *i.i.d.* with zero mean and variance σ_v^2 conditional on $(\phi_i, \boldsymbol{\zeta}_i)$ for all t;

(iv) $\mathbf{z}_{it}$ is a continuous $q \times 1$ random vector and has a Lebesgue probability density $f_t(z)$ for $t = 1, 2$, and $f(z) = f_1(z) + f_2(z) > 0$ for each $z \in \mathbb{R}^q$;

(v) $\Delta s_i \equiv (\mathbf{w}_{i1} - \mathbf{w}_{i2})'\boldsymbol{\gamma}$ is a continuous random variable for all i and t, and $(\Delta s_i, \mathbf{z}_{it})$ has a Lebesgue probability density $f_t(s, z)$ for $t = 1, 2$.

Assumption **D** (ii) implies $\widehat{\boldsymbol{\Gamma}}_h(z)\mathbf{D}\boldsymbol{\mu} = \mathbf{0}$. Hence, $B_N = \mathbf{0}$ in (3.14), and there is no bias associated with the unobserved fixed effects. Assumption **D** (iii) is not essential and is introduced to simplify the exposition of the limit result. The estimation methodology continues to work whether $\mathbf{z}_{it}$ and $\mathbf{w}_{it}$ are continuous variables or mixed with discrete variables. When mixed data are of concern, one can construct product kernels differently as explained in Li and Racine (2007, Ch.4).

Assumption S (*Curve Smoothness*).

(i) Define $s_{it} \equiv \mathbf{w}_{it}'\boldsymbol{\gamma} + \xi_i$. The unknown function $\Lambda(s_{i1}, s_{i2}, \boldsymbol{\zeta}_i) \equiv \mathbb{E}[u_{it}|d_{i1} = d_{i2} = 1, \boldsymbol{\zeta}_i] = \mathbb{E}[u_{it}|e_{i1} \leq s_{i1}, e_{i2} \leq s_{i2}, \boldsymbol{\zeta}_i]$ satisfies $\Lambda(s_{i1}, s_{i2}, \boldsymbol{\zeta}_i) - \Lambda(s_{i2}, s_{i1}, \boldsymbol{\zeta}_i) = d(s_{i1}, s_{i2}, \boldsymbol{\zeta}_i)(s_{i1} - s_{i2})$, and $d(s_{i1}, s_{i2}, \boldsymbol{\zeta}_i)$ has bounded and uniformly continuous partial derivatives up to the second order with respect to its first two arguments;

(ii) $\boldsymbol{\beta}(z)$ and $f_t(z)$ for $t = 1, 2$ have bounded and uniformly continuous partial derivatives up to the second order around the interior point of interest z;

(iii) $f_t(s,z)$, $m_t(s,z)$, $g_t(s,z)$, $\varkappa_t(s,z)$, and $\varphi_t(s,z)$ all have bounded and uniformly continuous partial derivatives up to the second order around the point $(0,z)$, where $m_t(s,z) \equiv \mathbb{E}[\phi_i \mathbf{x}_{it}\mathbf{x}_{it}'|\Delta s_i = s, \mathbf{z}_{it} = z]$, $g_t(s,z) \equiv \mathbb{E}[\phi_i(\mathbf{x}_{it}'\mathbf{x}_{it})^2|\Delta s_i = s, \mathbf{z}_{it} = z]$, and $\varkappa_t(s,z) \equiv \mathbb{E}\left[\phi_i|\Lambda(s_{i1}, s_{i2}, \boldsymbol{\zeta}_i)|^{i_1}\|\mathbf{x}_{it}\|^{i_2} \mid \Delta s_i = s, \mathbf{z}_{it} = z\right]$ for non-negative integers $i_1 \leq 1$ and $i_2 \leq 1$ and $i_1 + i_2 \leq 2$, and $\varphi_t(s,z) \equiv \mathbb{E}\left[\phi_i(\mathbf{x}_{it}'\mathbf{x}_{it})^{1+\delta/2} \mid \Delta s_i = s, \mathbf{z}_{it} = z\right]$ for some $\delta > 0$.

Assumption **S** (i) is an essential assumption that is required to remove the selection bias asymptotically. That is, the impact of λ_{it} (for all i and t) on the estimation of $\boldsymbol{\beta}(z)$ is asymptotically ignorable. Kyriazidou's (1997) "conditional exchangeability" assumption, i.e., $F(u_{i1}, u_{i2}, e_{i1}, e_{i2}|\boldsymbol{\zeta}_i) = F(u_{i2}, u_{i1}, e_{i2}, e_{i1}|\boldsymbol{\zeta}_i)$, implies Assumption **S** (i), and the latter is sufficient to asymptotically remove the endogenous selection bias $E_N^{-1}C_N$ in (3.14).

Assumption K (*Kernel Function*). The univariate kernel function $k(v)$ is a uniformly bounded, symmetric (around zero) probability density function with a compact support $[-1, 1]$, and $\mathcal{K}(v) = \prod_{s=1}^{q} k(v_s)$ is a product kernel.

Assumption B (*Bandwidth*). As $N \to \infty$, $h_0 \to 0$, $\|\mathbf{H}\| \to 0$ and $N|\mathbf{H}|h_0 \to \infty$, where $|\mathbf{H}| = h_1 \cdots h_q$ and $\|\mathbf{H}\| = \sqrt{\sum_{j=1}^{q} h_j^2}$.

While it is not necessary to use a kernel function with bounded support, we impose it to economize the proof. The bandwidth assumption ensures the consistency of the proposed estimator. However, one needs to undersmooth to remove the asymptotic bias in the limit distribution.

Assumption E (*Existence of the Estimator*).

$\boldsymbol{\Psi}(z) \equiv \lim_{N\to\infty}(|\mathbf{H}|h_0)^{-1}\sum_{t=1}^{2}\mathbb{E}\left[\phi_i\widehat{\psi}_i(1-\varpi_{it})\pi_{it}\mathbf{x}_{it}\mathbf{x}_{it}'\right]$ and

$\boldsymbol{\Gamma}(z) \equiv \lim_{N\to\infty}(|\mathbf{H}|h_0)^{-1}\sum_{t=1}^{2}\mathbb{E}\left[\phi_i\widehat{\psi}_i^2(1-\varpi_{it})^2\pi_{it}^2\mathbf{x}_{it}\mathbf{x}_{it}'\right]$ are both nonsingular $p \times p$ matrices,

where $\pi_{it} \equiv \mathcal{K}_h(\mathbf{z}_{it}, z) = \prod_{j=1}^{q} k((\mathbf{z}_{it} - z)/h_i)$ and $\varpi_{it} = \pi_{it}/\sum_{t=1}^{2}\pi_{it} \in (0,1)$ for all i and t.

Assumption **E** ensures the numerical feasibility of the proposed estimator. Since ϖ_{it} contains a random denominator, as in Sun et al. (2009), we cannot obtain a closed-form expression for $\boldsymbol{\Psi}(z)$ and $\boldsymbol{\Gamma}(z)$. The following theorem gives the asymptotic normality result for $\widehat{\boldsymbol{\beta}}(z)$.

Theorem 1 *Under Assumptions **D**, **S**, **K** and **B** and assuming that $\mathbb{E}|v_{it}|^{2+\delta} < \infty$ for some $\delta > 0$ and that $\sqrt{N|\mathbf{H}|h_0^3} = o(1)$ and $\sqrt{N|\mathbf{H}|h_0}\|\mathbf{H}\|^2 = o(1)$ as $N \to \infty$, for an interior point z*

$$\sqrt{N|\mathbf{H}|h_0}\left[\widehat{\boldsymbol{\beta}}(z) - \boldsymbol{\beta}(z)\right] \xrightarrow{d} \mathbb{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\beta(z)}\right) \ ,$$

where $\boldsymbol{\Sigma}_{\beta(z)} = \sigma_v^2\boldsymbol{\Psi}(z)^{-1}\boldsymbol{\Gamma}(z)\boldsymbol{\Psi}(z)^{-1}$. Moreover, a consistent estimator for $\boldsymbol{\Sigma}_{\beta(z)}$ is given by

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_{\beta(z)} &= \mathbf{S}_p\widehat{\boldsymbol{\Omega}}(z)^{-1}\widehat{\mathbf{J}}(z)\widehat{\boldsymbol{\Omega}}(z)^{-1}\mathbf{S}_p' \xrightarrow{p} \boldsymbol{\Sigma}_{\beta(z)} \\
\widehat{\boldsymbol{\Omega}}(z) &= (N|\mathbf{H}|h_0)^{-1}\mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\mathcal{X}(z) \\
\widehat{\mathbf{J}}(z) &= (N|\mathbf{H}|h_0)^{-1}\mathcal{X}(z)'\widehat{\mathbf{W}}_h(z)\widehat{\mathbf{V}}\widehat{\mathbf{V}}'\widehat{\mathbf{W}}_h(z)\mathcal{X}(z) \ ,
\end{aligned}$$

where $\mathbf{S}_p$ includes the first p rows of the $p(q+1)$ identify matrix, and a typical element of the $2N \times 1$ vector $\widehat{\mathbf{V}}$ equals $\widehat{v}_{it} = y_{it} - \mathbf{x}_{it}'\widehat{\boldsymbol{\beta}}(\mathbf{z}_{it})$.

For the proof of Theorem 1, see Appendix A. In the proof, we show that the nonparametric approximation bias term $\mathbf{S}_p E_N^{-1}A_N$ is of order $O_e\left(\|\mathbf{H}\|^2 + h_0^2\right)$, while the endogenous selection bias

term $\mathbf{S}_p E_N^{-1} C_N$ is of order $O_e(h_0)$.[16] Therefore, $\sqrt{Nh_0|\mathbf{H}|}\|\mathbf{H}\|^2 = o(1)$ and $\sqrt{N|\mathbf{H}|h_0^3} = o(1)$ are imposed to ensure a zero asymptotic bias for $\widehat{\boldsymbol{\beta}}(z)$. Consider an example when $h_j = cn^{-\delta}$ for $j = 0, 1, \ldots, q$. Then, Assumption **D**, $\sqrt{Nh_0|\mathbf{H}|}\|\mathbf{H}\|^2 = o(1)$ and $\sqrt{N|\mathbf{H}|h_0^3} = o(1)$ require that $(3+q)^{-1} < \delta < (1+q)^{-1}$.

Theorem 1 studies the limit results of our proposed estimator $\widehat{\boldsymbol{\beta}}(z)$ when $\boldsymbol{\gamma}$ is assumed to be known. Theorem 2 below shows that these results continue to hold under some reasonable conditions, when $\boldsymbol{\gamma}$ is *unknown*. In particular, we need the following additional assumptions.

Assumption G (*Existence of First-Step Estimator*). For any given parameter value $\boldsymbol{\gamma} \in \mathcal{A}$, there exists an estimator $\widehat{\boldsymbol{\gamma}}_N$ such that $\sup_{\boldsymbol{\gamma}\in\mathcal{A}} \|\widehat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}\| = O_p(N^{-\eta})$, where $\mathcal{A}$ is a compact subset of $\mathbb{R}^l$ and $\eta \in (0, 1/2]$.

Assumption K2 (*Kernel Function*). The univariate kernel function $k(v)$ is continuously differentiable over its compact support [-1,1].

Assumption S2 (*Curve Smoothness*).

(i) $\mathbb{E}\left[\phi_i \|\Delta \mathbf{w}_i\| \mathbf{x}_{it}' \mathbf{x}_{it} | \mathbf{z}_{it} = z\right]$ and $\mathbb{E}\left[\phi_i \|\Delta \mathbf{w}_i\| \ \|\mathbf{x}_{it}\| | \mathbf{z}_{it} = z\right]$ have bounded and uniformly continuous partial derivatives up to the second order around the point of interest z;

(ii) $\mathbb{E}\left[\phi_i \|\Delta \mathbf{w}_i\|^{i_1} \|\mathbf{x}_{it}\|^{i_2} | \Delta s_i = s, \mathbf{z}_{it} = z\right]$ for $i_1 \le 2$ and $i_2 \le 2$ and
$\mathbb{E}\left[\phi_i \|\mathbf{w}_{it}\|^2 \|\mathbf{x}_{it}\| \mid \Delta s_i = s, \mathbf{z}_{it} = z\right]$ have bounded and uniformly continuous partial derivatives up to the second order around $(0, z)$;

(iii) $\Lambda(s_{i1}, s_{i2}, \boldsymbol{\zeta}_i)$ is continuously differentiable; and $\Lambda(s_{i1}, s_{i2}, \boldsymbol{\zeta}_i)$ and its first-order partial derivatives with respect to its first two arguments are all uniformly bounded over its domain.

Theorem 2 *Under the assumptions given in Theorem 1, Assumptions **G**, **K2** and **S2** and $\|\widehat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}\|/h_0^2 = o_p(1)$, the limit result in Theorem 1 continues to hold.*

The proof of Theorem 2 is provided in Appendix B. Let $h_0 \sim N^{-\alpha_0}$ and $h_j \sim N^{-\alpha_1}$ with $\alpha_0 = \vartheta a_1$ for some $\vartheta > 0$. Then, $\|\widehat{\boldsymbol{\gamma}}_N - \boldsymbol{\gamma}\|/h_0^2 = o_p(1)$ holds if $\eta > 2\alpha_0$, which implies $\alpha_0 \in (0, 1/4)$ as $\eta \le 1/2$. The conditions that $N|\mathbf{H}|h_0 \to \infty$, $\sqrt{N|\mathbf{H}|h_0}\|\mathbf{H}\|^2 \to 0$ and $\sqrt{N|\mathbf{H}|h_0^3} \to 0$ as $N \to \infty$ imply that $\max\{1-(q+4)\alpha_1, (1-q\alpha_1)/3\} < \alpha_0 < 1-q\alpha_1$. We then obtain $2\vartheta/\min\{q+3\vartheta, q+4+\vartheta\} < \eta \le 1/2$, which implies $0 < \vartheta < \min\{q, (q+4)/3\}$. Hence, h_0 and h_j can be of the same order if and only if $q \ge 2$; and h_0 converges to zero at a slower speed than h_1 when $q = 1$.

5 Monte Carlo Study

In order to study the finite sample performance of our estimator (3.13), we conduct some Monte Carlo simulations. We use the following data generating process (DGP) for a model under *binary sample selection* [of the form in (2.1)]:

$$
\begin{aligned}
y_{it} &= \begin{cases} x_{it}\beta(z_{it}) + \mu_i + u_{it} & \text{if} \quad d_{it} = 1 \\ - & \text{otherwise} \end{cases} \\
d_{it}^* &= w_{1,it}\gamma_1 + w_{2,it}\gamma_2 + \xi_i + e_{it} \\
d_{it} &= \mathbb{1}\{d_{it}^* \ge 0\} \ ,
\end{aligned} \tag{5.1}
$$

where the varying coefficient in the outcome equation is $\beta(z_{it}) = \sin(\pi z_{it})$ and the constant coefficients in the selection equation are $\gamma_1 = \gamma_2 = 1$. The exogenous covariates are generated as

[16] We use $A_N = O_e(a_N)$ to denote $A_N = O_p(a_N)$ but not $A_N = o_p(a_N)$, where $a_N > 0$ is a sequence of constants.

follows: $(w_{1,it}, w_{2,it}) \sim i.i.d.\ \mathbb{N}(0,1)$, $x_{it} = w_{1,it}$ and $z_{it} \sim i.i.d.\ \mathbb{U}(0, 0.5\pi)$. The fixed effects are $\xi_i = \overline{w}_{1,i} + \overline{w}_{2,i} + 0.5\varsigma_i$ and $\mu_i = \overline{x}_i + \overline{z}_i + \varrho_i - 0.5 - 0.25\pi$,[17] where $(\varsigma_i, \varrho_i) \sim i.i.d.\ \mathbb{U}(0,1)$ and $\overline{w}_{1,i} = T^{-1}\sum_{t=1}^{T} w_{1,it}$ is the time average of $w_{1,it}$ with similarly defined $(\overline{w}_{2,i}, \overline{x}_i, \overline{z}_i)$. The error in the selection equation e_{it} is $i.i.d$ logistically distributed with location and scale parameters set to zero and one, respectively; the error in the outcome equation is $u_{it} = -e_{it} + \zeta_{it}$, where $\zeta_{it} \sim i.i.d.\ \mathbb{N}(0,1)$. Both errors share a non-zero correlation by design. Throughout this section, we use second-order Gaussian kernels; the bandwidths are selected via cross-validation as described in Section 3.

We set $T = 2$ so that there is only one pair of time periods for each cross-section i to consider. The above DGP implies that, on average, $\Pr[d_{i1} + d_{i2} = 1] \approx 0.37$ and $\Pr[d_{i1} = d_{i2} = 1] \approx 0.35$, i.e., about 37 and 35 percent of the sample is used in the first and second stages, respectively. Note that (i) cross-sections that "switch" their selection status only are used in the first-stage conditional logit estimation, and (ii) we can estimate the outcome equation only for those individuals who are selected into the sample *at least* in two periods.[18] We consider sample sizes $N = \{100,\ 200,\ 400\}$. For each N, we replicate the design 500 times.

Table 1 reports the results of the experiment. We study the finite sample performance of our estimator (3.13) in comparison to a "naive" varying coefficient model with fixed effects that ignores endogenous selection (labelled "A"). The "naive" estimator is that of Sun et al. (2009), which is likely to produce inconsistent estimates of the unknown coefficient function $\beta(\cdot)$ due to its inability to account for the presence of sample selection effects. It is convenient to think of this estimator as a limiting case of our estimator (3.13) with the bandwidth h_0 (for a single index in the selection equation $\Delta\mathbf{w}_i'\widehat{\boldsymbol{\gamma}}$) equal to infinity. In order to gauge the sensitivity of our estimator's performance to a sampling error in the first-stage estimates, we also re-estimate our model using true values of γ_1 and γ_2 (labelled "B"), i.e., with the first stage skipped. Our proposed two-stage estimator (3.13) is labelled "C".

[*insert Table 1*]

For each estimator in each simulation, we compute the root mean squared error (RMSE):

$$RMSE = \left(\frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{\phi}_i d_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} \widetilde{\phi}_i d_{it} \left(\beta(z_{it}) - \widehat{\beta}(z_{it}) \right)^2 \right)^{1/2}, \tag{5.2}$$

where $\widehat{\beta}(\cdot)$ is the estimate of $\beta(\cdot)$ from either of the three estimators A, B and C; $\widetilde{\phi}_i \equiv \mathbb{1}\left\{\sum_{t=1}^{T} d_{it} \geq 2\right\}$ "picks" cross-sections that are selected into the sample at least in two periods.

We summarize the results in Figure 1, where we plot distributions (across simulations) of the RMSE for each estimator and each sample size in the form of boxplots. We also report the averages of the RMSE computed over 500 simulations in Table 1. The results show that our estimator (C) is less biased than a "naive" estimator (A), which ignores endogenous selection. Comparing estimators B and C, we find that the results do not change dramatically if we use true or estimated values of γ_1 and γ_2; the first-stage estimation seems to not distort the results obtained in the second stage.[19] We also observe that the estimation (across all three estimators) becomes more stable as the sample size increases.

[*insert Figure 1*]

We next examine the finite sample performance of our estimator in the presence of *polychotomous switching*. Specifically, we consider $R = 3$. To make the experiment even more general, we set $T = 3$

[17] μ_i is generated so that $\mathbb{E}[\mu_i] = 0$.

[18] Exactly in two periods in this experiment, because $T = 2$.

[19] Kyriazidou (1997) documents a similar finding in a completely parametric formulation of our model (2.1).

(the case of $T > 2$). The DGP used is as follows [in line with (2.2)].

$$
\begin{aligned}
y_{r,it} &= \begin{cases} x_{r,it}\beta_r(z_{r,it}) + \mu_{r,i} + u_{r,it} & \text{if} \quad d_{r,it} = 1 \\ - & \text{otherwise} \end{cases} \\
d^*_{r,it} &= w_{1,it}\gamma_{r,1} + w_{2,it}\gamma_{r,2} + \xi_{r,i} + e_{r,it} \\
d_{r,it} &= \mathbb{1}\{d^*_{r,it} \geq \max_{j=1,\ldots,R;\ j\neq r} \{d^*_{j,it}\}\} \ ,
\end{aligned} \tag{5.3}
$$

where the varying coefficients in the outcome equations are $\beta_1(z_{1,it}) = \sin(\pi z_{1,it})$, $\beta_2(z_{2,it}) = 1 + z_{2,it} + (z_{2,it})^2$ and $\beta_3(z_{3,it}) = 1 + (z_{3,it})^3$ for regimes 1, 2 and 3, respectively. The constant coefficients in each of the three selection equations are $\gamma_{r,1} = \gamma_{r,2} = 1$ for $r = 1,\ 2,\ 3$. The exogenous covariates are generated as follows: $(w_{1,it}, w_{2,it}) \sim i.i.d.\ \mathbb{N}(0,1)$, $x_{r,it} = w_{1,it}$ and $z_{r,it} \sim i.i.d.\ \mathbb{U}(0, 0.5\pi)$ for $r = 1,\ 2,\ 3$. The fixed effects are $\xi_{r,i} = \overline{w}_{1,i} + \overline{w}_{2,i} + 0.5\varsigma_{r,i}$ and $\mu_{r,i} = \overline{x}_{r,i} + \overline{z}_{r,i} + \varrho_{r,i} - 0.5 - 0.25\pi$, where $(\varsigma_{r,i}, \varrho_{r,i}) \sim i.i.d.\ \mathbb{U}(0,1)$ for $r = 1,\ 2,\ 3$. The error terms in the selection equations $e_{r,it}$ are $i.i.d$ the type I extreme-value (Gumbel) distributed with location and scale parameters set to zero and one, respectively. The disturbances in the outcome equations are generated as $u_{r,it} = -e_{r,it} + \zeta_{r,it}$, where $\zeta_{r,it} \sim i.i.d.\ \mathbb{N}(0,1)$ for $r = 1,\ 2,\ 3$. We consider sample sizes $N = \{150,\ 300,\ 600\}$, for each of which we simulate the design 500 times.

[*insert Figure 2*]

Since $T = 3$ in this design, in the second stage we estimate (3.13) for $\mathcal{C}(3,2) = 3$ unique pairs of the time periods and then average $\widehat{\beta}_r(\cdot)$ for each $z_{r,it}$, as discussed in Section 3.[20] Also, note that the second stage is estimated for each regime separately. Figure 2 and Table 2 summarize the results from the three estimators, for each of the three regimes. The results are similar to those obtained in the case of binary sample selection. We find our estimator (C) to be less biased than a "naive" estimator (A) across all three regimes. The results do not seem to be sensitive to whether we use true or estimated values of the parameters in the selection equations (compare RMSE for estimators B and C). Importantly, the estimation becomes more stable as the sample size increases.

[*insert Table 2*]

6 Empirical Application: the Case of U.S. Credit Unions

In this section, we investigate the U.S. retail credit union production technologies in the period from 2002 to 2006. Using our proposed estimator (3.13), we are able to produce more robust estimates of credit union production technologies by controlling for (i) parameter heterogeneity in the cost function across credit unions of different sizes, (ii) endogenous selectivity as represented by differing service menus offered by credit unions and (iii) unobserved credit-union-specific heterogeneity. Before we proceed, we note that the notation used in this section has no connection to that in previous sections, unless specified otherwise.

6.1 Framework and Data Description

Given that, due to their cooperative nature, credit unions are not profit-maximizers, researchers usually think of them as maximizing service provision to their members in terms of quantity, price

[20] In order to facilitate comparability of the results across estimators, we similarly estimate the "naive" estimator of Sun et al. (2009) using the three unique pairs of the time periods separately and then averaging the obtained estimates for each $z_{r,it}$. However, one can, technically, estimate it using all "selected" observations at once.

and variety of services (Smith, 1984). We therefore follow a common practice in the credit union literature (Frame et al., 2003; Wheelock and Wilson, 2011, 2013) and adopt a "service provision approach." According to this framework, given the type of their production technology,[21] credit unions minimize non-interest, variable cost subject to the levels and types of services (outputs), the competitive prices of variable inputs and the levels of quasi-fixed netputs.

We consider four types of financial services that credit unions offer to their customers: real estate loans ($y1$), business and agricultural loans ($y2$), consumer loans ($y3$) and investments ($y4$). These are output quantities. We further follow Frame et al. (2003) and Wheelock and Wilson (2011, 2013) and include two quasi-fixed netputs to capture the price dimension of the service provision by credit unions: the average interest rate on saving deposits ($\widetilde{y}5$) and (the inverse of) the average interest rate on loans ($\widetilde{y}6$). The motivation here is to capture the cooperative nature of credit unions that, among other things, seek to offer the highest deposit rates and lowest loan rates possible to their members. Like Wheelock and Wilson (2011, 2013), one thus may prefer thinking of these price variables as quasi-fixed outputs. We therefore consider the *inverse* of the average interest rate on loans to enforce positive monotonicity (in outputs) of the cost function. Like Frame et al. (2003), we define two variable inputs: financial capital ($x1$) and labor ($x2$) with the vector of corresponding competitive prices $\mathbf{w} = (w1, w2)$. All of the above variables are taken as arguments of the dual variable, non-interest cost function of a credit union.

As pointed out in the Introduction, the data on credit unions contain a large number of observations for which the reported values of some types of services are zeros, which indicates the presence of significant differences among credit unions in terms of the service menu they offer to members. Ignoring this observed heterogeneity in the provision of services amounts to making a strong and rather unrealistic assumption that all credit unions share the same technology that is invariant to the menu of services they provide. This assumption of homogeneous technology across credit unions is likely to result in the loss of information and the misspecification of the econometric model, which is further aggravated if the choice of the differing service menus by credit unions is endogenous. Malikov et al. (2013) document that the overwhelming majority of U.S. retail, or so-called natural-person, credit unions (more than 99%) offer one of the following three service menus to their members: (i) consumer loans and investments ($y3,\ y4$); (ii) real estate and consumer loans as well as investments ($y1,\ y3,\ y4$); and (iii) all types of services: real estate, business and consumer loans, and investments ($y1,\ y2,\ y3,\ y4$). We label these service menus (output mixes) as "1", "2" and "3", respectively, and refer to corresponding credit unions as "Type 1", "Type 2" and "Type 3". We hereafter use credit union and service menu types interchangeably when referring to credit unions and their production technologies.

The data come from year-end Call Reports available from the National Credit Union Administration (NCUA), a federal regulatory body that supervises all state and federally chartered credit unions in the U.S. In this study, we focus on the period prior to the 2008 financial crisis, in an attempt to minimize the influence of potential structural changes in the industry during the crisis and in its aftermath on the estimation results. In particular, we consider a five-year period from 2002 and 2006. We focus on retail credit unions only[22] and therefore exclude corporate credit unions (whose customers are the retail credit unions) from the sample to minimize noise in the data due to apparent non-homogeneity between these two types of unions.[23]

[21] That is, given the mix of financial services that credit unions choose to provide to their members.

[22] That is, we focus on retail credit unions of Types 1, 2 or 3. Credit unions that offer other service menus (less than a percent of observations) likely contain either outliers or reporting errors.

[23] We also discard observations with negative values of outputs and total cost. In addition, we exclude observations with non-positive values of variable input prices, quasi-fixed netputs, equity capital, total assets, reserves and total liabilities. Since $\widetilde{\mathbf{y}}$ and $w1$ are interest rates, we also eliminate those observations for which values of these variables

Recall that, in order to make our proposed estimator feasible, one needs (i) cross-sections to switch regimes (to estimate the selection equations in the first stage) and (ii) a given regime to be selected by a cross-sectional unit at least in two time periods (to estimate the outcome equation in the second stage). We therefore confine our analysis to credit unions that meet the above two requirements. Also, to avoid a potential impact of entries and exits, we examine only continuously operating credit unions. Lastly, given significant computational intensity of our proposed estimator (particularly, a cross-validation procedure in the second stage), we select a pseudo-random representative subsample of credit unions satisfying all above criteria, which renders a balanced panel of 500 units continuously observed over 5 years. The procedure does not significantly affect the distribution of key variables and the composure over credit union types.[24]

Table 3 reports summary statistics of the variables used in our analysis. We deflate all nominal stock variables to 2011 U.S. dollars using the GDP Implicit Price Deflator. A comparison of sample mean and median estimates of variables shows clear differences between the credit union types. As expected, the size of the credit unions (proxied either by total assets or the number of members) increases as one moves from Type 1 to Type 3. The dramatic differences between the three types favor our view that the assumption of homogeneous credit union technology *across* service menu types is likely to result in the loss of information and the misspecification of the econometric model. Moreover, credit unions technology is also unlikely to be homogenous *within* a given service menu type, which, if overlooked, can distort results as well.

[*insert Table 3*]

To put the problem of modeling credit union technologies into perspective of the estimator that we consider in this paper, there are three distinct types of retail credit unions, as defined by their differing service menus. These types are what we have referred to in the previous sections as polychotomous "regimes". Since there are no legal restrictions on which of the four financial services (outputs) a credit union may offer to its members, it is natural to view these credit union types as an outcome of endogenous decision-making. The data seem to suggest that the variables capturing a credit union's size, financial strength and potential for growth may be particularly relevant to a choice of the service menu. A careful examination of the credit union literature suggests considering the following variables: the number of current and potential members, equity capital,[25] reserves and the leverage ratio, defined as the ratio of total debt to total assets (Bauer, 2008; Bauer et al., 2009; Goddard et al., 2002, 2008).[26] These are the variables entering the selection equations. For their summary statistics, see Table 3.

Further, it has been argued in the literature that the size of a credit union (commonly measured by its total assets) matters considerably in shaping its cost structure and that any parametric specification of the cost function that overlooks this relationship is thus likely to suffer from parameter instability (Wheelock and Wilson, 2011). We concur with this sentiment and agree that it may be inappropriate to assume that the cost structure of a small credit union is the same as that of a large credit union. To accommodate this technological heterogeneity among credit unions of different sizes, we allow parameters of the *credit-union-type*-specific cost function to also be varying with (the log of) credit union's total assets. Such a specification yields *credit-union*-specific estimates of

lie outside the unit interval. These excluded observations are likely to be the result of erroneous data reporting. All variables are constructed following the instructions in the appendix of Malikov et al. (2013).

[24] We have tried numerous pseudo-random subsample, all of which yield qualitatively unchanged results. The composition of the sample by credit union types is 28%, 60% and 12% for Types 1, 2 and 3, respectively.

[25] We note that, since credit unions are mutual organizations, they cannot raise "equity" via public offering *per se*. The equity is instead raised by retaining earnings.

[26] For more on the motivation of choosing these variables, see Malikov et al. (2013).

the cost function parameters.

6.2 Estimation and Empirical Results

We consider a VC model of heterogeneous credit union production technologies with polychotomous endogenous switching and fixed effects in both the selection and outcome equations. In this paper, we assume that the credit-union-type-specific dual cost function takes a semiparametric analogue of the translog specification, under which parameters are unknown smooth functions of the size. We cast the model in the form of (2.2), i.e.,

$$\begin{aligned}\ln C^r_{it} = {} & \mu^r(z_{it}) + \eta^r_1(z_{it})\ t + \frac{1}{2}\eta^r_{11}(z_{it})\ t^2 + \\ & \sum_{m=1}^{M^r} \beta^r_m(z_{it}) \ln y^r_{m,it} + \frac{1}{2}\sum_{m=1}^{M^r}\sum_{h=1}^{M^r} \beta^r_{mh}(z_{it}) \ln y^r_{m,it} \ln y^r_{h,it} + \sum_{m=1}^{M^r} \rho^r_{m1}(z_{it}) \ln y^r_{m,it}\ t + \\ & \sum_{k=1}^{K} \omega^r_k(z_{it}) \ln \widetilde{y}_{k,it} + \frac{1}{2}\sum_{k=1}^{K}\sum_{g=1}^{K} \omega^r_{kg}(z_{it}) \ln \widetilde{y}_{k,it} \ln \widetilde{y}_{g,it} + \sum_{k=1}^{K} \rho^r_{k2}(z_{it}) \ln \widetilde{y}_{k,it}\ t + \\ & \sum_{j=1}^{J} \delta^r_j(z_{it}) \ln w_{j,it} + \frac{1}{2}\sum_{j=1}^{J}\sum_{s=1}^{J} \delta^r_{js}(z_{it}) \ln w_{j,it} \ln w_{s,it} + \sum_{j=1}^{J} \rho^r_{j3}(z_{it}) \ln w_{j,it}\ t + \\ & \sum_{m=1}^{M^r}\sum_{j=1}^{J} \theta^r_{mj}(z_{it}) \ln y^r_{m,it} \ln w_{j,it} + \sum_{k=1}^{K}\sum_{j=1}^{J} \vartheta^r_{kj}(z_{it}) \ln \widetilde{y}_{k,it} \ln w_{j,it} + \\ & \sum_{m=1}^{M^r}\sum_{k=1}^{K} \varrho^r_{mk}(z_{it}) \ln y^r_{m,it} \ln \widetilde{y}_{k,it} + \mu^r_i + u^r_{it} \qquad (\text{if} \quad d^r_{it} = 1) \end{aligned} \tag{6.1a}$$

$$d^{r*}_{it} = \sum_{h=1}^{H} \gamma^r_h \ln q_{h,it} + \xi^r_i + e^r_{it}\ , \qquad (i = 1, \ldots, N;\ t = 1, \ldots, T;\ R = 1, \ldots, R) \tag{6.1b}$$

where, for each credit union type $r = 1, \ldots, R$ ($R = 3$), C^r_{it} is the variable, non-interest cost; $y^r_{m,it} \in \mathbf{y}^r_{it}$ is the output specific to a given type of credit unions, i.e., $\mathbf{y}^1_{it} \equiv (y3_{it}, y4_{it})$, $\mathbf{y}^2_{it} \equiv (y1_{it}, y3_{it}, y4_{it})$, $\mathbf{y}^3_{it} \equiv (y1_{it}, y2_{it}, y3_{it}, y4_{it})$ with the corresponding values of $M^r = \{2,\ 3,\ 4\}$. The variable input prices $w_{j,it} \in (w1, w2)$, quasi-fixed netputs $\widetilde{y}_{k,it} \in (\widetilde{y}5, \widetilde{y}6)$ and the log of total assets z_{it} are invariant to credit union type and thus do not have superscript r (also, $J = K = 2$). To capture temporal changes in the cost frontiers, we also include the time trend t in (6.1a). The rth cost function is observed if a credit union selects the rth type of the service mix, as captured by the binary indicator d^r_{it}. The selection is governed by (6.1b) that assumes that the propensity to select the rth service mix type is a function of $\mathbf{q}_{it}$ that includes the number of current and potential members, equity capital, reserves and the leverage ratio ($H = 5$). We control for unobserved unit-specific heterogeneity among credit unions by including fixed effects μ^r_i and ξ^r_i in the cost and selection equations, respectively.

We estimate the model in two stages as outlined in Section 3. In the first stage, we transform (6.1b) into its binary selection analogue as described in Section 2.2, which is then estimated via Chamberlain's (1980) conditional multinomial logit. We use the obtained estimates of parameters $\widehat{\boldsymbol{\gamma}}^r$ in the second stage, in which we apply our proposed estimator (3.13) onto the cost functions (6.1a) for each of the credit union types separately. Since the design of our estimator is such that

it uses pairs of the time periods, we consider $\mathcal{C}(5,2)$ unique pairs, the results for which we then average, as described in Section 3.[27]

In order to analyze to what extent the results are distorted if (i) one assumes that selection of the credit union type is "ignorable" and thus the selectivity bias need not be accounted for or (ii) a homogeneous cost function is fitted for all credit unions of a given type under the assumption of parameters in (6.1a) being constant, we also estimate two auxiliary models. Clearly, both models are special cases of the one we consider in this paper, as discussed in Section 2. For the ease of discussion, below we define all three models we estimate.

Model I. The semiparametric varying coefficient model with endogenous selection and fixed effects; given by (6.1) and estimated via our proposed estimator (3.13) in two stages.

Model II. The semiparametric varying coefficient model with fixed effects under the assumption of "ignorable" (exogenous) selection; estimated in one stage via Sun et al.'s (2009) estimator applied onto (6.1a). Any differences between models I and II are likely due to selectivity bias in the latter.

Model III. The parametric model with endogenous selection and fixed effects. This model assumes that credit union technology is homogeneous within a given credit union type, i.e., parameter functions in the cost function (6.1a) are assumed to be constant across credit unions. The model is estimated in two stages via Kyriazidou's (1997) estimator. Any differences between models I and III are likely due to parametric misspecification in model III.

In each model, we impose the symmetry and linear homogeneity (in input prices) restrictions onto the cost functions. The homogeneity is imposed by dividing the variable cost and input prices by the price of labor ($w2$). Given the complexity of all three estimators, it is unlikely to expect the fitted cost functions to properly satisfy all monotonicity properties, which may result in misleading results. Therefore, we also impose positive monotonicity in input prices and outputs (including quasi-fixed outputs) onto the cost functions. We do so *post*-estimation via quadratic programming technique as proposed by Hall and Huang (2001) and Du et al. (2013). The idea is to reweigh observations used in estimation so that all constraints are observation-wise satisfied. In this paper, we follow Du et al. (2013) whose method allows weights to be non-positive which has some desirable implications. Although the above method is developed to be applied in a nonparametric setting [models I and II], it can be easily extended to a parametric specification [model III].[28] To conserve space, we do not report the results from the first stage and directly proceed to the discussion of the main results from the cost functions.

[*insert Tables 4.1-4.3*]

Tables 4.1-4.3 report summary statistics of elasticity estimates [derivatives of the cost function in (6.1a) with respect to the covariates] obtained from models I through III for each of the three credit union types. As expected, the results are more similar across flexible semiparametric models I and II, than across the latter two and the parametric model III. This can be clearly seen from Figures 3.1-3.3 that plot kernel densities of these elasticity estimates. The densities are constructed using the Gaussian kernel and Silverman's (1986) "rule-of-thumb" bandwidth.

[*insert Figures 3.1-3.3*]

In the case of the Type 1 credit unions, the empirical evidence indicate the presence of economically negligible selectivity bias which is suggested by little differences between the results from

[27] We do the same when estimating models II and III described below. While the procedure is quite natural in the case of model III, we note that model II can be estimated using the entire sample period at once. We however opt to use pairs of time periods, in order to facilitate comparability of the results across the three models.

[28] For more details on constrained estimation in the case of nonparametric regression, see Du et al. (2013).

models I and II (see Figure 3.1). In particular, using the Li (1996) test, we fail to reject the null of equality of the two densities in the cases of the input price elasticity (with respect to $w1$) and the technical change [elasticity of the cost function in (6.1a) with respect to t]: the bootstrap p-values are 0.505 and 0.192, respectively. The differences between the models are however amplified when investigating the production of the Type 2 credit unions. When compared to a benchmark model I, we find relatively large negative biases in all output elasticities produced by model II. We attribute these differences to the sample selection effects that the latter model takes for granted. For instance, the median estimate of output elasticity for $y3$ (consumer loans) from model II is reported to be 27% less than that from model I: 0.082 vs. 0.113 (see Table 4.2). The sign of the selection bias in the estimates from model II however generally changes in the case of the input price elasticity. The elasticity estimate densities are consistently statistically different across the two models with p-values less than 10^{-6}. For credit unions of Type 3, we similarly find evidence of negative selection biases in output elasticities with respect to $y1$ and $y4$ (see Figure 3.3). However, we find that the kernel densities of the remaining elasticities are statistically equal at the 5% significance level.

A comparison of the estimates produced by model I and its parametric counterpart III enables us to analyze the implications of imposing parameter homogeneity onto credit union technology as done by model III. The differences in elasticity estimates are striking across all three credit union types. For instance, model III tends to over-estimate output elasticity for $y3$ and under-estimate the elasticity in the case of another output $y4$ in the case of Type 1 credit unions, whereas the biases in output elasticity estimates are all uniformly positive for Type 2 credit unions. Notably in the case of credit unions of Type 3, the results appear to be less distorted around the medians of distributions of output elasticities due to a smaller variation in the elasticity estimates (over credit unions) from model III. The latter likely results from parameter homogeneity implied by parametric model III. Expectedly, the densities are all statistically different across models I and III.

As noted earlier, it has been argued in the literature that the size of a credit union matters considerably in shaping its cost structure and that any parametric specification of the cost function will overlook this relationship. It is therefore of interest to compare the relationship between estimated credit union technologies and the asset size of credit unions implied by both models I and III. Recall that the underlying difference lies in the fact that the former model explicitly recognizes the above relationship, while model III does not. We compare the two models by looking at the estimates of scale economies, computed as one minus the sum of output elasticities. The defined measure is intuitive because its positive (negative) value indicates the presence of the scale economies (diseconomies). We scatterplot the estimates of scale economies against the log of total assets for all three credit union technology types in Figure 4. The figure also graphs the fitted (kernel) local-constant mean of these estimates given the asset size.[29]

[*insert Figure 4*]

We find significant differences across the two models. The parametric model III generally suggests a negative relationship between the scale economies and the overall size of a credit union, a pattern that one would normally expect to see in the data. The relationship is quite strong for credit unions of Types 1 and 2, whereas no clear pattern is detected among Type 3 credit unions. Based on the semiparametric model I, we however find evidence in favor of a more nonlinear (inverted-U-shaped) relationship between scale economies and the size. In particular, the results from model I suggest that scale economies tend to increase in the first stages of a Type 1 credit union's growth,

[29] At first glance, it may seem that there are little differences across the three types of credit unions in terms of the asset size as indicated by the range of values on the horizontal axis, which contradicts our findings in the previous subsection (see Table 3). However, Figure 4 plots scale economies against the asset size *scaled down* by it type-specific mean.

which seems to be quite puzzling. However, recall that credit unions of Type 1 tend to be small in size in general (see Table 3): about half of credit unions in this technology group are as small as an entity with no more than 2 full-time equivalent employees. Therefore, we may expect that, as these credit unions grow, so do their resources. An increase in available resources would enable credit unions to adopt new information technologies — internet banking, automated teller machines, electronic money systems and access to members' credit history through the credit rating bureaus — that initially result in large fixed costs but, once adopted, are substantial cost-savers. One would therefore expect to see scale economies to be increasing through the early expansion of unions and, as the impact of the above financial constraints wears out, to eventually start declining as credit unions continue to grow. The latter is exactly the pattern that we observed based on model I. Indeed, the scale economies continue to decline as credit unions evolve from Type 1 to Type 2. However, economies seem to be increasing yet again among smaller Type 3 credit unions. A greater diversification enjoyed by these (generally larger) credit unions is a potential explanation for this. The diversification is due to a growing number of members, a larger range of financial services as well as an opportunity to engage in more advanced financial operations (Wilcox, 2005, 2006). All three factors are conducive to a decline in average risk which, in turn, is likely to lead to a smaller average cost of screening, risk-monitoring and other risk management activities.[30]

The above findings call for extra caution when researchers first estimate a parametric model of credit union production technologies (even after controlling for selection into groups of different service menus) and then analyze how the estimated technological metrics change with the size of entities. Our findings indicate that the relationship with the asset size implied by such models [like model III] may deviate substantially from that predicted by models that engineer the relationship explicitly [like model I].

Lastly, we contrast the estimates of scale economies from models I-III qualitatively by testing for their statistical significance. That is, scale economies that are found to be less/equal to/greater than zero at the 95% significance level are informative of decreasing/constant/increasing returns to scale. Given the two-stage nature of all models as well as the presence of varying coefficients in models I and II, we use jackknife standard errors for the inference.[31] All models I-III provide overwhelming evidence in favor of scale economies across all credit unions of Type 1 and 2. In the case of Type 3 credit unions, model III similarly indicates scale economies uniformly enjoyed by all credit union in this category. However, based on model I and II we find that 38% and 24% of credit-union-years exhibit no scale economies (or sometimes scale diseconomies), respectively.

7 Conclusion

In this paper, we consider a flexible panel data sample selection model in which (i) the outcome equation is permitted to take a semiparametric VC form to capture potential parameter heterogeneity in the relationship of interest, (ii) both the outcome and selection equations contain unobserved fixed effects and (iii) selection is generalized to a polychotomous case.

We propose estimating this model in two stages. Given consistent parameter estimates of the selection equation obtained in the first stage, we estimate the VC outcome equation using data for observed individuals (cross-sections) whose estimated likelihood of being selected into the sample

[30] In fact, recent studies in banking also report scale economies that tend to increase with the size of entities (Hughes and Mester, 2013).

[31] We use the panel data extension of the traditional jackknife, in which a cross-section, rather than an observation, is deleted in each loop. Also, given the computational intensity of our and Sun et al.'s (2009) estimator, we perform the jackknife with random subsampling (for details, see Shao and Tu, 1995). We perform 100 iterations.

stays approximately the same over time. For such individuals, the sample selection bias would be approximately time-invariant and thus can be treated as another component of fixed effects present in the outcome equation. Given that there are unlikely to be many (if any at all) cross-sections with exactly the same selection probabilities over time, we adopt the idea of Ahn and Powell (1993) and Kyriazidou (1997) and weigh these cross-sections based on "closeness" of their respective selection probabilities (and thus their selectivity biases) to being the same over time. The weighted semiparametric outcome equation can then be estimated in a manner similar to that proposed by Sun et al. (2009). The selection bias term is then "asymptotically" removed from the equation along with fixed effects using kernel-based weights. We show that our proposed estimator is consistent and asymptotically normal.

We showcase our model by applying it to study heterogeneous production technologies of U.S. retail credit unions in the 2002-2006 period. However, the model we consider is not tailored to production analysis only. The framework can be applied to study numerous economic issues, where endogenous selectivity is of a concern and one desires to explore parameter heterogeneity in the relationship of interest. One such example would be a study of the wage differential and labor force participation that we have used to motivate our paper. We also note that, while in this paper we confine our analysis to the selection equations of a linear parametric form, it however may be generalized to a semiparametric form as well. We leave the latter for future research.

Appendix

A Proof of Theorem 1

Throughout the Appendix we use $A_N \approx B_N$ to denote that B_N is the leading term of A_N, i.e., $A_N = B_N + (s.o.)$, where $(s.o.)$ denotes terms having probability order smaller than that of B_N; we use $A_N \sim B_N$ to denote that A_N and B_N have exactly the same stochastic order. In addition, we use $A_N = O_e(a_N)$ to denote $A_N = O_p(a_N)$ but not $A_N = o_p(a_N)$, where $a_N > 0$ is a sequence of constants. We also use M to denote a generic positive constant which can take different values at different places. Moreover, let a and b be two $2N \times 1$ vectors. Then, a simple calculation gives

$$a'\widehat{W}_h(z)\, b = \sum_{i=1}^{N} \widehat{\psi}_i \sum_{t=1}^{2} \pi_{it} a_{it} b_{it} + \sum_{i=1}^{N}\sum_{j=1}^{N} \widehat{\psi}_i \widehat{\psi}_j q_{ij} \sum_{t=1}^{2} \pi_{it} a_{it} \sum_{s=1}^{2} \pi_{js} b_{js} \ . \tag{A.1}$$

By (3.14), the conditional bias and variance of $\mathrm{vec}\{\widehat{\Theta}(z)\}$ are given as follows:

$$Bias\left[\mathrm{vec}\left\{\widehat{\Theta}(z)\right\} |\phi_i, \zeta_i\right] = \left[\mathcal{X}(z)' \widehat{W}_h(z) \mathcal{X}(z)\right]^{-1} \mathcal{X}(z)' \widehat{W}_h(z) \left[\Pi(z)/2 + \lambda\right] \tag{A.2}$$

and, under Assumption **D** (iii),

$$Var\left[\mathrm{vec}\{\widehat{\Theta}(z)\}|\phi_i, \zeta_i\right] = \sigma_v^2 \left[\mathcal{X}(z)' \widehat{W}_h(z) \mathcal{X}(z)\right]^{-1} \left[\mathcal{X}(z)' \widehat{W}_h^2(z) \mathcal{X}(z)\right] \left[\mathcal{X}(z)' \widehat{W}_h(z) \mathcal{X}(z)\right]^{-1} . \tag{A.3}$$

As the proofs given in the Appendix closely follow those given in Sun et al. (2009), for ease of cross-reference, we introduce some notation used in Sun et al. (2009): $G_{it}(z, H) = \mathcal{D}_H \mathcal{Z}_{it}(z)$ and $[\cdot]_{it,js} = G_{it}(z, H) G_{js}(z, H)'$, where $\mathcal{D}_H = \mathrm{diag}\{1, h_1, \ldots, h_q\}$ is a $(q+1) \times (q+1)$ diagonal matrix, so that the $(l+1)$th element of $G_{it}(z, H)$ is $G_{it,l} = (z_{it,l} - z_l)/h_l$ for $l = 1, \cdots, q$. In addition, for each i and t, we denote $\pi_{it} \equiv \mathcal{K}_h(z_{it}, z)$ and $c_H(z_i, z)^{-1} = \pi_{i1} + \pi_{i2}$.

Lemma 1 *Under Assumptions* ***D****,* ***S*** *(ii),* ***K*** *and* ***B****, we have*

$$\begin{aligned}
&(N\,|H|\,h_0)^{-1}\,\mathcal{D}_H^{-1}\mathcal{X}\,(z)'\,\widehat{W}_h\,(z)\,\mathcal{X}\,(z)\,\mathcal{D}_H^{-1} \\
\approx\;& (|H|h_0)^{-1}\sum_{t=1}^{2}\mathbb{E}\left[(1-\varpi_{it})\phi_i\widehat{\psi}_i\pi_{it}[\cdot]_{it,it}\otimes(x_{it}x_{it}')\right]=O_e(1)\ , \qquad \text{(A.4)}
\end{aligned}$$

where $\varpi_{it}\equiv\pi_{it}/\sum_{t=1}^{2}\pi_{it}\in(0,1)$ *for all* i *and* t*.*

Proof: By (A.1) we have

$$\begin{aligned}
\mathcal{D}_H^{-1}E_N\mathcal{D}_H^{-1} &\equiv \mathcal{D}_H^{-1}\mathcal{X}\,(z)'\,\widehat{W}_h\,(z)\,\mathcal{X}\,(z)\,\mathcal{D}_H^{-1} \\
&= \sum_{i=1}^{N}\phi_i\widehat{\psi}_i\sum_{t=1}^{2}\pi_{it}[\cdot]_{it,it}\otimes\left(x_{it}x_{it}'\right)-\sum_{i=1}^{N}\phi_i\widehat{\psi}_iq_{ii}\sum_{t=1}^{2}\sum_{s=1}^{2}\pi_{it}\pi_{is}[\cdot]_{it,is}\otimes\left(x_{it}x_{is}'\right) \\
&\quad-\sum_{i=1}^{N}\phi_i\widehat{\psi}_i\sum_{j\neq i}^{N}\phi_j\widehat{\psi}_jq_{ji}\sum_{t=1}^{2}\sum_{s=1}^{2}\pi_{it}\pi_{js}[\cdot]_{it,js}\otimes\left(x_{it}x_{js}'\right) \\
&\approx \sum_{i=1}^{N}\phi_i\widehat{\psi}_i\sum_{t=1}^{2}(1-\omega_{it})\pi_{it}[\cdot]_{it,it}\otimes\left(x_{it}x_{it}'\right)\ ,
\end{aligned}$$

where we obtain the last line by following the proof of Lemma A.2 in Sun et al. (2009), and

$$q_{ii} = c_H\,(z_i,z)-c_H\,(z_i,z)^2\,/\sum_{i=1}^{N}c_H\,(z_i,z)\ , \qquad \text{(A.5)}$$

$$q_{ij} = -c_H\,(z_i,z)\,c_H\,(z_j,z)\,/\sum_{i=1}^{N}c_H\,(z_i,z)\quad\text{for}\quad i\neq j\ . \qquad \text{(A.6)}$$

First, we consider $\mathcal{A}_N\equiv(N\,|H|\,h_0)^{-1}\sum_{i=1}^{N}\phi_i\widehat{\psi}_i\sum_{t=1}^{2}\pi_{it}[\cdot]_{it,it}\otimes(x_{it}x_{it}')$. Letting $v=\Delta s_i/h_0$ and $\omega=H^{-1}\,(z_{it}-z)$ and applying the change of variables yield

$$\begin{aligned}
&(|H|\,h_0)^{-1}\sum_{t=1}^{2}\mathbb{E}\left[\phi_i\widehat{\psi}_i\pi_{it}[\cdot]_{it,it}\otimes\left(x_{it}x_{it}'\right)\right] \\
=\;& (|H|\,h_0)^{-1}\,\mathbb{E}\left\{\mathbb{E}\left[\phi_i[\cdot]_{it,it}\otimes\left(x_{it}x_{it}'\right)\middle|\,\Delta s_i,z_{it}\right]\pi_{it}\widehat{\psi}_i\right\} \\
=\;& \int\int k(v)\mathcal{K}(\omega)\mathbb{E}\left(\phi_i\begin{bmatrix}1 & \omega'\\ \omega & \omega\omega'\end{bmatrix}\otimes\left(x_{it}x_{it}'\right)\middle|\,vh_0,H\omega+z\right)f_t\,(vh_0,H\omega+z)\,dvd\omega \\
=\;& R_{K,2}\otimes\mathbb{E}\left(\phi_ix_{it}x_{it}'\middle|\,\Delta s_i=0,z_{it}=z\right)f_t(0,z)+o(1)\ ,
\end{aligned}$$

where $R_{K,2}=\mathrm{diag}\,\{1,\kappa_2,\ldots,\kappa_2\}$ is a $(q+1)\times(q+1)$ diagonal matrix and $\kappa_2=\int k(v)v^2dv$. Then, we have

$$\mathbb{E}\,(\mathcal{A}_N)=\sum_{t=1}^{2}\mathbb{E}\left[\phi_iR_{K,2}\otimes\left(x_{it}x_{it}'\right)\middle|\,0,z\right]f_t(0,z)+o(1)\ . \qquad \text{(A.7)}$$

Next, $Var\,(\mathcal{A}_N)$ is dominated by $2N^{-1}\,(|H|\,h_0)^{-2}\,\mathbb{E}\left[\phi_i\widehat{\psi}_i^2\sum_{t=1}^{2}\pi_{it}^2\,\|[\cdot]_{it,it}\otimes(x_{it}x_{it}')\|^2\right]$, where $\|\cdot\|$ denotes a Euclidian norm. As $\|A\|=\sqrt{\mathrm{tr}\,(AA')}$ and $\mathrm{tr}\,(A\otimes B)=\mathrm{tr}(A)\mathrm{tr}(B)$, we have

$$\left\|[\cdot]_{it,it}\otimes\left(x_{it}x_{it}'\right)\right\|=\left(1+\sum_{l=1}^{q}G_{it,l}^2\right)x_{it}'x_{it}\ . \qquad \text{(A.8)}$$

Applying the same method from above, we can show that $Var\left(\mathcal{A}_N\right) = O\left(\left(N\left|H\right|h_0\right)^{-1}\right)$ if $g_t(s,z) \equiv \mathbb{E}\left[\phi_i \left(x_{it}'x_{it}\right)^2 \left|\Delta s_i = s, z_{it} = z\right.\right]$ is continuous and bounded around point $(0,z)$ for $t = 1,2$. Hence, we obtain

$$A_N \approx \sum_{t=1}^{2} \mathbb{E}\left[\phi_i R_{K,2} \otimes \left(x_{it}x_{it}'\right)\middle| 0, z\right] f_t(0,z) + o_p(1) \ , \tag{A.9}$$

if $h_0 \to 0$, $\|H\| \to 0$, and $N\left|H\right|h_0 \to \infty$ as $N \to \infty$.

We cannot obtain a closed-form limit result for $\left(N\left|H\right|h_0\right)^{-1}\sum_{i=1}^{N}\phi_i\widehat{\psi}_i\sum_{t=1}^{2}\omega_{it}\pi_{it}[\cdot]_{it,it}\otimes(x_{it}x_{it}')$ as $\omega_{it} = \pi_{it}/\left(\pi_{i1}+\pi_{i2}\right)$ contains a random denominator. However, as ω_{it} always lies between 0 and 1, this term has the same stochastic order as A_N. This completes the proof of this lemma.

Lemma 2 *Under Assumptions **D**, **S**, **K** and **B**, we have*

$$\left(N\left|H\right|h_0\right)^{-1}\mathcal{D}_H^{-1}\mathcal{X}(z)'\widehat{W}_h(z)\lambda = O_e(h_0) \ . \tag{A.10}$$

Proof: By (A.1) we have

$$\begin{aligned}\mathcal{D}_H^{-1}C_N &\equiv \mathcal{D}_H^{-1}\mathcal{X}(z)'\widehat{W}_h(z)\lambda \\ &= \sum_{i=1}^{N}\phi_i\widehat{\psi}_i\sum_{t=1}^{2}\pi_{it}\left(G_{it}\otimes x_{it}\right)\left[\phi_i\widehat{\psi}_i\lambda_{it} - \sum_{j=1}^{N}\phi_j\widehat{\psi}_j\left(\pi_{j1}\lambda_{j1}+\pi_{j2}\lambda_{j2}\right)q_{ji}\right] \ .\end{aligned} \tag{A.11}$$

For $t \neq s$, Assumption **S** (i) means $\lambda_{js} = \lambda_{jt} + \Delta s_j d\left(s_{js}, s_{jt}, \zeta_j\right)$, and we obtain $\pi_{jt}\lambda_{jt} + \pi_{js}\lambda_{js} = \left(\pi_{jt}+\pi_{js}\right)\lambda_{jt} + \pi_{js}\Delta s_j d\left(s_{js}, s_{jt}, \zeta_j\right)$. Then, by $c_H\left(z_i, z\right)^{-1} = \pi_{i1}+\pi_{i2}$, (A.5) and (A.6), we have

$$\begin{aligned}&\phi_i\widehat{\psi}_i\lambda_{it} - \sum_{j=1}^{N}\phi_j\widehat{\psi}_j\left(\pi_{j1}\lambda_{j1}+\pi_{j2}\lambda_{j2}\right)q_{ji} \\ =\ & \phi_i\widehat{\psi}_i\lambda_{it} - \sum_{j=1}^{N}\frac{\phi_j\widehat{\psi}_j\lambda_{jt}q_{ji}}{c_H\left(z_j,z\right)} - \sum_{j=1}^{N}\phi_j\widehat{\psi}_j q_{ji}\Delta s_j d\left(s_{js}, s_{jt}, \zeta_j\right) \\ =\ & \phi_i\widehat{\psi}_i\lambda_{it}\frac{c_H\left(z_i,z\right)}{\sum_{i=1}^{N}c_H\left(z_i,z\right)} + \sum_{j\neq i}^{N}\phi_j\widehat{\psi}_j\lambda_{jt}\frac{c_H\left(z_i,z\right)}{\sum_{i=1}^{N}c_H\left(z_i,z\right)} \\ & - \sum_{j=1}^{N}\phi_j\widehat{\psi}_j q_{ji}\Delta s_j d\left(s_{js}, s_{jt}, \zeta_j\right) \ ,\end{aligned} \tag{A.12}$$

where $s \neq t$, and we know s given t (since there are only two periods). Therefore, we have

$$\begin{aligned}\mathcal{D}_H^{-1}C_N &= \sum_{i=1}^{N}\phi_i\widehat{\psi}_i^2\sum_{t=1}^{2}\lambda_{it}\pi_{it}\left(G_{it}\otimes x_{it}\right)\frac{c_H\left(z_i,z\right)}{\sum_{i=1}^{N}c_H\left(z_i,z\right)} \\ &\quad + \sum_{i=1}^{N}\phi_i\widehat{\psi}_i\sum_{t=1}^{2}\pi_{it}\left(G_{it}\otimes x_{it}\right)\sum_{j\neq i}^{N}\phi_j\widehat{\psi}_j\lambda_{jt}\frac{c_H\left(z_i,z\right)}{\sum_{i=1}^{N}c_H\left(z_i,z\right)} \\ &\quad - \sum_{i=1}^{N}\phi_i\widehat{\psi}_i\sum_{t=1}^{2}\pi_{it}\left(G_{it}\otimes x_{it}\right)\sum_{j=1}^{N}\phi_j\widehat{\psi}_j q_{ji}\Delta s_j d\left(s_{js}, s_{jt}, \zeta_j\right) \\ &\equiv D_{N1} + D_{N2} - D_{N3} \ ,\end{aligned}$$

where the definition of D_{Nj} $(j = 1, 2, 3)$ should be clear from the following context. Again, note that $s \neq t$ is known given t.

First, we have $\left\|(|H|\, h_0)^{-1} D_{N1}\right\| \sim (N\,|H|\, h_0)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{2} \phi_i \widehat{\psi}_i^2 \pi_{it} \,\|(G_{it} \otimes x_{it})\, \lambda_{it}\|$. Letting $v = \Delta s_i / h_0$ and $\omega = H^{-1}(z_{it} - z)$ and applying the change of variables yield

$$
\begin{aligned}
& (|H|\, h_0)^{-1} \sum_{t=1}^{2} \mathbb{E}\left(\phi_i \widehat{\psi}_i^2 \pi_{it} \,\|(G_{it} \otimes x_{it})\, \lambda_{it}\|\right) \\
= \;& (|H|\, h_0)^{-1}\, \mathbb{E}\{\mathbb{E}\left[\phi_i \,\|\Lambda\left(s_{it}, s_{is}, \zeta_i\right) G_{it} \otimes x_{it}\| \,|\Delta s_i, z_{it}\right] \pi_{it} \widehat{\psi}_i^2\} \\
= \;& \int\int k^2(v) \mathcal{K}(\omega) \mathbb{E}\left[\phi_i \left|\Lambda\left(s_{is} + v h_0, s_{is}, \zeta_i\right)\right| \left\| \begin{bmatrix} x_{it} \\ \omega \otimes x_{it} \end{bmatrix} \right\| \middle| v h_0, H\omega + z\right] f_t\left(v h_0, H\omega + z\right) dv d\omega \\
\leq \;& M \mathbb{E}\left[\phi_i \left|\Lambda\left(s_{is}, s_{is}, \zeta_i\right)\right| \|x_{it}\| \,|0, z\right] f_t(0, z) + o(1)\ .
\end{aligned}
$$

It then follows $\left\|(|H|\, h_0)^{-1} D_{N1}\right\| = O_p(1)$ by Assumption **S**. Hence, $(N\,|H|\, h_0)^{-1} D_{N1} = O_p(N^{-1})$.

Next, applying the change of variables approach we obtain

$$
\begin{aligned}
\tfrac{1}{N|H|h_0} \|D_{N2}\| \;\sim\; & \frac{1}{N\,|H|\, h_0} \sum_{i=1}^{N} \phi_i \widehat{\psi}_i \sum_{t=1}^{2} \pi_{it} \,\|G_{it} \otimes x_{it}\| \,\frac{1}{N} \sum_{j \neq i}^{N} \phi_j \widehat{\psi}_j \,|\lambda_{jt}| \\
& \leq M h_0 \sum_{t=1}^{2} f_t(0, z) \mathbb{E}\left(\phi_i \,\|x_{it}\| \,|0, z\right) \mathbb{E}\left(\,\phi_i \left|\Lambda\left(s_{it}, s_{it}, \zeta_i\right)\right| \,|\, 0, z\right) + o_p(h_0)\ ,
\end{aligned}
$$

where we give an inequality result in the second line to simplify the mathematical expression, although it is evident that $(N\,|H|\, h_0)^{-1} D_{N2} = O_e(h_0)$.

Similarly, we can show that $(N\,|H|\, h_0)^{-1} D_{N3} = O_p(h_0^2)$. This completes the proof of this lemma.

Lemma 3 *Under Assumptions **D**, **S**, **K** and **B**, we have*

$$
\begin{aligned}
& (N\,|H|\, h_0)^{-1} \mathcal{D}_H^{-1} \mathcal{X}(z)' \widehat{W}_h(z) \Pi(z) \\
\approx \;& (|H|\, h_0)^{-1} \sum_{t=1}^{2} \mathbb{E}\left[\phi_i \widehat{\psi}_i^2 \left(1 - \varpi_{it}\right) \pi_{it} \left(G_{it} \otimes x_{it}\right) x_{it}' r(\widetilde{z}_{it}, z)\right] = O_e(\|H\|^2)\ . \qquad \text{(A.13)}
\end{aligned}
$$

Proof: By (A.1), we have

$$
\begin{aligned}
\mathcal{D}_H^{-1} A_N \;\equiv\; & \mathcal{D}_H^{-1} \mathcal{X}(z)' \widehat{W}_h(z) \Pi(z) \\
= \;& \sum_{i=1}^{N} \phi_i \widehat{\psi}_i \sum_{t=1}^{2} \pi_{it} \left(G_{it} \otimes x_{it}\right) \left[\phi_i \widehat{\psi}_i x_{it}' r(\widetilde{z}_{it}, z) - \sum_{j=1}^{N} \phi_j \widehat{\psi}_j q_{ji} \sum_{s=1}^{2} \pi_{js} x_{js}' r(\widetilde{z}_{js}, z)\right] \\
\approx \;& \sum_{i=1}^{N} \phi_i \left(1 - \omega_{it}\right) \widehat{\psi}_i^2 \sum_{t=1}^{2} \pi_{it} \left(G_{it} \otimes x_{it}\right) x_{it}' r(\widetilde{z}_{it}, z)\ ,
\end{aligned}
$$

where we obtain the third line by following the proof of Lemma A.3 in Sun et al. (2009).

First, we consider $\Delta_N \equiv \sum_{i=1}^{N} \phi_i \widehat{\psi}_i^2 \sum_{t=1}^{2} \pi_{it} \left(G_{it} \otimes x_{it}\right) x_{it}' r(\widetilde{z}_{it}, z)$. Letting $v = \Delta s_i / h_0$ and

$\omega = H^{-1}(z_{it} - z)$ and applying the change of variables yield

$$\begin{aligned}
& (|H|\, h_0)^{-1}\, \mathbb{E}\left[\phi_i \widehat{\psi}_i^2 \pi_{it}\, (G_{it} \otimes x_{it})\, x_{it}' r(\widetilde{z}_{it}, z)\right] \\
= \;& (|H|\, h_0)^{-1}\, \mathbb{E}\left\{\mathbb{E}\left[\phi_i x_{it}' r(\widetilde{z}_{it}, z)\, (G_{it} \otimes x_{it}) \,|\Delta s_i, z_{it}\right] \pi_{it} \widehat{\psi}_i^2\right\} \\
= \;& \int\int k^2(v) \mathcal{K}(\omega) \mathbb{E}\left[\phi_i x_{it}' r(\delta_{it} H\omega + z, z) \left[\begin{array}{c} x_{it} \\ \omega \otimes x_{it} \end{array}\right] \middle| v h_0, H\omega + z\right] f_t\,(v h_0, H\omega + z)\, dv d\omega \\
= \;& \left[\int k^2(v) dv \int k(v) v^2 dv\right] f_t(0,z) \mathbb{E}\left(\left[\begin{array}{c} \phi_i x_{it} x_{it}' \Theta_H(z) \\ 0_{q\times p} \end{array}\right] \middle| 0, z\right) + O\left(h_0^2 \|H\|^2 + \|H\|^4\right), \quad \text{(A.14)}
\end{aligned}$$

where $\widetilde{z}_{it} = \delta_{it} z_{it} + (1 - \delta_{it})\, z = \delta_{it}\,(z_{it} - z) + z$ for some $\delta_{it} \in (0,1)$, $0_{q\times p}$ is a $q \times p$ matrix of zeros, and

$$\Theta_H(z) = \left[\mathrm{tr}\left(H \frac{\partial^2 \beta_1(z)}{\partial z \partial z'} H\right), \cdots, \mathrm{tr}\left(H \frac{\partial^2 \beta_p(z)}{\partial z \partial z'} H\right)\right]' .$$

Hence, $\mathbb{E}\left[(N\,|H|\, h_0)^{-1} \Delta_N\right] = \left[\int k^2(v) dv \int k(v) v^2 dv\right] \sum_{t=1}^2 f_t(0,z) \mathbb{E}\left(\left[\begin{array}{c} \phi_i x_{it} x_{it}' \Theta_H(z) \\ 0_{q\times p} \end{array}\right] \middle| 0, z\right) + O\left(h_0^2 \|H\|^2 + \|H\|^4\right)$. Similarly, we can show that $Var\left((N\,|H|\, h_0)^{-1} \Delta_N\right) = O\left((N\,|H|\, h_0)^{-1} \|H\|^4\right)$ if $\sum_{t=1}^2 \mathbb{E}\left(\phi_i \|x_{it}\|^4 \,|\Delta s_i = 0, z_{it} = z\right) < M < \infty$.

Next, since ϖ_{it} contains a random denominator, we cannot obtain a closed-form limit result for $(N\,|H|\, h_0)^{-1} \sum_{i=1}^N \sum_{t=1}^2 \phi_i \widehat{\psi}_i^2 \varpi_{it} \pi_{it}\, (G_{it} \otimes x_{it})\, x_{it}' r\,(\widetilde{z}_{it}, z)$. However, as $\varpi_{it} \in (0,1)$ for all i and t, this term is of the same stochastic order as $(N\,|H|\, h_0)^{-1} \Delta_N = O_e\left(\|H\|^2\right)$. This completes the proof of this lemma.

Lemma 4 *Under Assumptions **D**, **S**, **K** and **B**, we have*

$$\begin{aligned}
& (N\,|H|\, h_0)^{-1}\, \mathcal{D}_H^{-1} \mathcal{X}(z)' \widehat{W}_h^2(z) \mathcal{X}(z) \mathcal{D}_H^{-1} \\
\approx \;& (|H|\, h_0)^{-1} \sum_{t=1}^2 \mathbb{E}\left[\phi_i \widehat{\psi}_i^2\, (1 - \varpi_{it})^2\, \pi_{it}^2 [\cdot]_{it} \otimes \left(x_{it} x_{it}'\right)\right] = O_e(1)\ . \quad \text{(A.15)}
\end{aligned}$$

Proof: By (A.1) and following the proof of Lemma A.5 in Sun et al. (2009, p.125), we obtain

$$\mathcal{D}_H^{-1} \mathcal{X}(z)' \widehat{W}_h^2(z) \mathcal{X}(z) \mathcal{D}_H^{-1} \approx \sum_{i=1}^N \phi_i \widehat{\psi}_i^2\, (1 - \varpi_{it})^2\, \pi_{it}^2 [\cdot]_{it} \otimes \left(x_{it} x_{it}'\right)\ . \quad \text{(A.16)}$$

Then, applying the change of variables approach as in the proof of Lemma 1, we can show that

$$(N\,|H|\, h_0)^{-1} \sum_{i=1}^N \phi_i \widehat{\psi}_i^2 \pi_{it}^2 [\cdot]_{it} \otimes \left(x_{it} x_{it}'\right) \approx \kappa_2 \sum_{t=1}^2 \mathbb{E}\left[\phi_i R_{K,4} \otimes \left(x_{it} x_{it}'\right) \middle| 0, z\right] f_t(0,z) + o_p(1)\ ,$$

where $R_{K,4} = \mathrm{diag}\,\{1, \varsigma_2, \ldots, \varsigma_2\}$ is a $(q+1) \times (q+1)$ diagonal matrix and $\varsigma_2 = \int k^2(v) v^2 dv$. This completes the proof of this lemma.

Proof of Theorem 1: First, by (A.2), (A.4), (A.10) and (A.13), we have $Bias\left[\mathrm{vec}\left\{\widehat{\Theta}(z)\right\} |\phi_i, \zeta_i\right] = O_e\left(\|H\|^2\right) + O_e(h_0)$, where the selection bias is of order $O_e(h_0)$ and cannot be improved in general.

Second, the asymptotic normality result is obtained from $E_N^{-1}D_N$, where $D_N \equiv \mathcal{X}(z)'\widehat{W}_h(z)V$ and $E_N \equiv \mathcal{X}(z)'\widehat{W}_h(z)\mathcal{X}(z)$ as defined in (3.14). As in Lemma 3, we have $(N|H|h_0)^{-1/2}\mathcal{D}_H^{-1}D_N \equiv (N|H|h_0)^{-1/2}\mathcal{D}_H^{-1}\mathcal{X}(z)'\widehat{W}_h(z)V \approx \sum_{i=1}^N Z_{Ni}$, where $Z_{Ni} \equiv (N|H|h_0)^{-1/2}\phi_i\widehat{\psi}_i\sum_{t=1}^2(1-\omega_{it})\pi_{it}(G_{it}\otimes x_{it})v_{it}$. Assumption **D** indicates that $\{Z_{Ni}\}$ is an *i.i.d.* array as h_0 and H depend on the sample size N. We will apply Cramér-Wold device and Liaponuov's CLT to derive the asymptotic normality result. In doing so, we only need to check that for some $\delta > 0$

$$\mathbb{E}\|Z_{Ni}\|^{2+\delta} < M < \infty \qquad \text{and} \qquad \lim_{N\to\infty}\sum_{i=1}^N \mathbb{E}\|Z_{Ni}\|^{2+\delta} = 0 \tag{A.17}$$

as

$$\begin{aligned} Var\left(\sum_{i=1}^N Z_{Ni}\right) &= \sigma_v^2(N|H|h_0)^{-1}\mathbb{E}\left[\mathcal{D}_H^{-1}\mathcal{X}(z)'\widehat{W}_h^2(z)\mathcal{X}(z)\mathcal{D}_H^{-1}\right] \\ &\approx \sigma_v^2(|H|h_0)^{-1}\sum_{t=1}^2\mathbb{E}\left[\phi_i\widehat{\psi}_i^2(1-\varpi_{it})^2\pi_{it}^2[\cdot]_{it}\otimes\left(x_{it}x_{it}'\right)\right]+o(1) \end{aligned}$$

by Lemma 4.

By Proposition 3.8 in White (2001, p.35) we have[32]

$$\begin{aligned} \mathbb{E}\|Z_{Ni}\|^{2+\delta} &= \frac{1}{(N|H|h_0)^{1+\delta/2}}\mathbb{E}\left[\left\|\phi_i\widehat{\psi}_i\sum_{t=1}^2(1-\omega_{it})\pi_{it}(G_{it}\otimes x_{it})v_{it}\right\|^{2+\delta}\right] \\ &\le \frac{2^{1+\delta}}{(N|H|h_0)^{1+\delta/2}}\sum_{t=1}^2\mathbb{E}\left[\left\|\phi_i\widehat{\psi}_i\pi_{it}(G_{it}\otimes x_{it})v_{it}\right\|^{2+\delta}\right], \end{aligned}$$

where we obtain

$$\|(G_{it}\otimes x_{it})\| = \sqrt{x_{it}'x_{it}\left(1+\sum_{l=1}^q G_{it,l}^2\right)}. \tag{A.18}$$

Letting $v = \Delta s_i/h_0$ and $\omega = H^{-1}(z_{it}-z)$ and applying the change of variables yield

$$\begin{aligned} &(|H|h_0)^{-1}\mathbb{E}\left[\left\|\phi_i\widehat{\psi}_i\pi_{it}(G_{it}\otimes x_{it})v_{it}\right\|^{2+\delta}\right] \\ &= \int\int k^{2+\delta}(v)\mathcal{K}^{2+\delta}(\omega)\left(1+\sum_{l=1}^q\omega_l^2\right)^{1+\delta/2}\mathbb{E}\left[\phi_i|v_{it}|^{2+\delta}\left(x_{it}'x_{it}\right)^{1+\delta/2}\Big|\, vh_0, H\omega+z\right]f_t(vh_0, H\omega+z)\,dvd\omega \\ &= f_t(0,z)\int\int k^{2+\delta}(v)\mathcal{K}^{2+\delta}(\omega)\left(1+\sum_{l=1}^q\omega_l^2\right)^{1+\delta/2}dvd\omega\,\mathbb{E}\left[\phi_i|v_{it}|^{2+\delta}\left(x_{it}'x_{it}\right)^{1+\delta/2}\Big|\, 0, z\right]+o(1) \\ &\le M<\infty. \end{aligned}$$

Hence, we obtain $\mathbb{E}\|Z_{Ni}\|^{2+\delta} = O\left(N^{-1}(N|H|h_0)^{-\delta/2}\right)$. (A.17) holds accordingly as $N|H|h_0 \to \infty$ when $N\to\infty$. Combining the above with Lemmas 1 and 4 completes the proof of this theorem.

[32] $\mathbb{E}|X+Y|^r \le 2^{r-1}(\mathbb{E}|X|^r + \mathbb{E}|Y|^r)$ for $r>1$.

B Proof of Theorem 2

From (3.14), it is clear that γ affects the performance of $\widehat{\beta}(z)$ via the term $\widehat{\psi}_i = k\left(\Delta w_i'\gamma/h_0\right)$. We now replace $\widehat{\psi}_i$ by $\widetilde{\psi}_{N,i} = k\left(\Delta w_i'\widehat{\gamma}_N/h_0\right)$ and denote $\widetilde{K}_{N,h}(z) = \text{diag}\left\{\widetilde{\psi}_{N,1}K_h(z_1,z),\ldots,\widetilde{\psi}_{N,N}K_h(z_N,z)\right\}$, $\widehat{\Gamma}_{N,h}(z) = I_{2N} - D\left(D'\widetilde{K}_{N,h}(z)D\right)^{-1}D'\widetilde{K}_{N,h}(z)$ and $\widetilde{W}_{N,h}(z) = \widetilde{\Gamma}_{N,h}(z)'\widetilde{K}_{N,h}(z)\widetilde{\Gamma}_{N,h}(z)$. Accordingly, we denote $\widetilde{A}_N \equiv \mathcal{X}(z)'\widetilde{W}_{N,h}(z)\Pi(z)$, $\widetilde{B}_N \equiv \mathcal{X}(z)'\widetilde{W}_{N,h}(z)D\mu$, $\widetilde{C}_N \equiv \mathcal{X}(z)'\widetilde{W}_{N,h}(z)\lambda$, $\widetilde{D}_N \equiv \mathcal{X}(z)'\widetilde{W}_{N,h}(z)V$ and $\widetilde{E}_N \equiv \mathcal{X}(z)'\widetilde{W}_{N,h}(z)\mathcal{X}(z)$, where $\widetilde{B}_N = 0$ as $\widetilde{\Gamma}_{N,h}(z)D\mu = 0$.

We consider these terms one by one below.

Lemma 5 *Under Assumptions **D**, **S**, **S2**, **K**, **K2** and **B** and if $\|\widehat{\gamma}_N - \gamma\|/h_0^2 \to 0$ as $N \to \infty$, we have*

$$\left(N|H|h_0\right)^{-1}\mathcal{D}_H^{-1}\left(\widetilde{E}_N - E_N\right) = O_e\left(\|\widehat{\gamma}_N - \gamma\|/h_0^2\right), \tag{B.1}$$

$$\left(N|H|h_0\right)^{-1}\mathcal{D}_H^{-1}\left(\widetilde{A}_N - A_N\right) = O_e\left(\|\widehat{\gamma}_N - \gamma\|\,\|H\|^2/h_0\right), \tag{B.2}$$

$$\left(N|H|h_0\right)^{-1}\mathcal{D}_H^{-1}\left(\widetilde{C}_N - C_N\right) = O_e\left(\|\widehat{\gamma}_N - \gamma\|/h_0\right), \tag{B.3}$$

$$\left(N|H|h_0\right)^{-1/2}\mathcal{D}_H^{-1}\left(\widetilde{D}_N - D_N\right) = O_e\left(\|\widehat{\gamma}_N - \gamma\|/h_0^2\right). \tag{B.4}$$

Proof: By Assumptions **K** and **K2**, we have

$$\left|\widetilde{\psi}_{N,i} - \widehat{\psi}_i\right| \equiv \left|k\left(\frac{\Delta w_i'\widehat{\gamma}_N}{h_0}\right) - k\left(\frac{\Delta w_i'\gamma}{h_0}\right)\right| \leq M\frac{\|\Delta w_i\|\,\|\widehat{\gamma}_N - \gamma\|}{h_0}, \tag{B.5}$$

$$\left|\widetilde{\psi}_{N,i}^2 - \widehat{\psi}_i^2\right| \leq M\frac{\|\Delta w_i\|\,\|\widehat{\gamma}_N - \gamma\|}{h_0}k\left(\frac{\Delta w_i'\widetilde{\gamma}_i}{h_0}\right), \tag{B.6}$$

where $\Delta w_i'\widetilde{\gamma}_i$ lies between $\Delta w_i'\widehat{\gamma}_N$ and $\Delta w_i'\gamma$.

First, following the proof of Lemma 1, we have

$$\left(N|H|h_0\right)^{-1}\mathcal{D}_H^{-1}\left(\widetilde{E}_N - E_N\right) \approx \frac{1}{N|H|h_0}\sum_{i=1}^N\phi_i\left(\widetilde{\psi}_{N,i} - \widehat{\psi}_i\right)\sum_{t=1}^2\left(1-\omega_{it}\right)\pi_{it}[\cdot]_{it,it}\otimes\left(x_{it}x_{it}'\right) \equiv \Delta_N .$$

Then by (B.5) we have $\|\Delta_N\| \leq M\left(N|H|h_0^2\right)^{-1}\|\widehat{\gamma}_N - \gamma\|\sum_{i=1}^N\phi_i\|\Delta w_i\|\sum_{t=1}^2\pi_{it}\|[\cdot]_{it,it}\otimes(x_{it}x_{it}')\|$. By (A.8) and applying the change of variables approach yields

$$|H|^{-1}\mathbb{E}\left[\phi_i\|\Delta w_i\|\pi_{it}\left(1+\sum_{l=1}^q G_{it,l}^2\right)x_{it}'x_{it}\right] = \left[1+(q+1)\kappa_2\right]\mathbb{E}\left[\phi_i\|\Delta w_i\|x_{it}'x_{it}\middle|z_{it}=z\right]f_t(z)+o(1),$$

which implies $\|\Delta_N\| = O_e\left(\|\widehat{\gamma}_N - \gamma\|/h_0^2\right)$.

Second, following the proof of Lemma 3, we have

$$\left(N|H|h_0\right)^{-1}\mathcal{D}_H^{-1}\left(\widetilde{A}_N - A_N\right) \approx \sum_{i=1}^N\phi_i\left(\widetilde{\psi}_{N,i}^2 - \widehat{\psi}_i^2\right)\sum_{t=1}^2\omega_{it}\pi_{it}\left(G_{it}\otimes x_{it}\right)x_{it}'r\left(\widetilde{z}_{it},z\right) \equiv \Delta_N .$$

Then, by (A.18), (B.5) and (B.6) we have

$$\begin{aligned}\|\Delta_N\| &\leq \frac{M\|\widehat{\gamma}_N - \gamma\|}{N|H|h_0^2}\sum_{i=1}^N\phi_i\|\Delta w_i\|k\left(\frac{\Delta w_i'\widetilde{\gamma}_i}{h_0}\right)\sum_{t=1}^2\pi_{it}\|G_{it}\otimes x_{it}\|\left|x_{it}'r\left(\widetilde{z}_{it},z\right)\right| \\ &= O_e\left(\|\widehat{\gamma}_N - \gamma\|\,\|H\|^2h_0^{-1}\right),\end{aligned}$$

where we use

$$
\begin{aligned}
& |H|^{-1}\,\mathbb{E}\left[\phi_i\,\|\Delta w_i\|\,k\left(\frac{\Delta w_i'\widetilde{\gamma}_i}{h_0}\right)\pi_{it}\sqrt{\left(1+\sum_{l=1}^{q}G_{it,l}^2\right)x_{it}'x_{it}}\,\left|x_{it}'r\left(\widetilde{z}_{it},z\right)\right|\right] \\
\leq\ & M\,\|H\|^2\,\mathbb{E}\left[\phi_i\,\|\Delta w_i\|\,\|x_{it}\|^2\middle|\,\Delta s_i=0,z_{it}=z\right]f_t(0,z) \\
& \times\sum_{l_1=1}^{q}\sum_{l_2=1}^{q}\left\|\frac{\partial^2\beta(z)}{\partial z_{l_1}\partial z_{l_2}}\right\|\int\mathcal{K}(\omega)\sqrt{\left(1+\sum_{l=1}^{q}\omega_l^2\right)}\,|\omega_{l_1}\omega_{l_2}|\,d\omega\,[1+o(1)]\ .
\end{aligned}
$$

Third, following the proof of Lemma 2, we have

$$
\mathcal{D}_H^{-1}\left(\widetilde{C}_N-C_N\right)\approx\sum_{i=1}^{N}\phi_i\sum_{t=1}^{2}\pi_{it}\left(G_{it}\otimes x_{it}\right)\sum_{j\neq i}^{N}\phi_j\left(\widetilde{\psi}_{N,i}\widetilde{\psi}_{N,j}-\widehat{\psi}_i\widehat{\psi}_j\right)\lambda_{jt}\frac{c_H\left(z_i,z\right)}{\sum_{i=1}^{N}c_H\left(z_i,z\right)}\equiv\Delta_N\ .
$$

A simple calculation gives

$$
\widetilde{\psi}_{N,i}\widetilde{\psi}_{N,j}-\widehat{\psi}_i\widehat{\psi}_j=\left(\widetilde{\psi}_{N,i}-\widehat{\psi}_i\right)\widehat{\psi}_j+\left(\widetilde{\psi}_{N,j}-\widehat{\psi}_j\right)\widetilde{\psi}_{N,i}\ .
$$

By (B.5), we then obtain

$$
\begin{aligned}
\frac{\|\Delta_N\|}{N\,|H|\,h_0} &\leq M\frac{\|\widehat{\gamma}_N-\gamma\|}{N\,|H|\,h_0^2}\sum_{i=1}^{N}\phi_i\sum_{t=1}^{2}\pi_{it}\,\|G_{it}\otimes x_{it}\|\sum_{j\neq i}^{N}\phi_j\,|\lambda_{jt}|\,\frac{c_H\left(z_i,z\right)}{\sum_{i=1}^{N}c_H\left(z_i,z\right)} \\
&\quad\times\left[\widehat{\psi}_j\,\|w_i\|+\widetilde{\psi}_{N,i}\,\|\Delta w_j\|\right] \\
&= O_e\left(\|\widehat{\gamma}_N-\gamma\|/h_0\right)\ .
\end{aligned}
$$

Next, following the proof of Lemma 3, we can show that

$$
\mathcal{D}_H^{-1}\left(\widetilde{D}_N-D_N\right)\approx\sum_{i=1}^{N}\phi_i\left(\widetilde{\psi}_{N,i}^2-\widehat{\psi}_i^2\right)\sum_{t=1}^{2}\omega_{it}\pi_{it}\left(G_{it}\otimes x_{it}\right)v_{it}\equiv\Delta_N\ ,
$$

where $\mathbb{E}\left(\Delta_N\right)=0$ and $Var\left(\left(N\,|H|\,h_0\right)^{-1/2}\Delta_N\right)=O\left(\|\widehat{\gamma}_N-\gamma\|^2/h_0^2\right)$. Hence, (B.4) holds. This completes the proof of this lemma.

Proof of Theorem 2: Let $\widetilde{\beta}(z)$ be the estimator of $\beta(z)$ when γ is replaced by $\widehat{\gamma}_N$. Then, taking Lemmas 1 to 5 together, we obtain

$$
\begin{aligned}
\widetilde{\beta}(z)-\widehat{\beta}(z) &= S_p\widetilde{E}_N^{-1}\left(\widetilde{A}_N+\widetilde{C}_N+\widetilde{D}_N\right)-S_pE_N^{-1}\left(A_N+C_N+D_N\right) \\
&= S_p\widetilde{E}_N^{-1}\left[\left(\widetilde{A}_N-A_N\right)+\left(\widetilde{C}_N-C_N\right)+\left(\widetilde{D}_N-D_N\right)\right] \\
&\quad+S_p\left(\widetilde{E}_N^{-1}-E_N^{-1}\right)\left(A_N+C_N+D_N\right) \\
&= O_e\left(\|\widehat{\gamma}_N-\gamma\|/h_0^2\right)O_e\left(\|H\|^2+h_0+\left(N\,|H|\,h_0\right)^{-1/2}\right)\ ,
\end{aligned}
$$

if $\|\widehat{\gamma}_N-\gamma\|/h_0^2=o_p(1)$. It then follows that $\sqrt{N\,|H|\,h_0}\left[\widetilde{\beta}(z)-\widehat{\beta}(z)\right]=O_e\left(\|\widehat{\gamma}_N-\gamma\|/h_0^2\right)$ as $\sqrt{N|H|h_0}\,\|H\|^2=o_p(1)$ and $\sqrt{N|H|h_0^3}=o_p(1)$. This completes the proof of this theorem.

References

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1):3–29.

Baltagi, B. (2013). *Econometric Analysis of Panel Data*. 5th edition.

Bauer, K. (2008). Detecting abnormal credit union performance. *Journal of Banking & Finance*, 32(4):573–586.

Bauer, K. J., Miles, L. L., and Nishikawa, T. (2009). The effect of mergers on credit union performance. *Journal of Banking & Finance*, 33(12):2267–2274.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136(1):163–188.

Cai, Z., Das, M., Xiong, H., and Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics*, 133(1):207–241.

Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902.

Cai, Z. and Li, Q. (2008). Nonparametric estimation of varying coefficient dynamic panel data models. *Econometric Theory*, 24(5):1321.

Cai, Z. and Xiong, H. (2012). Partially varying coefficient instrumental variables models. *Statistica Neerlandica*, 66(2):85–110.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47(1):225–238.

Charlier, E., Melenberg, B., and van Soest, A. (2001). An analysis of housing expenditure using semiparametric models and panel data. *Journal of Econometrics*, 101(1):71–107.

Das, M. (2003). Identification and sequential estimation of panel data models with insufficient exclusion restrictions. *Journal of Econometrics*, 114(2):297–328.

Das, M. (2004). Simple estimators for nonparametric panel data models with sample attrition. *Journal of Econometrics*, 120(1):159–180.

Das, M. (2005). Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal of Econometrics*, 124(2):335–361.

Das, M., Newey, W. K., and Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies*, 70(1):33–58.

Du, P., Parmeter, C. F., and Racine, J. S. (2013). Nonparametric kernel regression with multiple predictors and multiple shape constraints. *Statistica Sinica*, 23(3):1347–1371.

Dustmann, C. and Rochina-Barrachina, M. E. (2007). Selection correction in panel data models: An application to the estimation of females' wage equations. *Econometrics Journal*, 10(2):263–293.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications*. Chapman & Sons, New York.

Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *RAND Journal of Economics*, 38(4):1002–1019.

Frame, S. W., Karels, G. V., and McClatchey, C. A. (2003). Do credit unions use their tax advantage to benefit members? Evidence from a cost function. *Review of Financial Economics*, 12(1):35–47.

Fried, H. O., Lovell, C., and Yaisawarng, S. (1999). The impact of mergers on credit union service provision. *Journal of Banking & Finance*, 23(2):367–386.

Goddard, J. A., McKillop, D. G., and Wilson, J. O. (2002). The growth of US credit unions. *Journal of Banking & finance*, 26(12):2327–2356.

Goddard, J. A., McKillop, D. G., and Wilson, J. O. (2008). The diversification and financial performance of US credit unions. *Journal of Banking & Finance*, 32(9):1836–1849.

Hall, P. and Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Annals of Statistics*, 29(3):624–647.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796.

Heckman, J. (1974). Shadow prices, market wages and labor supply. *Econometrica*, 42(4):679–694.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Henderson, D. J., Carroll, R. J., and Li, Q. (2008). Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, 144(1):257–275.

Henderson, D. J. and Ullah, A. (2005). A nonparametric random effects estimator. *Economics Letters*, 88(3):403–407.

Honoré, B. E., Kyriazidou, E., and Powell, J. L. (2000). Estimation of tobit-type models with individual specific effects. *Econometric Reviews*, 19(3):341–366.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–531.

Hughes, J. P. and Mester, L. J. (2013). Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function. *Journal of Financial Intermediation*, 22(4):559–585.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65(6):1335–1364.

Kyriazidou, E. (2001). Estimation of dynamic panel data sample selection models. *Review of Economic Studies*, 68(3):543–572.

Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*, 15(3):261–274.

Li, Q., Huang, C. J., Li, D., and Fu, T.-T. (2002). Semiparametric smooth coefficient models. *Journal of Business & Economic Statistics*, 20(3):412–422.

Li, Q. and Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14(2):485–512.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, Princeton.

Li, Q. and Stengos, T. (1996). Semiparametric estimation of partially linear panel data models. *Journal of Econometrics*, 71(1):389–397.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 68:69–88.

Maddala, G. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

Malikov, E., Restrepo-Tobón, D. A., and Kumbhakar, S. C. (2013). Are all US credit unions alike? A generalized model of heterogeneous technologies with endogenous switching and fixed effects. Working Paper, Binghamton University.

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

McFadden, D. (1974). Analysis of qualitative choice behavior. In Zarembka, P., editor, *Econometrics*. Academic Press, New York.

National Credit Union Administration (2011). *2011 Annual Report*.

Rochina-Barrachina, M. E. (1999). A new estimator for panel data sample selection models. *Annales d'Economie et de Statistique*, 55-56:153–181.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Cambridge University Press, Cambridge.

Silverman, B. W. (1986). *Density Estimation*. Chapman and Hall, London.

Smith, D. J. (1984). A theoretic framework for the analysis of credit union decision making. *Journal of Finance*, 39(4):1155–1168.

Su, L. and Ullah, A. (2006). Profile likelihood estimation of partially linear panel data models with fixed effects. *Economics Letters*, 92(1):75–81.

Sun, Y., Carroll, R. J., and Li, D. (2009). Semiparametric estimation of fixed-effects panel data varying coefficient models. *Advances in Econometrics*, 25:101–129.

Wheelock, D. C. and Wilson, P. W. (2011). Are credit unions too small? *Review of Economics and Statistics*, 93(4):1343–1359.

Wheelock, D. C. and Wilson, P. W. (2013). The evolution of cost-productivity and efficiency among US credit unions. *Journal of Banking & Finance*, 37:75–88.

White, H. (2001). *Asymptotic Theory for Econometricians*. Academic Press.

Wilcox, J. A. (2005). Economies of scale and continuing consolidation of credit unions. *FRBSF Economic Letter*, (Nov 4).

Wilcox, J. A. (2006). Performance divergence of large and small credit unions. *FRBSF Economic Letter*, (Aug 4).

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68(1):115–132.

Yan, J. (2013). A smoothed maximum score estimator for multinomial discrete choice models. Working Paper, University of Wisconsin - Madison.

Tables and Figures

Table 1: Average RMSE across Monte Carlo Simulations for the Case of Binary Sample Selection

Estimator	$N = 100;\ T = 2$	$N = 200;\ T = 2$	$N = 400;\ T = 2$
A	1.3272	1.0350	0.8289
B	0.9702	0.8069	0.7085
C	0.9876	0.8039	0.7064

Note: The results are based on 500 simulations.

Table 2: Average RMSE across Monte Carlo Simulations for the Case of Polychotomous Switching

Estimator	$N = 150;\ T = 3$	$N = 300;\ T = 3$	$N = 600;\ T = 3$
		Regime #1	
A	1.2999	0.8709	0.6421
B	1.0417	0.7318	0.5707
C	1.0778	0.7599	0.6088
		Regime #2	
A	1.3010	0.8730	0.6192
B	0.9997	0.6332	0.4445
C	1.0259	0.6954	0.4994
		Regime #3	
A	1.3259	0.8738	0.6288
B	1.0454	0.7347	0.5730
C	1.0924	0.7677	0.5965

Note: The results are based on 500 simulations.

Table 3: Summary Statistics

Variable	Mean	Min	Q_{25}	Q_{50}	Q_{75}	Max
			Credit Unions of Type 1			
Cost	173.3	9.6	58.3	117.4	203.8	1,815.0
$y3$	2,537.0	45.5	890.0	1,825.0	3,177.0	16,880.0
$y4$	1,536.0	0.1	129.9	550.7	1,809.0	26,090.0
$\widetilde{y}5$	0.017	0.001	0.011	0.016	0.021	0.045
$1/\widetilde{y}6$	0.089	0.036	0.074	0.085	0.099	0.187
$w1$	0.021	0.004	0.014	0.019	0.025	0.089
$w2$	38.7	0.5	24.6	37.2	48.4	154.4
Total Assets	4,860.0	218.9	1,599.0	3,488.0	6,365.0	36,740.0
Equity	737.8	27.9	230.8	458.6	982.2	4,619.0
Leverage	0.007	0.000	0.001	0.003	0.007	0.128
Reserves	184.8	16.0	58.9	117.5	247.5	1,057.0
Current Members	1,164	53	494	777	1,449	19,520
Potential Members	6,991	100	750	1,500	3,200	241,800
			Credit Unions of Type 2			
Cost	2,545.0	29.5	425.7	917.1	2,454.0	76,720.0
$y1$	16,560.0	0.2	1,043.0	3,997.0	13,580.0	503,200.0
$y3$	27,280.0	85.1	4,714.0	9,026.0	23,990.0	1,012,000.0
$y4$	23,810.0	9.5	2,473.0	6,610.0	17,040.0	879,300.0
$\widetilde{y}5$	0.017	0.001	0.012	0.017	0.021	0.056
$1/\widetilde{y}6$	0.085	0.022	0.072	0.082	0.094	0.185
$w1$	0.022	0.003	0.016	0.022	0.027	0.074
$w2$	49.8	4.0	41.1	48.1	56.7	122.9
Total Assets	77,380.0	1,079.0	11,710.0	27,450.0	63,880.0	1,899,000.0
Equity	8,990.0	88.6	1,535.0	3,236.0	7,790.0	209,100.0
Leverage	0.010	0.000	0.002	0.005	0.009	0.282
Reserves	2,802.0	6.6	372.4	883.3	2,045.0	99,410.0
Current Members	9,796	249	2,155	4,248	10,570	185,200
Potential Members	126,400	300	5,000	10,000	50,000	8,383,000
			Credit Unions of Type 3			
Cost	14,790.0	187.2	1,939.0	6,799.0	16,640.0	121,700.0
$y1$	193,900.0	2.0	20,850.0	59,820.0	159,400.0	3,337,000.0
$y2$	11,630.0	1.3	517.9	2,152.0	11,570.0	194,300.0
$y3$	145,200.0	496.7	16,030.0	42,700.0	165,000.0	1,768,000.0
$y4$	108,800.0	103.1	7,298.0	24,920.0	82,850.0	2,418,000.0
$\widetilde{y}5$	0.020	0.005	0.015	0.019	0.024	0.035
$1/\widetilde{y}6$	0.076	0.051	0.066	0.075	0.085	0.109
$w1$	0.020	0.004	0.015	0.020	0.025	0.042
$w2$	55.4	25.6	45.7	53.2	64.2	89.3
Total Assets	510,400.0	4,398.0	60,600.0	165,500.0	527,800.0	7,456,000.0
Equity	54,170.0	519.5	7,118.0	19,910.0	57,260.0	844,500.0
Leverage	0.028	0.000	0.005	0.010	0.032	0.264
Reserves	18,960.0	170.8	1,617.0	5,094.0	12,450.0	522,600.0
Current Members	44,120	1,100	7,127	21,600	47,300	354,300
Potential Members	466,500	2,000	30,000	98,610	364,400	6,973,000

Notes: *Cost*, $y1$, $y2$, $y3$, $y4$, $w2$, *Total Assets, Equity, Reserves* are in thousands of real 2011 US dollars; $\widetilde{y}5$, $\widetilde{y}6$, $w1$, *Leverage* are unit-free interest rates. The numbers of *Current* and *Potential Members* are in terms of number of people. Despite that minima of several variables are reported to be zeros (due to rounding), they are not exactly equal to zeros.

Table 4.1: Summary of Elasticity Estimates for Credit Unions of Type 1

Model	Mean	St.Dev.	D_{10}	Q_{25}	Q_{50}	Q_{75}	D_{90}
				$y3$			
I	0.152	0.071	0.080	0.111	0.140	0.179	0.233
II	0.147	0.083	0.074	0.096	0.130	0.170	0.242
III	0.207	0.056	0.133	0.169	0.208	0.249	0.279
				$y4$			
I	0.044	0.036	0.010	0.020	0.036	0.053	0.086
II	0.041	0.038	0.011	0.017	0.028	0.048	0.096
III	0.021	0.007	0.011	0.016	0.021	0.026	0.030
				$\widetilde{y}5$			
I	0.095	0.060	0.037	0.057	0.077	0.115	0.185
II	0.085	0.054	0.032	0.049	0.069	0.110	0.159
III	0.057	0.020	0.032	0.041	0.054	0.072	0.086
				$\widetilde{y}6$			
I	0.167	0.082	0.077	0.106	0.154	0.209	0.273
II	0.163	0.095	0.062	0.093	0.143	0.205	0.305
III	0.086	0.025	0.053	0.067	0.087	0.105	0.119
				$w1$			
I	0.622	0.117	0.472	0.555	0.624	0.695	0.772
II	0.614	0.111	0.470	0.544	0.614	0.694	0.749
III	0.545	0.090	0.434	0.485	0.550	0.596	0.658
				t			
I	0.015	0.047	-0.042	-0.014	0.013	0.043	0.076
II	0.007	0.048	-0.062	-0.021	0.011	0.041	0.067
III	-0.007	0.013	-0.025	-0.017	-0.006	0.002	0.010

Table 4.2: Summary of Elasticity Estimates for Credit Unions of Type 2

Model	Mean	St.Dev.	D_{10}	Q_{25}	Q_{50}	Q_{75}	D_{90}
				$y1$			
I	0.049	0.028	0.024	0.033	0.043	0.058	0.074
II	0.040	0.029	0.017	0.023	0.032	0.048	0.071
III	0.088	0.018	0.062	0.077	0.089	0.101	0.110
				$y3$			
I	0.128	0.065	0.061	0.083	0.113	0.160	0.213
II	0.095	0.051	0.050	0.062	0.082	0.110	0.157
III	0.306	0.037	0.260	0.279	0.304	0.331	0.357
				$y4$			
I	0.043	0.033	0.016	0.024	0.035	0.052	0.078
II	0.036	0.022	0.016	0.022	0.031	0.043	0.061
III	0.079	0.017	0.059	0.071	0.081	0.090	0.098
				$\widetilde{y}5$			
I	0.086	0.058	0.036	0.048	0.067	0.109	0.155
II	0.072	0.052	0.029	0.040	0.055	0.087	0.132
III	0.061	0.018	0.039	0.047	0.058	0.073	0.087
				$\widetilde{y}6$			
I	0.116	0.077	0.051	0.068	0.094	0.135	0.221
II	0.108	0.065	0.040	0.063	0.095	0.134	0.186
III	0.016	0.012	0.000	0.007	0.016	0.025	0.033
				$w1$			
I	0.749	0.060	0.681	0.717	0.747	0.780	0.823
II	0.752	0.055	0.690	0.726	0.755	0.782	0.809
III	0.685	0.034	0.639	0.664	0.689	0.709	0.724
				t			
I	0.006	0.143	-0.174	-0.067	0.012	0.076	0.190
II	0.002	0.138	-0.172	-0.067	0.010	0.085	0.179
III	0.006	0.008	-0.004	0.001	0.007	0.012	0.017

Table 4.3: Summary of Elasticity Estimates for Credit Unions of Type 3

Model	Mean	St.Dev.	D_{10}	Q_{25}	Q_{50}	Q_{75}	D_{90}
				$y1$			
I	0.327	0.219	0.084	0.160	0.276	0.465	0.643
II	0.237	0.143	0.054	0.138	0.215	0.323	0.453
III	0.256	0.064	0.183	0.211	0.247	0.305	0.340
				$y2$			
I	0.078	0.057	0.024	0.041	0.063	0.102	0.151
II	0.074	0.050	0.022	0.038	0.065	0.095	0.140
III	0.028	0.012	0.012	0.020	0.028	0.035	0.043
				$y3$			
I	0.218	0.097	0.097	0.147	0.219	0.280	0.335
II	0.219	0.114	0.072	0.133	0.223	0.279	0.355
III	0.216	0.050	0.144	0.189	0.216	0.249	0.277
				$y4$			
I	0.160	0.077	0.069	0.108	0.153	0.204	0.258
II	0.140	0.075	0.058	0.090	0.123	0.187	0.230
III	0.146	0.043	0.085	0.123	0.150	0.180	0.197
				$\widetilde{y}5$			
I	0.170	0.108	0.055	0.085	0.141	0.237	0.311
II	0.210	0.135	0.069	0.110	0.180	0.281	0.400
III	0.133	0.038	0.085	0.106	0.132	0.162	0.178
				$\widetilde{y}6$			
I	0.200	0.132	0.061	0.100	0.162	0.279	0.377
II	0.217	0.139	0.072	0.121	0.182	0.282	0.453
III	0.208	0.069	0.117	0.164	0.209	0.256	0.298
				$w1$			
I	0.503	0.215	0.223	0.340	0.515	0.680	0.785
II	0.484	0.207	0.183	0.344	0.505	0.625	0.740
III	0.583	0.089	0.458	0.515	0.584	0.647	0.702
				t			
I	0.023	0.067	-0.050	-0.019	0.019	0.054	0.100
II	0.021	0.049	-0.033	-0.011	0.019	0.046	0.084
III	0.029	0.013	0.012	0.020	0.030	0.039	0.045

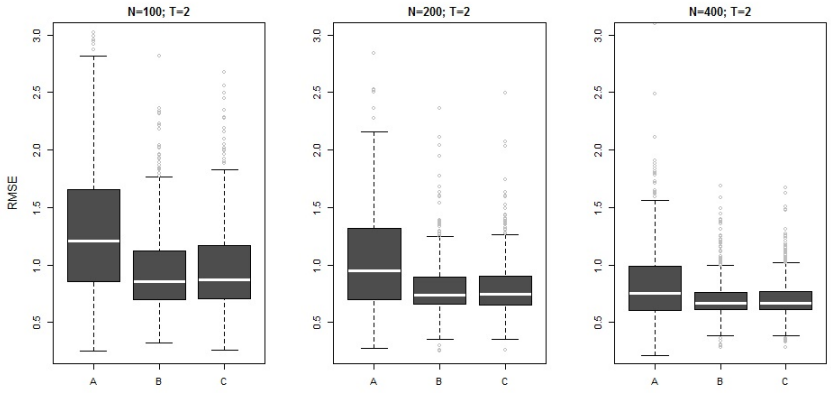

Figure 1: RMSE across Monte Carlo Simulations for Estimators A, B and C for the Case of Binary Sample Selection

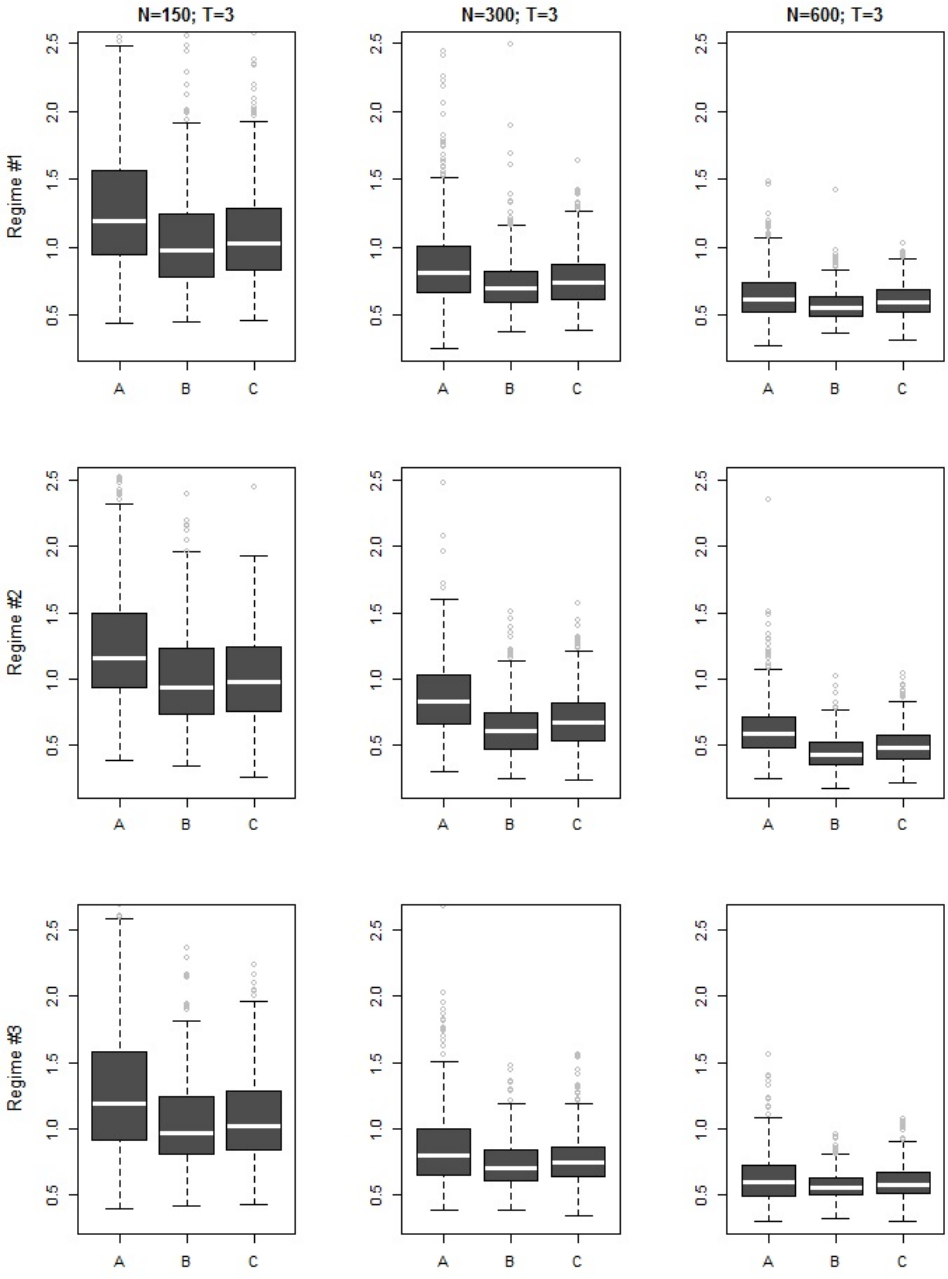

Figure 2: RMSE across Monte Carlo Simulations for Estimators A, B and C for the Case of Polychotomous Switching

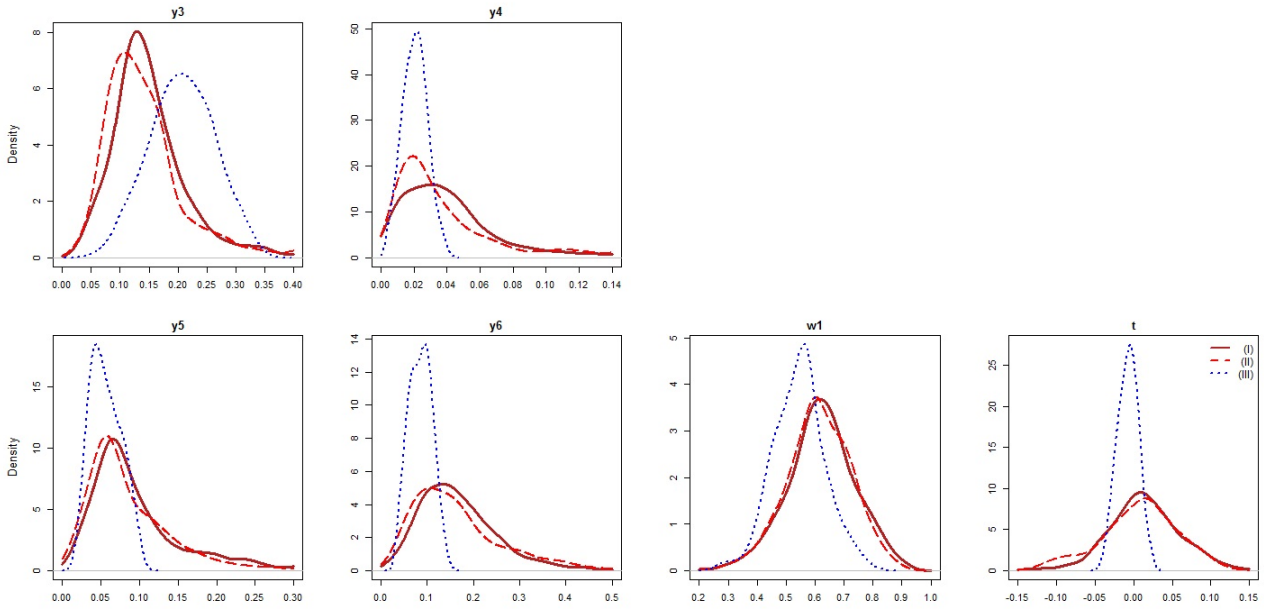

Figure 3.1: Kernel Densities of Elasticity Estimates for
Type 1 Credit Unions from Models I-III

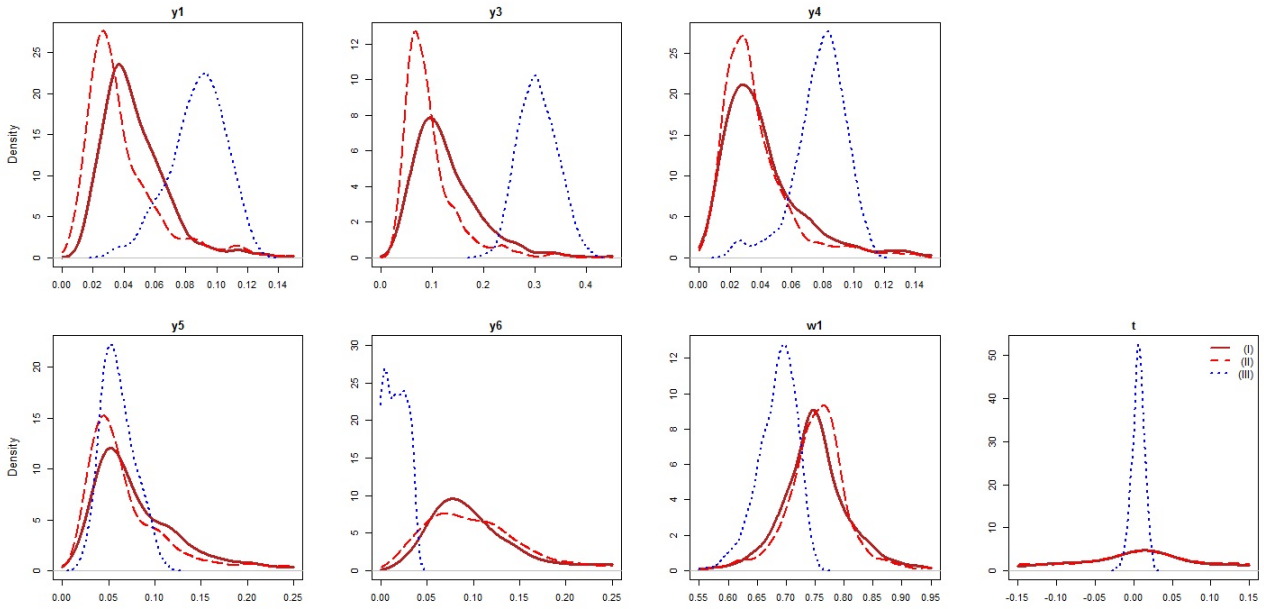

Figure 3.2: Kernel Densities of Elasticity Estimates for
Type 2 Credit Unions from Models I-III

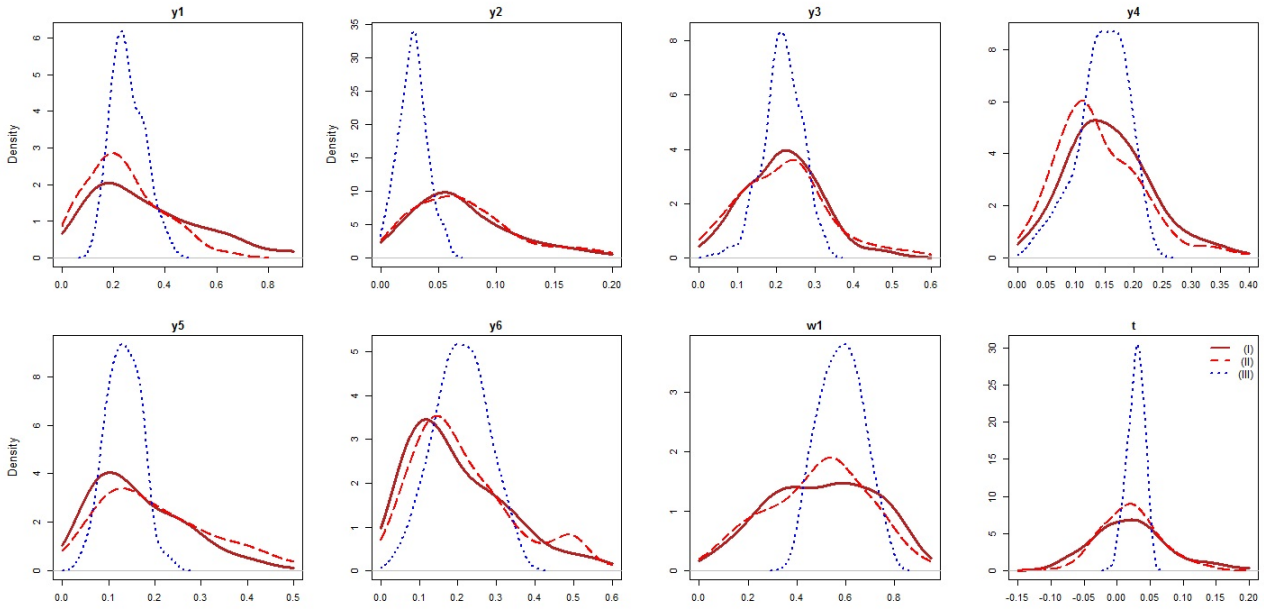

Figure 3.3: Kernel Densities of Elasticity Estimates for
Type 3 Credit Unions from Models I-III

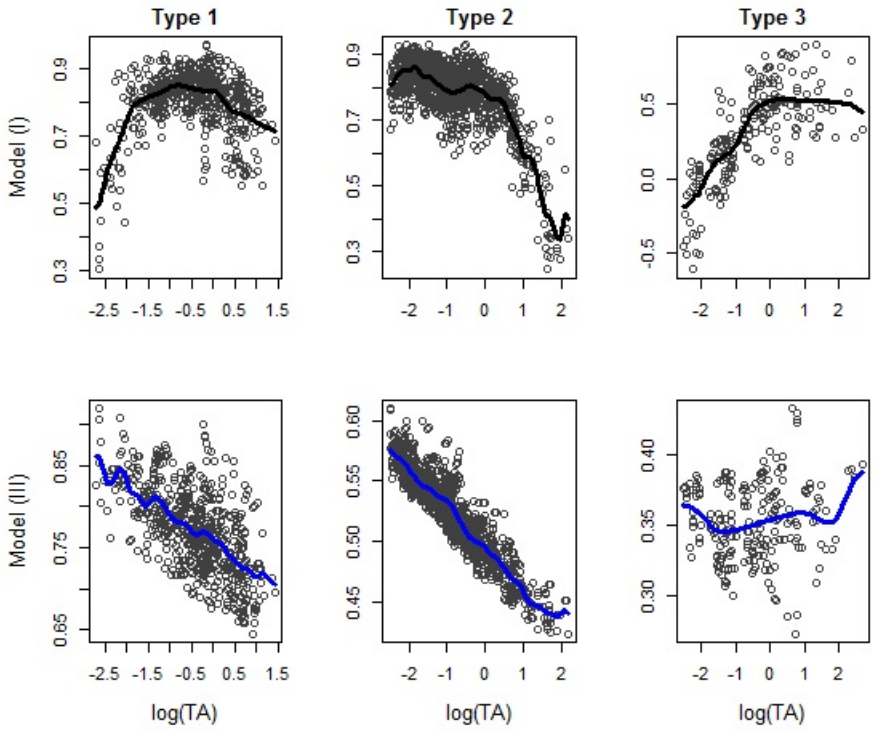

Figure 4: The Relationship between Scale Economies and the Asset Size of Credit Unions based on Models I and III