Estimating Heterogeneous Effects in Binary Response Panel Data Models

PRELIMINARY AND INCOMPLETE. COMMENTS ARE WELCOME.

Anastasia Semykina
Department of Economics
Florida State University
Tallahassee, FL 32306-2180, USA
E-mail: asemykina@fsu.edu
Phone: +1 850-644-4557
Fax: +1 850-644-4535

September 19, 2019

Abstract

The paper considers estimating heterogeneous effects on binary outcomes in different population subgroups. Applications include the evaluation of heterogeneous treatment effects, as well as estimation of causal effects of other policy-relevant variables. In the existing literature, it is common to divide the sample into group-specific subsamples and perform estimation separately for each group. In this paper, we argue that this estimation approach generally results in inconsistent estimators and present estimation methods that produce consistent estimators of causal effects in heterogeneous populations. The theoretical argument is illustrated with an empirical example.

JEL Classifications: C31, C33, C35
Keywords: binary response, heterogeneous effects, nonrandom sorting

1 Introduction

Researchers are frequently interested in estimating heterogeneous effects on binary outcomes in different population subgroups. Examples include studying dropout rates among high school students by race, gender, and socio-economic status, examining labor market participation decisions among married and single women, and investigating self-employment outcomes by age and education level. In the empirical literature, it is common to estimate such group-specific parameters by dividing the sample into corresponding subsamples and performing the estimation separately for each group. While this approach is intuitively appealing, it generally results in inconsistent estimators when sorting into groups is not random. Similar to linear models (Vella, 1988), consistent estimators of heterogeneous parameters can only be obtained if the full information set is utilized, i.e. when each group is considered as a part of the entire population. Estimation is further complicated when considering panel data models, which are characterized by unobserved hererogeneity at the unit level and cross-group transitions over time. The present paper discusses methods that address nonrandom sorting and produce consistent estimators of heterogeneous parameters and partial effects in binary response panel data models.

The related literature includes studies of linear switching regression models (Goldfeld and Quandt, 1973; Lee 1978; Maddala and Nelson, 1975; Maddala 1983). Such models specify two equations, where the applicability of either equation depends on the endogenous switching from one regime to the other. Another relevant strand of the literature includes studies of program evaluation and estimation of treatment effects. Analogous to switching regression models, program evaluation studies focus on addressing endogenous self-selection into treatment. One parameter of interest is the effect of treatment on the treated, which can be formulated within either a switching regression or self-selection framework (Bjorklund and Moffitt, 1987; Heckman et al., 2006). Furthermore, several studies have proposed methods for estimating heterogeneous treatment effects using the instrumental variables methodology (Heckman et al., 2006; Basu, 2014, and others).

The problem of nonrandom selection is discussed in studies of sample selection, including the seminal paper by Heckman (1979). In those models, parameters are assumed to be the same for all units in the population, and the selection problem arises because the dependent variable is not observed for some part of the population. The existing literature discusses methods for addressing sample selection in linear and binary response models. Moreover, both cross section and panel data models have been considered (Heckman, 1979; Kyriazidou, 1997; Newey, 2009; Semykina and Wooldridge, 2017; Wooldridge, 1995, among others).

Regarding heterogeneous effects in binary response models, several studies discuss switching probit for cross section and panel data (Carrasco, 2001; Manski et al., 1992). Similar to linear models, the endogenous switching is between two regimes, and parameters are regime-specific. However, to the best of our knowledge, estimating heterogeneous effects models with an arbitrary number of groups (regimes) has not been considered so far. The present paper proposes methods for estimating heterogeneous effects in binary response panel data models with two or more groups. The models account for the presence of unobserved heterogeneity that may be correlated with explanatory variables and accommodate multiple ordered and unordered groups.

The rest of the paper is structured as follows. Section 2 presents binary response models with heterogeneous effects. Estimation of population parameters and partial effects is discussed in Section 3. Simulation results are presented in Section 4. Section 5 contains an empirical application, and Section 6 concludes.

2 Heterogeneity in binary response panel data models

2.1 General Setup

Consider a population that consists of J groups (or subpopulations). Assume that the number of periods, T, is fixed, and $N \to \infty$, where N is the cross section sample size. Consider the following binary response model with heterogeneous effects:

$$\begin{aligned} y_{itj}^* &= \mathbf{x}_{it}\boldsymbol{\beta}_j + c_{ij} + u_{itj}, \qquad (1) \\ y_{itj} &= 1[\mathbf{x}_{it}\boldsymbol{\beta}_j + c_{ij} + u_{itj} > 0], \quad t = 1, \ldots, T, \quad j = 1, \ldots, J, \end{aligned}$$

where y_{itj}^* is a continuous latent variable, y_{itj} is the observed binary outcome for unit i in group j in period t, and $1[\cdot]$ is an indicator function equal to one if the expression in brackets is true. The vector of explanatory variables, $\mathbf{x}_{it}$, is $1 \times K$, and $\boldsymbol{\beta}_j$ is a $K \times 1$ group-specific vector of parameters. Define $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ and make the following assumption:

Assumption 1 $u_{itj}|\mathbf{x}_i, c_{ij} \sim u_{itj}$.

The assumption implies that $\mathbf{x}_{it}$ is independent of the idiosyncratic error, but may be correlated with a time-constant group-specific unobserved effect c_{ij}. It also indicates that the observed covariates are strictly exogenous conditional on c_{ij}, i.e. past and future values of $\mathbf{x}_{it}$ do not affect the distribution of y_{it} after accounting for the current values of covariates.

Note that a given cross section unit may appear in different groups in different t. Transitions may occur due to changes in both time-varying covariates and idiosyncratic shocks and may be endogenous with respect to y_{itj}. For example, a shock to the main outcome may affect both $\mathrm{P}(y_{itj} = 1)$ and the probability of belonging to group j in period t. Even if j is constant across t, group sorting is not random if time-constant and/or time-varying unobservables affecting the group assignment are correlated with the unobservables in (1). As discussed in detail below, such endogeneity causes inconsistency in the estimators of $\boldsymbol{\beta_j}$ obtained by estimating (1) separately for each j.

Let d_{it} be a discrete random variable identifying groups, $d_{it} = \{1, 2, \ldots, J\}$. After defining dichotomous indicators for each group as $s_{itj} = 1[d_{it} = j]$, $t = 1, \ldots$, T, $j = 1, \ldots, J$, the outcome for unit i in a given period can be written as

$$y_{it} = \sum_{j=1}^{J} s_{itj} y_{itj}, \qquad t = 1, \ldots, T. \tag{2}$$

Apart from $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$, parameters of interest include partial effects. These can be of two types. We define the unconditional partial effect (PE_j^U) as a change in the probability of success in group j due to an increase in variable x for a randomly selected unit from the population. In the population, the unconditional partial effect of a continuous explanatory variable is

$$PE_{j,k}^U = \frac{\partial \mathrm{P}(y_j = 1|\mathbf{x}, c_j)}{\partial x_k}, \qquad j = 1, \ldots, J. \tag{3}$$

On the other hand, a conditional partial effect (PE_j^C) is a change in the probability of success due to an increase in x for a unit in group j. For a continuous covariate,

$$PE_{j,k}^C = \frac{\partial \mathrm{P}(y = 1|d = j, \mathbf{x}, c_j)}{\partial x_k} = \frac{\partial \mathrm{P}(y_j = 1|d = j, \mathbf{x}, c_j)}{\partial x_k} \qquad j = 1, \ldots, J. \tag{4}$$

Although both effects may be of interest, PE_j^U is often deemed more suitable for cross-group comparisons, whereas PE_j^U is useful when focusing on a particular group.

In practice, c_j is not observed, which makes it impossible to estimate $PE_{j,k}^U$ and $PE_{j,k}^C$. Instead, it is common to estimate average partial effects (APE) that are obtained by ‘averaging’ over the distribution of the unobserved effect, c_j:

$$\begin{aligned} APE^U_{j,k} &= \mathrm{E}_{c_j}\left[\frac{\partial \mathrm{P}(y_j = 1|\mathbf{x}, c_j)}{\partial x_k}\right], \\ APE^C_{j,k} &= \mathrm{E}_{c_j}\left[\frac{\partial \mathrm{P}(y_j = 1|d = j, \mathbf{x}, c_j)}{\partial x_k}\right], \qquad j = 1, \dots, J. \end{aligned} \tag{5}$$

If group assignment is random, then $\mathrm{P}(y_j = 1|\mathbf{x}, c_j) = \mathrm{P}(y_j = 1|d = j, \mathbf{x}, c_j)$, and consistent estimators of model parameters are obtained by estimating (1) separately for each j. However, because of self-selection and other factors sorting into groups may be nonrandom, which causes inconsistency. In this paper, we allow for a possibility that $\mathrm{P}(y_{itj} = 1|d_{it} = j, \mathbf{x}_{it}, c_{ij}) \neq \mathrm{P}(y_{itj} = 1|\mathbf{x}_{it}, c_{ij})$ and discuss how it can be addressed when obtaining consistent estimators of $\boldsymbol{\beta}_j$ and APE. We start by considering a simple case with only two groups and then discuss more general models with $J > 2$, where groups may be ordered or unordered.

2.2 Model for two groups

Let y_j be determined as in equation (1), where $J = 2$. Applications of such models include, for example, examining labor force participation among married and non-married women, as well as estimating the determinants of dropout incidents among economically disadvantaged and other students. Assume that sorting into groups is determined by the value of a latent variable d^*_{it},

$$\begin{aligned} d^*_{it} &= \mathbf{z}_{it}\boldsymbol{\delta} + b_i + v_{it}, \\ d_{it} &= 1 \text{ if } d^*_{it} \leq 0, \\ d_{it} &= 2 \text{ if } d^*_{it} > 0, \end{aligned} \tag{6}$$

where $\mathbf{z}_{it}$ is a $1 \times L$ vector of exogenous variables, b_i is a time-constant unobserved effect, and v_{it} is an idiosyncratic error. Setting the cut point at zero is at no cost, as long as z_{it} contains an intercept. Vector $\mathbf{z}_{it} = (\mathbf{x}_{it}, \mathbf{z}_{it1})$ contains at least one additional variable that is not in $\mathbf{x}_{it}$.[1] Similar to the main equation, define $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT})$ and assume that the following holds:

Assumption 2 $v_{it}|\mathbf{z}_i, b_i \sim v_{it}$.

Hence, $\mathbf{z}_{it}$ is strictly exogenous conditional on b_i, but may be correlated with b_i. This correlation causes an omitted variable problem that has to be resolved before addressing nonrandom sorting. Building upon the work by Mundlak (1978) and Chamberlain (1980), unobserved effects can be modeled as

$$\begin{aligned} c_{ij} &= \bar{\mathbf{z}}_i \boldsymbol{\psi}_{cj} + a_{cij}, \quad j = 1, 2, \\ b_i &= \bar{\mathbf{z}}_i \boldsymbol{\psi}_b + a_{bi}, \end{aligned} \tag{7}$$

where $\bar{\mathbf{z}}_i = \sum_{t=1}^{T} \mathbf{z}_{it}$, and (a_{ci1}, a_{ci2}, b_i) are independent of $\mathbf{z}_i$. This modeling approach has been previously used in both theoretical and applied work (Abrevaya and Dahl, 2008; Jäckle and Himmler, 2010; Papke and Wooldridge, 2008; Semykina, 2018; Semykina and Wooldridge, 2010, 2018; Wooldridge, 1995, among others).[2] Note that although $\mathbf{z}_{it}$ may contain time-constant covariates, equation (7) indicates that their causal effects cannot be distinguished from the impact of c_{ij} and b_i, unless they are independent of the unobserved effects. Nevertheless, it is important to include such variables as controls to prevent inconsistency due to an omitted variable problem.

[1] Strictly speaking, the exclusion restriction is not required for identification. However, when $\mathbf{z}_{it} = \mathbf{x}_{it}$, identification relies exclusively on the functional form of the likelihood function, which is less reliable.

[2] Mundlak (1978) has shown that when this method is used for estimating linear models, the resulting $\hat{\boldsymbol{\beta}}_j$ is identical to the fixed effects estimator of $\boldsymbol{\beta}_j$.

Using (7), equations (1) and (6) can be written as

$$
\begin{aligned}
y_{itj} &= 1[\mathbf{x}_{it}\boldsymbol{\beta}_j + \bar{\mathbf{z}}_i\boldsymbol{\psi}_{cj} + \eta_{itj} > 0], \quad t = 1, \ldots, T, \quad j = 1, 2, \\
d_{it} &= 1 \text{ if } \mathbf{z}_{it}\boldsymbol{\delta} + \bar{\mathbf{z}}_i\boldsymbol{\psi}_b + \epsilon_{it} \leq 0, \\
d_{it} &= 2 \text{ if } \mathbf{z}_{it}\boldsymbol{\delta} + \bar{\mathbf{z}}_i\boldsymbol{\psi}_b + \epsilon_{it} > 0.
\end{aligned} \tag{8}
$$

where $\eta_{itj} = a_{cij} + u_{itj}$, $j = 1, 2$, and $\epsilon_{it} = a_{bi} + v_{it}$. We formulate the following assumption:

Assumption 3

(i) (η_{tj}, ϵ_t) *are independent of* $\mathbf{z}_i$, $j = 1, 2$, $t = 1, \ldots, T$.

(ii) For each t,

$$
\begin{pmatrix} \eta_{tj} \\ \epsilon_t \end{pmatrix} \sim Normal\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \\ \rho_j & 1 \end{bmatrix} \right), \quad j = 1, 2. \tag{9}
$$

(iii) $0 < \frac{1}{T}\sum_{t=1}^{T} \mathrm{P}(d_t = j) < 1$, $j = 1, 2$.

The assumption is stated for the underlying population; hence, subscript i is dropped. However, by random sampling, it also holds for a randomly selected unit from the population. Part (i) imposes strict exogeneity of covariates in (8). The normality assumption in (ii), is rather standard in the literature and permits obtaining formulae for conditional probabilities and partial effects. Because each i can only belong to one group in a given t, $\mathrm{Corr}(\eta_{it1}, \eta_{it2})$ is not defined. Moreover, note that $\mathrm{Corr}(\eta_{it1}, \eta_{is2})$ and $\mathrm{Corr}(\eta_{itj}, \epsilon_{is})$, $t \neq s$, are not specified, but can be (and likely are) different from zero. Finally, part (iii) ensures that there are cross section units in each group in at least some periods in the population.

Under Assumption 3, the two-group model is a switching probit model, which is analogous to a linear switching regression model discussed in the literature (Carrasco, 2001; Lee 1978; Maddala and Nelson, 1975; Maddala 1983; Manski et al., 1992). In the linear case, nonrandom group assignment is usually addressed by constructing a correction term. In binary response models, however, this approach is inapplicable because of the nonlinearity of the conditional mean. Instead, using the properties of normal distributions, we can write

$$\begin{aligned} \eta_{itj} &= \rho_j \epsilon_{it} + e_{itj}, \\ e_{itj}|\mathbf{z}_i, \epsilon_{it} &\sim Normal(1, 1-\rho_j^2), \end{aligned} \tag{10}$$

so that $y_{itj} = 1[\mathbf{x}_{it}\boldsymbol{\beta}_j + \bar{\mathbf{z}}_i\boldsymbol{\psi}_{cj} + \rho_j\epsilon_{it} + e_{itj} > 0]$, $t = 1, \ldots, T$, $j = 1, 2$.

Define $\mathbf{w}_{it} = (\mathbf{x}_{it}, \bar{\mathbf{z}}_i)$, $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j', \boldsymbol{\psi}_{cj}')'$, $\mathbf{q}_{it} = (\mathbf{z}_{it}, \bar{\mathbf{z}}_i)$, and $\boldsymbol{\pi} = (\boldsymbol{\delta}', \boldsymbol{\psi}_{bj}')'$. Then, the conditional probability for $j = 1$, period t, can be written as

$$\begin{aligned} \mathrm{P}(y_{it} = 1|d_{it} = 1, \mathbf{z}_i) &= \frac{\mathrm{P}(-e_{it1} < \mathbf{w}_{it}\boldsymbol{\theta}_1 + \rho_1\epsilon_{it}, \epsilon_{it} \leq -\mathbf{q}_{it}\boldsymbol{\pi}|\mathbf{z}_i)}{\mathrm{P}(\epsilon_{it} \leq -\mathbf{q}_{it}\boldsymbol{\pi}|\mathbf{z}_i)} \\ &= \frac{\int_{-\infty}^{-\mathbf{q}_{it}\boldsymbol{\pi}} \Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}_1 + \rho_1\epsilon}{\sqrt{1-\rho_1^2}}\right)\phi(\epsilon)d\epsilon}{1 - \Phi(\mathbf{q}_{it}\boldsymbol{\pi})}, \end{aligned} \tag{11}$$

and the corresponding conditional probability for $j = 2$ is

$$\mathrm{P}(y_{it} = 1|d_{it} = 2, \mathbf{z}_i) = \frac{\int_{-\infty}^{\mathbf{q}_{it}\boldsymbol{\pi}} \Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}_2 + \rho_2\epsilon}{\sqrt{1-\rho_2^2}}\right)\phi(\epsilon)d\epsilon}{\Phi(\mathbf{q}_{it}\boldsymbol{\pi})}, \tag{12}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are standard normal density and cumulative distribution functions, respectively. Note that $\mathrm{P}(y_{itj} = 1|\mathbf{z}_i) = \Phi(\mathbf{w}_{it}\boldsymbol{\theta}_j)$ and is the same regardless of the number of groups and ordering. These probabilities can be used to obtain APE_j^U and APE_j^C, as will be discussed in Section 3.

2.3 Model for multiple ordered groups

Let the total number of groups, J, exceed two. Using the unobserved effects model in (7), define vectors $\mathbf{q}_{it}$ and $\boldsymbol{\pi}$ as in the previous section. Also, define d_{it}^* and d_{it} as

$$
\begin{aligned}
& d_{it}^* = \mathbf{q}_{it}\boldsymbol{\pi} + \epsilon_{it}, && (13)\\
& d_{it} = j \ \text{ if } \ C_{j-1} < d_{it}^* \leq C_j, \quad j = 1, \ldots, J, \\
& C_0 = -\infty, \ \text{ and } \ C_J = \infty.
\end{aligned}
$$

Such a model is applicable, for example, when the goal is to study the labor force participation or probability of self-employment by age or education level.

Similar to a two-group case, it is convenient to assume that the distribution of ϵ_{it} is normal, which results in an ordered probit model. Formally, let Assumption 3 hold for $j = 1, 2, \ldots, J$, so that the errors are independent of $\mathbf{z}_i$ and have a joint normal distribution. Then, using the argument similar to the one in Section 2.2, we can write

$$
\begin{aligned}
& y_{itj} = 1[\mathbf{w}_{it}\boldsymbol{\theta}_j + \rho_j\epsilon_{it} + e_{itj} > 0], \quad t = 1, \ldots, T, \quad j = 1, \ldots, J, && (14)\\
& e_{itj}|\mathbf{w}_{it}, \epsilon_{it} \sim Normal(1, 1 - \rho_j^2),
\end{aligned}
$$

where $\mathbf{w}_{it}$ and $\boldsymbol{\theta}_j$ are defined as in Section 2.2.

From (13) and (14), the conditional probabilities for each group are

$$
\begin{aligned}
\mathrm{P}(y_{it} = 1|d_{it} = j, \mathbf{z}_i) \ &= \ \frac{\int_{C_{j-1}-\mathbf{q}_{it}\boldsymbol{\pi}}^{C_j-\mathbf{q}_{it}\boldsymbol{\pi}} \Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}_j+\rho_j\epsilon}{\sqrt{1-\rho_j^2}}\right)\phi(\epsilon)d\epsilon}{\Phi(C_j - \mathbf{q}_{it}\boldsymbol{\pi}) - \Phi(C_{j-1} - \mathbf{q}_{it}\boldsymbol{\pi})}, \quad j = 2, \ldots, J-1, \\
& \quad C_0 = -\infty, \qquad C_J = \infty.
\end{aligned}
$$

2.4 Model for unordered multiple groups

In some cases, there may be multiple groups that are not ordered. For example, one might want to study the determinants of job promotion among workers in different occupations. Then, the choice of $d_{it} = j$ can be described in the context of a multinomial response model. To formalize ideas, define

$$d_{itj}^* = \mathbf{q}_{it}\boldsymbol{\pi}_j + \epsilon_{itj}, \quad t = 1, \dots, T, \quad j = 1, \dots, J, \tag{15}$$

where the parameter vector and error term now vary by group.

Following the standard formulation of a multinomial response model, the cross-section unit i will be in group j in period t if it has the highest chance of belonging to that group. In the case of self-selection, choice j is the best option in the available set:

$$d_{it} = j \;\text{ if }\; d_{itj}^* = \max\{d_{it1}^*, d_{it2}^*, \dots, d_{itJ}^*\} \tag{16}$$

The choice in (16) will be made if $\mathbf{q}_{it}\boldsymbol{\pi}_j + \epsilon_{itj} > \mathbf{q}_{it}\boldsymbol{\pi}_l + \epsilon_{itl}$ for all $l \neq j$. It is clearly seen that only differences between d_{itj}^* are identified, so that a reference category needs to be assigned – a feature that is common to all multinomial response models. We formulate the following assumption:

Assumption 4

(i) $(\eta_{tj}, \epsilon_{t1}, \dots, \epsilon_{tJ})$ *are independent of* $\mathbf{z}_i$, *for* $j = 1, \dots, J$, $t = 1, \dots, T$.

(ii) For each t,

$$\begin{pmatrix} \eta_{tj} \\ \epsilon_{t1} \\ \dots \\ \epsilon_{tj} \\ \dots \\ \epsilon_{tJ} \end{pmatrix} \sim Normal \left(\begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \\ \dots \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & \dots & \rho_j & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho_j & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{bmatrix} \right), \quad j = 1, \dots, J. \tag{17}$$

(iii) $0 < \frac{1}{T}\sum_{t=1}^{T} \mathrm{P}(d_t = j) < 1$, $j = 1, \dots, J$.

Note that Assumption 4 imposes restrictions on the variance-covariance matrix. Theoretically, one could allow the variance in part (ii) to be completely unrestricted. However, in practice it is usually necessary to impose restrictions to ensure feasibility of the estimation. In the present context, the imposed restriction are reasonable. Specifically, $\mathrm{Cov}(\epsilon_{tj}, \epsilon_{tl}) = 0$ is effectively true by independence across i because each cross section unit can only belong to one group in a given t. For the same reason, $\mathrm{Cov}(\eta_{tj}, \epsilon_{tl}) = 0$ holds. Importantly, $\mathrm{Cov}(\epsilon_{tj}, \epsilon_{sl})$ and $\mathrm{Cov}(\eta_{tj}, \epsilon_{sl})$, $s \neq l$, are left completely unrestricted, which is consistent with what would be observed in the population. Indeed, these covariances are likely different from zero because of transitions across groups over time.

Define $\tilde{\epsilon}_{itl} = \epsilon_{itj} - \epsilon_{itl}$, and $\tilde{\boldsymbol{\pi}}_l = \boldsymbol{\pi}_j - \boldsymbol{\pi}_l$, for $l \neq j$. Then, under Assumption 4, for group $j = 1$, for example, we obtain

$$\begin{aligned} \mathrm{P}(y_{it} = 1, d_{it} = 1 | \mathbf{z}_i) &= \int_{-\mathbf{w}_{it}\boldsymbol{\theta}_1}^{\infty} \int_{-\mathbf{q}_{it}\tilde{\boldsymbol{\pi}}_2}^{\infty} \dots \int_{-\mathbf{q}_{it}\tilde{\boldsymbol{\pi}}_J}^{\infty} \phi(e_{it1}, \tilde{\epsilon}_2 \dots \tilde{\epsilon}_J; \Sigma) du_1 d\tilde{\epsilon}_2 \dots d\tilde{\epsilon}_J, (18) \\ \mathrm{P}(d_{it} = 1 | \mathbf{z}_i) &= \int_{-\mathbf{q}_{it}\tilde{\boldsymbol{\pi}}_2}^{\infty} \dots \int_{-\mathbf{q}_{it}\tilde{\boldsymbol{\pi}}_J}^{\infty} \phi(\tilde{\epsilon}_2, \dots, \tilde{\epsilon}_J; \tilde{\Sigma}) d\tilde{\epsilon}_2 \dots d\tilde{\epsilon}_J, \end{aligned}$$

where Σ and $\tilde{\Sigma}$ are variance-covariance matrices of vectors $(e_{it1}, \tilde{\epsilon}_2 \dots \tilde{\epsilon}_J)'$ and $(\tilde{\epsilon}_2 \dots \tilde{\epsilon}_J)'$, respectively. Using (18), the conditional probability is obtained as $P(y_{it} = 1 | d_{it} = 1, \mathbf{z}_i) =$

$\frac{P(y_{it}=1,d_{it}=1|\mathbf{z}_i)}{P(d_{it}=1|\mathbf{z}_i)}$. Probabilities $P(y_{it}=1|d_{it}=j,\mathbf{z}_i)$, $j=2,\ldots,J$, are obtained similarly.

Because equation (18) does not have a closed form solution, one would need to numerically evaluate a J-dimensional integral. Although simulated likelihood methods have been helpful in addressing computational difficulties, the estimation may still be infeasible if there are more than four groups. Therefore, we also discuss a different approach.

The unordered multiple groups case can be considered in the context of selection models, where the choice is made between the best option (observed choice) and the second best alternative. Define a binary indicator for group j in period t as

$$\begin{aligned} \omega_{itj} &= 1[\mathbf{q}_{it}\boldsymbol{\pi}_j + \epsilon_{itj} > \bar{d}_{itj}], \\ \bar{d}_{itj} &= \max_{l\neq j}\{\mathbf{q}_{it}\boldsymbol{\pi}_l + \epsilon_{itl}\}, \end{aligned} \tag{19}$$

which can be re-written as

$$\omega_{itj} = 1[\mathbf{q}_{it}\bar{\boldsymbol{\pi}}_j + \bar{\epsilon}_{itj} > 0], \quad t=1,\ldots,T, \quad j=1,\ldots,J, \tag{20}$$

where $\bar{\boldsymbol{\pi}}_j$ is the difference between $\boldsymbol{\pi}_j$ and the vector of parameters that correspond to $\bar{d}_{itj}$, and $\bar{\epsilon}_{itj}$ is the difference between ϵ_{itj} and the error corresponding to $\bar{d}_{itj}$. Because in the unordered case the second best option is not known, $\bar{\boldsymbol{\pi}}_j$ is a weighted average of $\boldsymbol{\pi}_j - \boldsymbol{\pi}_l$, $l \neq j$, where weights depend on the probability that group l is the best alternative to j. Notice that in this model it is not possible to estimate $\boldsymbol{\pi}_j$. Fortunately, this does not affect our ability to consistently estimate parameters $\boldsymbol{\theta}$, which is the main goal of the estimation.

Assumption 5

(i) $(\eta_{tj}, \bar{\epsilon}_{tj})$ *are independent of* $\mathbf{z}_i$, $j = 1\ldots,J$, $t=1,\ldots,T$.

(ii) For each t,

$$\begin{pmatrix} \eta_{tj} \\ \bar{\epsilon}_{tj} \end{pmatrix} \sim Normal \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \\ \rho_j & 1 \end{bmatrix} \right), \quad j = 1, \ldots, J. \tag{21}$$

(iii) $0 < \frac{1}{T} \sum_{t=1}^{T} \mathrm{P}(d_t = j) < 1$, $j = 1, \ldots, J$.

Under Assumption 5, conditional probabilities for each j and t are obtained as

$$\mathrm{P}(y_{it} = 1 | d_{it} = j, \mathbf{z}_i) = \mathrm{P}(y_{it} = 1 | \omega_{itj} = 1, \mathbf{z}_i) = \frac{\int_{-\infty}^{\mathbf{q}_{it}\bar{\boldsymbol{\pi}}_j} \Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}j + \rho_j \bar{\epsilon}}{\sqrt{1-\rho_j^2}}\right) \phi(\bar{\epsilon}) d\bar{\bar{\epsilon}}}{\Phi(\mathbf{q}_{it}\bar{\boldsymbol{\pi}}_j)}. \tag{22}$$

3 Estimation

To estimate the models presented in Section 2, one can use the maximum likelihood estimator (MLE). Full MLE would be an efficient estimator, but it requires specifying conditional density $f(y_{i1}, ..., y_{iT} | d_{i1}, ..., d_{iT}, \mathbf{z}_i)$. Because $y_{i1}, ..., y_{iT}$ are likely not serially independent even after conditioning on $d_{i1}, ..., d_{iT}, \mathbf{z}_i$ (largely due to the presence of the time-constant unobserved effect), the joint density function would generally be very complicated. That would increase computational costs and can make the estimation infeasible, unless additional restrictions on the error variance-covariance matrix are imposed. In this paper, we use a more feasible partial MLE estimator, which only requires specifying the conditional density in a given t.

Consider a general model with $J \geq 2$, but for the moment ignore the second unordered groups model discussed at the end of Section 2.4. For each t, the likelihood function for observation i is

$$\mathcal{L}_{it}(\boldsymbol{\gamma}) = \mathrm{P}_{it,11}^{y_{it}s_{it1}} \cdot \mathrm{P}_{it,01}^{(1-y_{it})s_{it1}} \cdot \ldots \cdot \mathrm{P}_{it,1J}^{y_{it}s_{itJ}} \cdot \mathrm{P}_{it,0J}^{(1-y_{it})s_{itJ}}, \tag{23}$$

where $\mathrm{P}_{it,1j} = \mathrm{P}(y_{it} = 1, d_{it} = j|\mathbf{z}_i;\boldsymbol{\gamma})$, $\mathrm{P}_{it,0j} = \mathrm{P}(y_{it} = 0, d_{it} = j|\mathbf{z}_i;\boldsymbol{\gamma})$, $j = 1,...,J$, $\boldsymbol{\gamma} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_J, \boldsymbol{\pi}, \rho_1, \ldots, \rho_J)'$ for the models in Sections 2.2 and 2.3, and $\boldsymbol{\gamma} = (\boldsymbol{\theta}'_1, \ldots,$ $\boldsymbol{\theta}'_J$, $\bar{\boldsymbol{\pi}}_1, \ldots$, $\bar{\boldsymbol{\pi}}_J$, $\rho_1, \ldots, \rho_J)'$ for the first model in Section 2.4.

Joint probabilities for the two-group model are

$$\begin{aligned} \mathrm{P}(y_{it} = 1, d_{it} = j|\mathbf{z}_i;\boldsymbol{\gamma}) &= \int_{-\infty}^{-\mathbf{q}_{it}\boldsymbol{\pi}} \Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}_j + \rho_j\epsilon}{\sqrt{1-\rho_j^2}}\right)\phi(\epsilon)d\epsilon, \\ \mathrm{P}(y_{it} = 0, d_{it} = j|\mathbf{z}_i;\boldsymbol{\gamma}) &= \int_{-\infty}^{-\mathbf{q}_{it}\boldsymbol{\pi}} \left[1-\Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}_j + \rho_j\epsilon}{\sqrt{1-\rho_j^2}}\right)\right]\phi(\epsilon)d\epsilon, \quad j = 1, 2. \end{aligned} \tag{24}$$

For ordered multiple groups, the probabilities are

$$\begin{aligned} \mathrm{P}(y_{it} = 1, d_{it} = j|\mathbf{z}_i;\boldsymbol{\gamma}) &= \int_{C_{j-1}-\mathbf{q}_{it}\boldsymbol{\pi}}^{C_j-\mathbf{q}_{it}\boldsymbol{\pi}} \Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}_j + \rho_j\epsilon}{\sqrt{1-\rho_j^2}}\right)\phi(\epsilon)d\epsilon, \\ \mathrm{P}(y_{it} = 0, d_{it} = j|\mathbf{z}_i;\boldsymbol{\gamma}) &= \int_{C_{j-1}-\mathbf{q}_{it}\boldsymbol{\pi}}^{C_j-\mathbf{q}_{it}\boldsymbol{\pi}} \left[1-\Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}_j + \rho_j\epsilon}{\sqrt{1-\rho_j^2}}\right)\right]\phi(\epsilon)d\epsilon, \\ & C_0 = -\infty, \quad C_J = \infty, \quad j = 1, \ldots, J. \end{aligned} \tag{25}$$

In the unordered case, the first equation in (18) specifies $\mathrm{P}(y_{it} = 1, d_{it} = 1|\mathbf{z}_i;\boldsymbol{\gamma})$ for the first model in Section 2.4. Probabilities for $j = 2, \ldots, J$ are obtained similarly. Changing the limits of integration permits computing $\mathrm{P}(y_{it} = 0, d_{it} = j|\mathbf{z}_i;\boldsymbol{\gamma})$.

Partial MLE is obtained by solving the following optimization problem:

$$\max_{\boldsymbol{\gamma}} \sum_{i=1}^{N}\sum_{t=1}^{T} \ln \mathcal{L}_{it}(\boldsymbol{\gamma}). \tag{26}$$

The resulting partial MLE is consistent under Assumptions 3, 4, and 5.2 for the parameters in the corresponding models if T is fixed, $N \to \infty$, and standard MLE regularity conditions hold (see, for example, Wooldridge, 2010, Chapter 13). However, the information matrix equality does not hold because the likelihood function is specified for a given t. The score vectors are generally serially correlated, so that it is necessary to obtain fully-robust standard errors that account for serial correlation. Subsequently, inference can be performed using the fully-robust t and Wald test statistics. The likelihood ratio test can also be used.

Several null hypotheses may be of particular interest. For example, to check whether sorting is random, and the usual group-by-group estimation is valid, one can test $H_0 : \rho_1 = \ldots = \rho_J = 0$. Also, the equality of the coefficients in two or more groups can be tested either for each explanatory variable separately, or for the entire vector of parameters, $\boldsymbol{\theta}_j$.

As discussed in Section 2.4, parameters in the unordered multiple groups model can be estimated using a simpler approach that relies on Assumption 4.2. For each group j, we then write the likelihood function for observation i in period t as

$$\mathcal{L}_{it}(\boldsymbol{\gamma}_j) = \mathrm{P}_{it,11}^{y_{it}\omega_{itj}} \cdot \mathrm{P}_{it,01}^{(1-y_{it})\omega_{itj}} \cdot \mathrm{P}_{it,0}^{(1-\omega_{itj})}, \tag{27}$$

where

$$\begin{aligned}
\mathrm{P}_{it,11} &\equiv \mathrm{P}(y_{it} = 1, \omega_{itj} = 1|\mathbf{z}_i; \boldsymbol{\gamma}_j) = \int_{-\infty}^{\mathbf{q}_{it}\bar{\boldsymbol{\pi}}_j} \Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}j + \rho_j\bar{\epsilon}}{\sqrt{1-\rho_j^2}}\right)\phi(\bar{\epsilon})d\bar{\bar{\epsilon}}, \qquad (28)\\
\mathrm{P}_{it,01} &\equiv \mathrm{P}(y_{it} = 0, \omega_{itj} = 1|\mathbf{z}_i; \boldsymbol{\gamma}_j) = \int_{-\infty}^{\mathbf{q}_{it}\bar{\boldsymbol{\pi}}_j} \left[1 - \Phi\left(\frac{\mathbf{w}_{it}\boldsymbol{\theta}j + \rho_j\bar{\epsilon}}{\sqrt{1-\rho_j^2}}\right)\right]\phi(\bar{\epsilon})d\bar{\bar{\epsilon}},\\
\mathrm{P}_{it,0} &\equiv \mathrm{P}(\omega_{itj} = 0|\mathbf{z}_i; \boldsymbol{\gamma}_j) = 1 - \Phi(\mathbf{q}_{it}\bar{\boldsymbol{\pi}}_j).
\end{aligned}$$

Subsequently, $\boldsymbol{\gamma}_j$ can be estimated separately for each j by partial MLE. The limitation of this estimation approach is that hypothesis testing is complicated when parameters from different groups are involved. A relatively simple solution is to use panel bootstrap, where all $\boldsymbol{\gamma}_j$ are estimated using the same bootstrap sample in each replication. Then, it becomes relatively straightforward to obtain estimators of covariances and test statistics.

To obtain an estimator of APE_j^C, note that from (7) we can write

$$\mathrm{P}(y_j = 1|d = j, \mathbf{x}, c_j) = \mathrm{P}(y_j = 1|d = j, \mathbf{z}, \bar{\mathbf{z}}, a_{cj}, a_{bj}, v), \tag{29}$$

where we also use the fact that d is a deterministic function of $(\mathbf{z}, \bar{\mathbf{z}}, a_{bj}, v)$ in the population. After interchanging the integration and differentiation, it follows that conditional APE of a continuous variable x_k in group j is

$$\frac{\partial \mathrm{E}_{\bar{\mathbf{z}}, a_{cj}, a_{bj}, v}[\mathrm{P}(y_j = 1 | d = j, \mathbf{z}, \bar{\mathbf{z}}, a_{cj}, a_{bj}, v)]}{\partial x_k}. \tag{30}$$

Hence, for $j = 2$ in a two-group model, conditional APE can be estimated as

$$\begin{aligned} \widehat{APE}_{2,k}^{C} = & \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\hat{\delta}_k \cdot \frac{\phi(\mathbf{q}_{it}\hat{\boldsymbol{\pi}})}{\Phi(\mathbf{q}_{it}\hat{\boldsymbol{\pi}})} \cdot \Phi\left(\frac{\mathbf{w}_{it}\hat{\boldsymbol{\theta}}_2 + \hat{\rho}_2 \mathbf{q}_{it}\hat{\boldsymbol{\pi}}}{\sqrt{1 - \hat{\rho}_2^2}} \right) \right. \\ & \left. + \frac{1}{\Phi(\mathbf{q}_{it}\hat{\boldsymbol{\pi}})} \cdot \frac{\hat{\beta}_{2k}}{\sqrt{1 - \hat{\rho}_2^2}} \cdot \int_{-\infty}^{\mathbf{q}_{it}\hat{\boldsymbol{\pi}}} \phi\left(\frac{\mathbf{w}_{it}\hat{\boldsymbol{\theta}}_2 + \hat{\rho}_2 \epsilon}{\sqrt{1 - \hat{\rho}_2^2}} \right) \phi(\epsilon) d\epsilon - \hat{\delta}_k \cdot \frac{\phi(\mathbf{q}_{it}\hat{\boldsymbol{\pi}})}{\Phi(\mathbf{q}_{it}\hat{\boldsymbol{\pi}})} \cdot \hat{\mathrm{P}}_{it,12}^{C} \right], \end{aligned} \tag{31}$$

where $\hat{\delta}_k$, $\hat{\beta}_{2k}$, $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\theta}}_2$, and $\hat{\rho}_2$, are the estimators of δ_k, β_{2k}, $\boldsymbol{\pi}$, $\boldsymbol{\theta}_2$, and ρ_2, respectively, and $\hat{\mathrm{P}}_{it,12}^{C}$ is the estimator of $\mathrm{P}(y_{it} = 1 | d_{it} = 2, \mathbf{z}_i)$, which is defined in equation (12). Correspondingly, $\widehat{APE}_{1,k}$ is obtained by replacing $\hat{\boldsymbol{\pi}}$ and $\hat{\delta}_k$ with $-\hat{\boldsymbol{\pi}}$ and $-\hat{\delta}_k$, respectively, and changing $\hat{\boldsymbol{\theta}}$ and $\hat{\rho}$ subscripts to one. Notice that when the group assignment is random, the partial effects on the conditional probabilities are the same as the unconditional partial effects. However, they are different when $\rho_j \neq 0$.

For the ordered groups model, conditional APE for each j can be estimated using

$$\begin{aligned} \widehat{APE}_{j,k}^{C} = & \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \frac{\hat{\delta}_k}{\Phi(\hat{\alpha}_j) - \Phi(\hat{\alpha}_{j-1})} \cdot \left[\phi(\hat{\alpha}_{j-1}) \Phi\left(\frac{\mathbf{w}_{it}\hat{\boldsymbol{\theta}}_j + \hat{\rho}_j \hat{\alpha}_{j-1}}{\sqrt{1 - \hat{\rho}_j^2}} \right) \right. \right. \\ - & \left. \phi(\hat{\alpha}_j) \Phi\left(\frac{\mathbf{w}_{it}\hat{\boldsymbol{\theta}}_j + \hat{\rho}_j \hat{\alpha}_j}{\sqrt{1 - \hat{\rho}_j^2}} \right) \right] \\ + & \frac{1}{\Phi(\hat{\alpha}_j) - \Phi(\hat{\alpha}_{j-1})} \cdot \frac{\hat{\beta}_{jk}}{\sqrt{1 - \hat{\rho}_j^2}} \cdot \int_{-\hat{\alpha}_{j-1}}^{\hat{\alpha}_j} \phi\left(\frac{\mathbf{w}_{it}\hat{\boldsymbol{\theta}}_j + \hat{\rho}_j \epsilon}{\sqrt{1 - \hat{\rho}_j^2}} \right) \phi(\epsilon) d\epsilon \\ + & \left. \hat{\delta}_k \cdot \frac{\phi(\hat{\alpha}_j) - \phi(\hat{\alpha}_{j-1})}{\Phi(\hat{\alpha}_j) - \Phi(\hat{\alpha}_{j-1})} \cdot \hat{\mathrm{P}}_{it,1j}^{C} \right\}, \\ & \hat{C}_0 = -\infty, \quad \hat{C}_J = \infty, \quad \hat{\alpha}_j = C_j - \mathbf{q}_{it}\hat{\boldsymbol{\pi}}, \end{aligned} \tag{32}$$

where $\hat{C}_j$ and $\hat{\alpha}_j$ are the estimators of C_j and α_j, respectively, and $\hat{\mathrm{P}}^C_{it,1j}$ is the estimator of $\mathrm{P}(y_{it} = 1|d_{it} = j, \mathbf{z}_i)$ defined in equation (15).

Given the complexity of the conditional probability function for multiple unordered groups, APE^C_j for the first model in Section 2.4 would have to be evaluated numerically. However, APE^C_j for the second unordered groups model can be estimated using (31) after replacing $\hat{\boldsymbol{\pi}}$ with $\hat{\boldsymbol{\pi}}_j$.

From (31) and (32), it is seen that the sign of $\widehat{APE}^C_{j,k}$ does not necessarily coincide with the sign of β_{jk}. When x_k increases, it affects not only the probability that $y_j = 1$, but also the probability that $d = j$. Consequently, more or fewer units are induced into group j, so that the size and composition of the group changes. Hence, the direction of the change in $\mathrm{P(y_j = 1|d = j)}$ depends on both β_{jk} and δ_k (or, δ_{jk}) and is uncertain.

Unconditional APE can be estimated similarly. In all models,

$$\widehat{APE}^U_{j,k} = \hat{\beta}_{jk} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \phi(\mathbf{w}_{it} \hat{\boldsymbol{\theta}}_j), \quad j = 1, \ldots, J, \tag{33}$$

for a continuous variable x_k.

Average partial effects of discrete variables (e.g. binary indicators) are obtained as average changes in estimated probabilities. For a discrete variable h in group j

$$\widehat{APE}^M_{j,h} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\widehat{\mathrm{P}}^{M1}_{it,1j} - \widehat{\mathrm{P}}^{M0}_{it,1j} \right], \quad M = U, C, \tag{34}$$

where $\widehat{\mathrm{P}}^{U1}_{it,1j} = \mathrm{P}(y_{it} = 1|\mathbf{z}^l_i; \hat{\boldsymbol{\gamma}})$, $\widehat{\mathrm{P}}^{C1}_{it,1j} = \mathrm{P}(y_{it} = 1|d_{it} = j, \mathbf{z}^l_i; \hat{\boldsymbol{\gamma}})$, $l = 0, 1$, for $\mathbf{z}^l_i = (x^l_{it}, z_{it,1}, \bar{z}_i)$, $x^1_{it} = (x_{it,1}, \ldots, x_{it,h-1}, x^1_{it,h}, x_{it,h+1}, \ldots, x_{it,k})$, and $x^0_{it} = \Big(x_{it,1}, \ldots, x_{it,h-1}, x^0_{it,h},$ $x_{it,h+1}, \ldots, x_{it,k}, \Big)$.

Note that (34) can also be used to obtain conditional APE for continuous variables. One can simply consider a particular (e.g. one unit) increase in x_k from a given value, such as the sample mean of x_k. Given the complexity of formulas in (31) and (32), using

(34) may be preferred. It appears to be especially attractive when obtaining APE_j^C in a model with multiple unordered groups.

In the above, estimators of APE are obtained by averaging over the distribution of all covariates other then the one whose effect is being estimated. Alternatively, one can obtain APE evaluated at particular values of other explanatory variables ($\widetilde{\mathbf{z}}$), such as sample means or median values. Then, equation (34), for example, would become

$$\widetilde{APE}_{j,h}^{M} = \widetilde{\mathrm{P}}_{1j}^{M1} - \widetilde{\mathrm{P}}_{1j}^{M0}, \quad M = U, C, \tag{35}$$

where $\widetilde{\mathrm{P}}_{1j}^{U1} = \mathrm{P}(y = 1|\widetilde{\mathbf{z}}^l; \hat{\boldsymbol{\gamma}})$, $\widetilde{\mathrm{P}}_{1j}^{C1} = \mathrm{P}(y = 1|d = j, \widetilde{\mathbf{z}}^l; \hat{\boldsymbol{\gamma}})$, $l = 0, 1$, for some fixed values $\widetilde{\mathbf{z}}^l = (\widetilde{x}^l, \widetilde{z}_1, \widetilde{\bar{z}})$, $\widetilde{x}^1 = (\widetilde{x}_1, \ldots, \widetilde{x}_{h-1}, x_h^1, \widetilde{x}_{h+1}, \ldots, \widetilde{x}_k)$ and $\widetilde{x}^0 = (\widetilde{x}_1, \ldots, \widetilde{x}_{h-1}, x_h^0, \widetilde{x}_{h+1}, \ldots, \widetilde{x}_k)$. APE of continuous covariates are obtained similarly.

4 Monte Carlo Simulations

To study the performance of proposed estimators in finite samples we conduct limited Monte Carlo experiments. In addition to the methods discussed in Section 3, parameters in each group were also estimated by pooled probit, which is a commonly used method in applied research. In every regression, the list of covariates was augmented by variable time means, $\bar{z}_i$. Hence, the focus is on assessing the gains from accounting for nonrandom sorting.

Data were simulated for a two-group model, a model with three ordered groups, and the one with three unordered groups. Explanatory variables are $(1, x_{it}, \bar{x}_i, \bar{z}_i)$ in the main equations, and $(1, x_{it}, z_{it}, \bar{x}_i, \bar{z}_i)$ in the sorting equations. The covariates are generated as

$$\begin{aligned} x_{it} &= b_{i1} + \zeta_{it1}, \qquad (36) \\ z_{it} &= b_{i2} + \zeta_{it2}, \end{aligned}$$

$$\bar{x}_i = \sum_{t=1}^{T} x_{it}, \quad \bar{z}_i = \sum_{t=1}^{T} z_{it}, \tag{37}$$

where b_{ij} are independent across i, $b_{ij} \sim Normal(0, \sigma_b^2)$, $j = 1, 2$, and $\mathrm{Corr}(b_{i1}, b_{i2}) = 0.25$. Correspondingly, $zeta_{itj}$ are independent across i and t, $\zeta_{itj} \sim Normal(0, \sigma_\zeta^2)$, $j = 1, 2$, $\sigma_b^2 + \sigma_\zeta^2 = 1$, and $\frac{\sigma_b^2}{\sigma_b^2 + \sigma_\zeta^2} = 0.5$

The coefficients on time means are set at $\xi_{cj} = (-0.3, -0.3)'$, $j = 1, 2$, $\xi_b = (0.3, 0.3)'$ in all models. However, other population parameters vary by the model to ensure that cross-section units are approximately equally distributed across groups. In a two-group model, y_{itj} and d_{it} are generated as in (8), using $\beta_1 = (1, -1)'$, $\beta_2 = (0.5, 1)'$, $\delta = (0.1, 0.5, 1)'$. The response variables for the model with three ordered groups are created using $\beta_1 = (-0.5, -1)'$, $\beta_2 = (0.2, -2)'$, $\beta_3 = (-0.2, 2)'$, $\delta = (0.5, 0.5, 1)'$, and cut points $C_1 = -0.3$, $C_2 = 1.2$. In the model with three unordered groups, the parameters are set at $\beta_1 = (-0.5, -1)'$, $\beta_2 = (1, -2)'$, $\beta_3 = (1, 2)'$, $\delta_2 = (-0.5, 0.5, 1)'$, $\delta_2 = (-0.5, -0.5, 1.2)'$, and $j = 1$ is a base group.

For each j, error terms were generated as $\eta_{itj} = a_{cij} + u_{itj}$, $\epsilon_{it} = a_{bi} + v_{it}$, where $a_{cij} \sim Normal(0, \sigma_a^2)$, $a_{bi} \sim Normal(0, \sigma_a^2)$, $u_{itj} \sim Normal(0, \sigma_u^2)$, $v_{it} \sim Normal(0, \sigma_v^2)$, $\sigma_a^2 + \sigma_u^2 = \sigma_a^2 + \sigma_v^2 = 1$, $\frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_v^2} = 0.5$, and $\mathrm{Corr}(a_{cij}, a_{bi}) = \mathrm{Corr}(u_{itj}, v_{it}) = \rho_j$. In the two-group model, data are simulated using $\rho_1 = -0.5$, $\rho_2 = 0.5$, while in both three-group models the correlations were $\rho_1 = 0.5$, $\rho_2 = 0.5$, $\rho_3 = -0.5$. We also generated data for $\rho_j = 0$, $\forall j$. Simulations were done for $T = 3$, $N = 500$, using 1000 replications.

Simulation results are presented in Tables 1-4. In all tables, PMLE is the partial MLE estimator discussed in Section 3. As seen in Table 1, probit and joint PMLE estimators have small biases in a two-group model when $\rho_j = 0$. However, probit estimators have smaller standard errors and, therefore, smaller root mean-square errors (RMSE). When ρ_j are different from zero, joint PMLE estimators still have small biases. Also, their standard errors decrease compared to $\rho_j = 0$ case. In contrast, biases tend to be large for probit, and so do RMSE.

For the model with three ordered groups (Tables 2), results are a qualitatively the same. Joint PMLE outperforms probit in terms of smaller biases and RMSE when $\rho_j \neq 0$, although it has larger standard errors, on average. In the model with three unordered groups (Tables 3), both joint PMLE and PMLE for the second ("best alternative") model in Section 2.4 have smaller biases than probit when error correlation is different form zero. A notable exception is a relatively poor performance of the "best alternative" PMLE for estimating parameters in group $j = 1$ (noticeably larger biases and standard errors). It is not entirely clear what causes such a poor performance. However, it appears the "best alternative" PMLE is generally less reliable than the joint PMLE.

Table 4 displays simulation outcomes or the parameters in the sorting equations in all three models. All results are obtained using the joint PMLE method. In all cases biases are very small. Standard errors and RMSE of the estimators of slope parameters are also small. The estimators of the correlation coefficients tend to have larger standard errors suggesting that it is hard to estimate ρ_j with a high degree of precision.

5 Empirical Application

To illustrate the presented theoretical argument with an empirical example, we study the determinants of labor force participation among married and unmarried white women with and without children. In total, there are four groups that are likely characterized by heterogeneous effects: married women with children, married women without children, and unmarried women with and without children. Because previous studies find that fertility decisions are endogenous with respect to labor supply and labor force participation decisions, sorting into different groups is expected to be nonrandom. The errors in the main and sorting equations are likely correlated, which implies that the methodology presented above should be helpful. Given that there are four unordered groups, we use estimators presented in Section 2.4.

To perform estimation, we use data from the National Longitudinal Survey of Youth, 1979 (NLSY79). The initial sample is representative of all individuals who were 14 to 22 years old in 1979. To maximize the sample size we utilize 1990-1994 waves of the survey because the response rate was relatively high in those years (90% or higher). In 1990, all respondents were at least 25 years old, and the age of the oldest respondent in 1994 was 37. Because the supplemental sample of white women was discontinued in 1991, only the respondents from the main and white poor cross-section samples were included. Women in the military sample and those working in a family business were excluded. Limiting the data to female respondents who participated in all five waves of the survey (1990-1994) results in a balanced sample of 2,114 white women. After dropping the observations with missing information on any of the variables used in the analysis, the final sample includes 1,933 women, and a total of 9,577 person-year observations. About 25.6% of the women transitioned from one group to another at least once during the considered period.

The main dependent variable is an indicator equal to one if the woman worked for at least some time during the period since the last interview. The list of explanatory variables includes age, education, and urban location indicator. To control for individual differences in cognitive ability we include the woman's score on the Armed Forces Qualification Test (AFQT), which was administered in 1979. The AFQT score is standardized to have a zero mean and unit variance in the sample. Vector $\bar{z}_i$ includes the individual time mean of the urban indicator. For most women, education remained constant over time, which is why education was excluded from $\bar{z}_i$. In other words, we assume that after accounting for innate ability (measured by AFQT score), education is not correlated with the unobserved effect. The time mean of age was also omitted from $\bar{z}_i$ due to perfect collinearity with year dummies.[3] A year-specific intercept is included in all equations.

As mentioned earlier, it is necessary to have an exclusion restriction to ensure the

[3] Because age increases by one every year for all women, the individual time mean of age is not identified when year dummies are included.

reliability of the estimators. Such a restriction can be obtained by assuming that the probability of working is determined by economic factors (e.g. skills and educational qualifications), but personality traits may be of minor or no importance. On the other hand, personality may influence the probability of marriage and likelihood of having children. In the context of the presented analysis, we include self-esteem, a measure of risk aversion, as well as the respondent's ideal and desired number of children in the sorting equations, but not in the main (employment) equation. The self-esteem measure, developed by Rosenberg (Rosenberg, 1965), is aimed to assess the degree of approval or disapproval toward oneself. In the sample, the self-esteem measure is standardized to have a zero mean and unit variance. As a measure of risk aversion, we use an indicator that equals one if the woman responded affirmatively to the following question: "Other than for a minor traffic violation, have you ever been stopped by the police, but not picked up or arrested?" Because more frequent encounters with police indicate riskier behaviors, an affirmative answer indicates lower risk aversion.[4] This question was asked in 1980 and, hence, the corresponding measure is time constant. Similar to the employment equation, $\bar{z}_i$ includes the individual time mean of the urban indicator.

Summary statistics are presented in Table 5. As seen in the Table, women without children are much more likely to be employed and tend to have more years of schooling. They also on average are slightly younger than women with children and tend to reside in urban locations, especially if not married. Women without children are more likely to engage in risky behaviors (the percent stopped by police is higher). The most risk averse group is married women with children. Unmarried women with children have the lowest AFQT and self-esteem scores among all four groups. There are no discernible differences in the ideal number of kids across groups, but the number of desired children tends to be slightly higher among married women with children.

To obtain main results, the employment equation was first estimated separately for

[4] Similar measures were used by Fairlie (2002) and Semykina (2018) in studies of self employment.

each group of women by pooled probit. Subsequently, the same equations were estimated using the two methods described in Section 2.4. Estimated coefficients and standard errors are presented in Table 6. As expected, estimates vary by the estimation method, and in some cases differences are substantial. For example, the estimated effect of AFQT among non-married women without children is noticeably smaller for both partial MLE methods than for probit. However, PMLE produced slightly larger AFQT coefficients in equations for married women. Similarly, the estimated effects of education among non-married women with and without children are substantially larger when using PMLE.

Differences in coefficient estimates translate into qualitatively similar differences in estimated APE (Table 7). It is also interesting to compare conditional (last column in Table 7) and unconditional APE (first three columns in Table 7). The differences are especially stark for married women without children. In this group, conditional APE are substantially larger in magnitude and do not always agree in their signs with APE^U. The results in Table 8 help to understand some of these differences. For example, age has a negative effect on employment among married women without children, but it also reduces the probability of being married without kids. Hence, as age increases, the proportion of older women without children decreases, so that the overall effect on employment is slightly positive. On the other hand, the coefficient on education is positive in both employment and sorting equations. Therefore, when education increases, more of the educated women are induced into being married without children, which results in a larger positive conditional APE of education in this group of women. The argument is similar for the urban location and AFQT score.

6 Conclusion

This paper discusses the methodology for consistently estimating heterogeneous parameters in binary response panel data models. In addition to a two-group case, we consider estimating parameters for multiple heterogeneous groups, which may be ordered or unordered. Simulations show that considered methods perform well in finite samples. The computed biases remain small when the correlation between errors in the main and sorting equations increases. The RMSE are also smaller than probit RMSE when error correlations are different from zero. As an illustration, we estimate heterogeneous effects on employment outcomes of married and non-married women with and without children using NLSY79 data. We find that accounting for nonrandom group sorting produces different results as compared to simple group-by-group estimation. Moreover, conditional APE can be consistently estimated only when the full information set is utilized.

The proposed methods can be used for estimating heterogeneous effects using cross-section data. It would correspond to a special case with $T = 1$. Obviously, the Mundlak-Chamberlain model of the unobserved effect cannot be used in such a setting. Instead, one would need to include a sufficient set of controls to avoid inconsistencies resulting from an omitted variable problem.

References

Abrevaya, Jason, and Christian M. Dahl, 2008, The Effects of Birth Inputs on Birthweight. *Journal of Business and Economic Statistics* 26, 379-97.

Basu, Anirban, 2014, Estimating Person-Centered Treatment (PeT) Effects Using Instrumental Variables: An Application to Evaluating Prostate Cancer Treatments. *Journal of Applied Econometrics* 29, 671-691.

Bjorklund and Moffitt, 1987, The Estimation of Wage Gains and Welfare Gains in Self-Election Models. *Review of Economics and Statistics* 69(1), 42-49.

Carrasco, Raquel, 1999, Transitions to and from Self-Employment in Spain: An Empirical Analysis. *Oxford Bulletin of Economics and Statistics* 61(3), 315-41.

Chamberlain, G., 1980, Analysis with qualitative data. *Review of Economic Studies* 47, 225-238.

Fairlie, Robert W., 2002, Drug Dealing and Legitimate Self-Employment. *Journal of Labor Economics* 20(3), 538-567.

Goldfeld, S.M. and R.E. Quandt, 1973, *Nonlinear Methods in Econometrics.* Amsterdam: North Holland.

Heckman, James J., 1979, Sample Selection Bias as a Specification Error. *Econometrica* 47(1), 153-61.

Heckman, James J., Sergio Urzua and Edward Vytlacil, 2006, Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics* 88(3), 389-432.

Jäckle, Robert, and Oliver Himmler, 2010, Health and Wages: Panel Data Estimates Considering Selection and Endogeneity. *Journal of Human Resources* 45(2), 364-406.

Kyriazidou, E., 1997, Estimation of a panel data sample selection model. *Econometrica* 65, 1335-1364.

Lee, L.F., 1978, Unionism and Wage Rates: A Simultaneous Equation Model with Qualitative and Limited Dependent Variables. *International Economic Review* 19, 415-433.

Maddala G.S. and F. Nelson, 1975, Switching Regression Models with Exogenous and Endogenous Switching. *Proceedings of the American Statistical Association* (Business and Economics Section), 423-426.

Maddala, G.S., 1983, *Limited Dependent Variable and Qualitative Variables in Econometrics.* Cambridge, U.K.: Cambridge University Press.

Manski, C., D. Sandefur, S. McLanahan, and D. Powers, 1992, Alternative Estimates of Family Structure During Adolescence on High School Graduation, *Journal of the American Statistical Association* 87, 25-37.

Mundlak, Yair, 1978, On the Pooling of Time Series and Cross Section Data. *Econometrica* 46(1), 69-85.

Newey, W.K., 2009, Two-step series estimation of sample selection models. *Econometrics Journal* 12, S217-S229.

Semykina, A., 2018, Self-Employment among Women: Do Children Matter More Than We Previously Thought? *Journal of Applied Econometrics*, April/May 2018, vol. 33, no. 3, pp. 416-434.

Semykina, A. and J.M. Wooldridge, 2018, Binary Response Panel Data Models with Sample Selection and Self Selection. *Journal of Applied Econometrics*, March 2018, vol. 33, no. 2, pp. 179-197.

Vella, Frank, 1988, Generating Conditional Expectations from Models with Selectivity Bias. *Economics Letters* 28, 97-103.

Wooldridge, J.M., 1995, Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions. *Journal of Econometrics* 68, 115–132.

Table 1: Simulation results for J=2 ($T = 3$, $N = 500$)

	$\rho_1 = \rho_2 = 0$			$\rho_1 = -0.5$, $\rho_2 = 0.5$		
	Bias	Avg. Std. Err.	RMSE	Bias	Avg. Std. Err.	RMSE
β_{01}						
Probit	0.010	0.089	0.089	-0.288	0.081	0.299
Joint PMLE	0.003	0.163	0.165	-0.018	0.123	0.119
β_{11}						
Probit	-0.008	0.101	0.100	0.030	0.097	0.102
Joint PMLE	-0.001	0.105	0.104	-0.003	0.094	0.095
β_{02}						
Probit	0.005	0.078	0.080	-0.367	0.075	0.375
Joint PMLE	-0.011	0.163	0.162	-0.011	0.134	0.135
β_{12}						
Probit	0.006	0.098	0.098	-0.034	0.096	0.101
Joint PMLE	-0.003	0.103	0.102	0.003	0.092	0.091

Table 2: Simulation results for J=3, ordered groups ($T = 3$, $N = 500$)

	$\rho_1 = \rho_2 = \rho_3 = 0$			$\rho_1 = \rho_2 = 0.5$, $\rho_3 = -0.5$		
	Bias	Avg. Std. Err.	RMSE	Bias	Avg. Std. Err.	RMSE
β_{01}						
Probit	-0.004	0.113	0.116	-0.592	0.130	0.607
Joint PMLE	-0.004	0.425	0.267	-0.022	0.523	0.270
β_{11}						
Probit	-0.020	0.128	0.138	-0.173	0.137	0.226
Joint PMLE	-0.011	0.174	0.144	-0.014	0.241	0.161
β_{02}						
Probit	0.009	0.087	0.087	0.006	0.093	0.094
Joint PMLE	0.006	0.113	0.087	0.003	0.104	0.086
β_{12}						
Probit	-0.031	0.201	0.211	-0.419	0.238	0.492
Joint PMLE	-0.012	0.321	0.213	-0.048	0.566	0.303
β_{03}						
Probit	-0.007	0.117	0.122	-0.560	0.128	0.576
Joint PMLE	-0.007	0.357	0.278	-0.025	0.346	0.272
β_{13}						
Probit	0.037	0.187	0.201	0.233	0.202	0.310
Joint PMLE	0.012	0.278	0.201	0.016	0.343	0.230

Table 3: Simulation results for J=3, unordered groups ($T = 3$, $N = 500$)

	$\rho_1 = \rho_2 = \rho_3 = 0$			$\rho_1 = \rho_2 = 0.5$, $\rho_3 = -0.5$		
	Bias	Avg. Std. Err.	RMSE	Bias	Avg. Std. Err.	RMSE
β_{01}						
Probit	-0.009	0.081	0.084	0.355	0.076	0.363
Joint PMLE	0.020	0.344	0.357	0.062	0.301	0.321
Best alt. PMLE	0.091	0.471	0.694	0.211	0.453	0.704
β_{11}						
Probit	-0.015	0.126	0.127	-0.068	0.128	0.148
Joint PMLE	0.029	0.143	0.134	0.012	0.151	0.148
Best alt. PMLE	0.162	0.197	0.238	0.129	0.213	0.233
β_{02}						
Probit	0.025	0.140	0.146	0.452	0.161	0.484
Joint PMLE	0.018	0.290	0.296	0.024	0.320	0.332
Best alt. PMLE	0.019	0.297	0.302	0.015	0.331	0.341
β_{12}						
Probit	-0.035	0.201	0.208	-0.178	0.214	0.287
Joint PMLE	-0.008	0.211	0.208	-0.017	0.239	0.245
Best alt. PMLE	-0.008	0.212	0.209	-0.012	0.242	0.248
β_{03}						
Probit	0.017	0.129	0.138	-0.348	0.117	0.368
Joint PMLE	0.003	0.256	0.262	-0.006	0.205	0.212
Best alt. PMLE	0.004	0.264	0.271	0.003	0.211	0.218
β_{13}						
Probit	0.054	0.200	0.221	0.041	0.199	0.224
Joint PMLE	0.026	0.209	0.218	0.027	0.197	0.214
Best alt. PMLE	0.026	0.210	0.219	0.027	0.198	0.214

Table 4: Simulation results for parameters in sorting equations ($T = 3$, $N = 500$)

	Bias	Avg. Std. Err.	RMSE	Bias	Avg. Std. Err.	RMSE
J=2						
	$\rho_1 = \rho_2 = 0$			$\rho_1 = -0.5$, $\rho_2 = 0.5$		
δ_0	0.001	0.051	0.051	0.000	0.051	0.052
δ_1	-0.001	0.067	0.063	0.003	0.066	0.065
δ_2	0.007	0.076	0.076	0.003	0.076	0.076
ρ_1	-0.003	0.217	0.222	0.024	0.186	0.186
ρ_2	-0.016	0.192	0.197	-0.013	0.154	0.156
J=3, ordered						
	$\rho_1 = \rho_2 = \rho_3 = 0$			$\rho_1 = \rho_2 = 0.5$, $\rho_3 = -0.5$		
δ_1	0.003	0.058	0.050	0.001	0.063	0.050
δ_2	0.008	0.071	0.058	0.008	0.079	0.058
ρ_1	-0.004	0.359	0.220	-0.018	0.361	0.189
ρ_2	-0.003	0.309	0.175	-0.018	0.284	0.160
ρ_3	0.002	0.327	0.250	0.016	0.278	0.206
J=3, unordered						
	$\rho_1 = \rho_2 = \rho_3 = 0$			$\rho_1 = \rho_2 = 0.5$, $\rho_3 = -0.5$		
δ_{02}	-0.005	0.079	0.076	-0.009	0.079	0.082
δ_{12}	0.004	0.093	0.090	0.009	0.093	0.095
δ_{22}	0.004	0.107	0.105	0.009	0.107	0.109
δ_{03}	-0.003	0.079	0.076	-0.010	0.079	0.079
δ_{13}	-0.002	0.093	0.088	-0.002	0.093	0.091
δ_{23}	-0.006	0.107	0.102	-0.010	0.107	0.109
ρ_1	-0.007	0.458	0.478	-0.077	0.417	0.443
ρ_2	-0.010	0.333	0.333	-0.016	0.296	0.307
ρ_3	0.000	0.296	0.316	0.006	0.257	0.273

Estimation was performed using joint PMLE estimator.

Table 5: Summary Statistics

Variable	married, has children	not married, has children	married, no children	not married, no children
Working (%)	77.59	79.01	95.51	96.33
Age	31.42	31.12	30.13	30.41
	(2.56)	(2.69)	(2.48)	(2.60)
Education	13.26	12.18	14.53	14.47
	(2.19)	(1.85)	(2.32)	(2.47)
Urban location (%)	71.54	71.26	76.97	84.16
AFQT score	0.02	-0.48	0.21	0.22
Ever stopped by police (%)	6.46	11.74	9.39	11.14
Self-esteem	0.04	-0.31	0.08	0.10
Ideal number of children	2.78	2.79	2.74	2.78
	(1.12)	(1.12)	(0.95)	(1.09)
Desired number of children	2.63	2.51	2.46	2.54
	(1.36)	(1.48)	(1.33)	(1.66)
Number of observations	5,189	1,482	1,246	1,660

Table 6: Estimated Coefficients and Standard Errors for Probability of Being Employed

	Probit	Best alt. PMLE	Joint PMLE
		Married women with children	
Age	-0.008	0.020	0.005
	(0.016)	(0.023)	(0.024)
Education	0.025	-0.004	0.013
	(0.019)	(0.027)	(0.027)
Urban	0.081	0.045	0.066
	(0.136)	(0.128)	(0.136)
AFQT	0.085**	0.114***	0.100**
	(0.043)	(0.043)	(0.047)
ρ		0.524	0.313
		Non-married women with children	
Age	-0.004	-0.014	-0.014
	(0.027)	(0.012)	(0.023)
Education	0.067*	0.138***	0.126***
	(0.038)	(0.018)	(0.041)
Urban	0.220	0.089	0.139
	(0.215)	(0.173)	(0.201)
AFQT	0.318***	0.318***	0.346***
	(0.082)	(0.036)	(0.076)
ρ		-0.998	-0.862
		Married women without children	
Age	-0.027	-0.053	-0.047
	(0.051)	(0.087)	(0.071)
Education	0.061	0.090	0.085
	(0.054)	(0.088)	(0.079)
Urban	0.681*	0.666*	0.676*
	(0.368)	(0.348)	(0.358)
AFQT	-0.051	-0.046	-0.046
	(0.121)	(0.119)	(0.119)
ρ		0.327	0.313
		Non-married women without children	
Age	-0.085*	-0.090***	-0.092***
	(0.050)	(0.025)	(0.035)
Education	-0.040	0.087***	0.077
	(0.067)	(0.033)	(0.055)
Urban	0.312	0.092	0.122
	(0.268)	(0.154)	(0.189)
AFQT	0.520***	0.191*	0.282*
	(0.179)	(0.107)	(0.160)
ρ		0.949	1.062

Table 7: Estimated Average Partial Effects in the Employment Equations

	Probit	Best alt. PMLE	Joint PMLE	Joint PMLE
	APE^U	APE^U	APE^U	APE^C
		Married women with children		
Age	-0.002	0.008	0.002	-0.001
Education	0.007	-0.002	0.004	0.001
Urban	0.024	0.017	0.022	0.019
AFQT	0.024	0.042	0.033	0.017
		Non-married women with children		
Age	-0.001	-0.001	-0.001	0.029
Education	0.018	0.006	0.008	-0.104
Urban	0.066	0.006	0.012	0.022
AFQT	0.069	0.012	0.017	-0.015
		Married women without children		
Age	-0.007	-0.013	-0.010	0.003
Education	0.016	0.020	0.016	0.139
Urban	0.198	0.180	0.160	0.404
AFQT	-0.013	-0.011	-0.009	0.046
		Non-married women without children		
Age	-0.023	-0.029	-0.033	-0.018
Education	-0.010	0.029	0.028	0.026
Urban	0.086	0.029	0.043	0.013
AFQT	0.114	0.065	0.104	0.022

Table 8: Estimated coefficients in sorting equations, joint PMLE

	Married, has children	Not married, has children	Married, no children
Age	0.121***	0.090***	-0.050**
	(0.021)	(0.025)	(0.023)
Education	-0.188***	-0.275***	0.014
	(0.025)	(0.031)	(0.025)
Urban location	-0.080	0.066	0.049
	(0.121)	(0.143)	(0.235)
AFQT score	0.061	-0.140**	-0.002
	(0.060)	(0.069)	(0.063)
Ever stopped by police	-0.561***	-0.196	-0.109
	(0.147)	(0.166)	(0.154)
Self-esteem	0.031	-0.092	0.016
	(0.051)	(0.057)	(0.054)
Ideal number of kids	-0.050	-0.031	-0.017
	(0.054)	(0.062)	(0.055)
Desired number of kids	0.107**	0.039	-0.008
	(0.048)	(0.058)	(0.047)