Estimation of Average Treatment Effects Using Panel Data when Treatment Effects Are Heterogeneous by Unobserved Fixed Effects

Shosei Sakaguchi *†
Graduate School of Economics, Kyoto University

April 3, 2017

Abstract

This paper proposes a new approach to identifying and estimating the time-varying average treatment effect (ATE) using panel data to control for unobserved fixed effects. The proposed approach allows for treatment effect heterogeneity induced by unobserved fixed effects. Under such heterogeneity, while existing panel data approaches identify and estimate the ATEs only for limited subpopulations, the proposed approach identifies and estimates the ATE for the entire population. The proposed approach requires two conditions: (i) The proportion of additive unobserved fixed effects terms in the treated and untreated potential outcome models is constant across units and time, and (ii) We have exogenous variables that correlate with unobserved fixed effects conditional on the assigned treatment. Under these conditions, the approach first identifies observed covariates parameters and the proportion of fixed effects terms. The approach then identifies the ATE by combining observed data with them to predict and adjust unobserved potential outcome for each treated and untreated unit. Based on the identification result, this paper proposes an estimator of the ATE, which takes the form of a generalized method of moments. I apply the estimator to estimate the impact of a mother smoking during pregnancy on her child's birth weight.

Keywords: Potential outcome, Program evaluation, Time-varying treatment, Treatment effect heterogeneity, Unobserved heterogeneity.

JEL classification: C21, C23.

---

*I am grateful to my supervisors Yoshihiko Nishiyama, Ryo Okui, and Ken Yamada for their helpful comments and encouragement. I would also like to thank Kotaro Hitomi, Naoya Sueishi, Susumu Imai, and seminar participants at Kyoto University, Hitotsubashi University, and the 2015 Japanese Economic Association Spring Meeting for their comments. This work was supported by JSPS KAKENHI Grant Number JP16J01170. Address for correspondence: Graduate School of Economics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan. Email: sakaguchi.shosei.75x@st.kyoto-u.ac.jp.

†Research Fellow of Japan Society of the Promotion of Science.

1 Introduction

This paper proposes a new approach to identifying and estimating the time-varying average treatment effect (ATE) using panel data to control for unobserved time-invariant confounders (i.e., fixed effects). The proposed approach can identify and consistently estimate the ATE for the entire population, rather than for a limited subpopulation, even when treatment effects are heterogeneous by unobserved unit characteristics.

Evaluating policy effects based on non-randomized studies often assumes that treatment assignments are independent of potential outcomes conditional on observed covariates.[1] However, this frequently used assumption is often violated due to the presence of unobserved confounding variables. For example, in job training program evaluations, such unobserved confounding variables may include individual ability. In the presence of unobserved confounding variables, econometric methods based on this assumption fail to provide consistent estimations for the interested parameters.

Panel data, the multi-period observation of each unit, is often used to solve the issue of unobserved confounding variables. In the panel data literature, the time-invariant unobserved confounding variables are refereed to as unobserved fixed effects, and panel data enables us to control for them. For example, in job training program evaluations, unobserved fixed effects may include individual ability. Popularly used methods in panel data treatment effect analysis are a fixed-effects (FE) estimation and a difference-in-differences (DID) estimation (e.g., Abrevaya, 2006; Card, 1996; LaLonde, 1986), which can control for additive unobserved fixed effects. Even when we cannot observe some additive time-invariant confounding variables, these estimation methods enable us to consistently estimate the ATE by exploiting panel data in some situations.

Although the FE estimation and the DID estimation are frequently used to control for unobserved fixed effects, they fail to consistently estimate the ATE in some situations. The FE estimation cannot consistently estimate the ATE when treatment effects are heterogeneous by unobserved fixed effects (e.g., Angrist and Pischke, 2008, Chapter 5; Lechner, 2011). The FE estimation requires the homogeneity of treatment effects among unob-

[1] Imbens (2004) and Imbens and Wooldridge (2009) summarize econometric methods based on this conditional independence assumption.

served fixed effects, but this is a very restrictive assumption because there are many cases where treatment effects are heterogeneous by unobserved fixed effects. For example, in job training evaluations, the training effects on wages may be heterogeneous by individual's unobserved ability, which can be considered as unobserved fixed effects. Gibbons, Suárez Serrato, and Urbancic (2014) present a number of empirical examples where the FE estimations are biased due to the presence of treatment effect heterogeneity by unobserved fixed effects. On the other hand, the standard DID method is applicable regardless whether treatment effects are heterogeneous by unobserved fixed effects or not, however, it requires the availability of the pretreatment data and identifies the ATE for the treated (ATT), rather than for the entire population, when treatment effects are heterogeneous by unobserved fixed effects.[2] In several program evaluation studies, the ATT does not coincide with the parameter of interest while the ATE does (e.g., Björklund and Moffitt, 1987).

The proposed approach in this paper can identify and consistently estimate the ATE for the entire population, rather than for a limited subpopulation, even when treatment effects are heterogeneous by unobserved fixed effects. In the presence of the heterogeneity, while the DID approach identifies and estimates only the ATT, the proposed approach identifies and estimates the ATE. Björklund and Moffitt (1987) stressed that, in many cases, the treatment effect for the treated does not coincide with the treatment effect for the population of the program evaluation interest. Note also that, while the DID approach requires the availability of pretreatment data, the proposed approach does not require such data restriction. Further, the proposed approach allows the ATE to vary across time periods and can identify and estimate the ATE at each time period.

To identify the ATE in the presence of the heterogeneity, the proposed identification approach relies on two conditions: (i) The proportion of additive unobserved fixed effects terms in the treated and untreated potential outcomes is constant across both units and time periods, and (ii) We have exogenous variables that correlate with unobserved fixed effects conditional on the assigned treatment. Condition (i) holds when unobserved fixed effects are represented as single unobserved confounding factor with a multiplicative coefficient. In the potential outcome models considered in this paper, constant terms are

[2]For proof of this statement, see, for example, Lechner (2011).

separated from unobserved fixed effects terms, thereby they are not involved with the constant proportion restriction of unobserved fixed effects terms. Condition (ii) holds even when we do not have exogenous variables that originally correlate with the unobserved fixed effects but have exogenous variables that correlate with them after conditioning on the assigned treatment.

Under these conditions, the proposed approach identifies the ATE through predicting and adjusting an unobserved potential outcome for each treated and untreated unit. To predict and adjust the unobserved potential outcome for each unit, the approach first identifies observed covariate parameters in both the potential outcome models and the proportion of unobserved fixed effects terms, which is supposed to be constant under condition (i), by exploiting time variation of panel data. The observed covariate parameters are identified through the within transformation of each the potential outcome, which controls for unobserved fixed effects. The proportion of unobserved fixed effects terms is then identified from the remaining terms of potential outcome models, which are subtractions of the observed covariate terms form the potential outcomes. In this identification step, we focus on the subpopulation of units called *movers* in Chamberlain (1982), who experience both treatment and no-treatment across time periods, because the remaining terms for both the treated and untreated outcomes are obtained only for them. The variables that satisfy condition (ii) are also used in this identification step to identify the proportion of unobserved fixed effects from the remaining terms, which comprise not only unobserved fixed effects terms. The ATE is finally identified by combining observed data with the identified parameters and proportion of fixed effects to predict and adjust unobserved potential outcome for each unit. Building on the moment conditions derived from the identification result, this paper provides an estimator of the ATE, which takes the form of the generalized method of moments (GMM).

As related works, when treatment effects are heterogeneous by unobserved fixed effects, one way to identify and estimate the ATE is to identify and estimate the average partial effect (APE) of the unit-specific slope in the correlated random coefficient (CRC) panel data model using a generalized within-group approach (e.g., Chamberlain, 1992).[3] However-

[3]Wooldridge (2010, p. 968) illustrates that the APE of the unit specific-slope in the binary treatment CRC panel data model generally corresponds to the ATE.

ever, although the CRC panel data model captures the heterogeneity of treatment effects, the generalized within-group approach identifies and estimates the APE only for a subpopulation of movers. In many cases, the subpopulations of movers are small and do not correspond to the populations of program evaluation interests. In contrast, the approach proposed in this paper identifies and estimates the ATE for the entire population, rather than for the subpopulation of movers.

Wooldridge (2005), Arellano and Bonhomme (2012), and Graham and Powell (2012) also study the CRC panel data model. Wooldridge (2005) provides conditions under which the usual linear FE estimator is consistent for the APE despite the presence of correlated random coefficients. However, Chernozhukov et al. (2013, p. 546–547) illustrate the difficulty to justify one of the conditions under the discrete treatment model, which includes the model considered in this paper. Arellano and Bonhomme (2012) study the identification and estimation of higher-order moments and densities of the random coefficients. As they maintain Chamberlain's (1992) conditions, the identification and estimation are limited on the subpopulation of movers. Graham and Powell (2012) focus on continuous regressors and provide identification and estimation of the APE in the CRC panel data model under milder conditions than Chamberlain's (1992) conditions.

Among other related works, Chernozhukov et al. (2013) and Sakaguchi (2016) also study the identification and estimation of the ATE in panel data model. Chernozhukov et al. (2013) consider point and partial identification and estimation for nonseparable panel data models built on the time homogeneity condition of the period specific disturbance term. Under the same kind of time homogeneity condition, Jun, Lee, and Shin (2016) derive the sharp identifiable bounds of potential outcome distribution. While these identification and estimation results are based on the time homogeneity condition on the period specific disturbance term, the proposed approach in this paper does not require such time homogeneity condition. Sakaguchi (2016) proposes an approach to identify and estimate the ATE for the entire population as an extension of the DID approach. The approach requires uniquely structured panel data wherein the treatment exposure expands from no units to all units across time. The proposed approach in this paper does not require such restriction on the data structure.

The remaining of this paper is structured as follows. Section 2 describes the setting and considered potential outcome models. Section 3 outlines the identification approach. In this section, I describe the identification approach dividing three steps. Built on the identifiction result, Section 4 then provides the estimator as well as its large sample properties. Section 5 presents Monte Carlo simulation results to show the finite sample behavior of the proposed estimator. Section 6 provides an empirical example where I apply the proposed estimator to estimate the impact of mother smoking during pregnancy on her child's birth weight using data from Abrevaya (2006). I conclude this paper with some remarks in Section 7.

2 Setup and Model

We suppose that $\{Y_{it}, D_{it}, X_{it}, Z_{it}\}$ is observed for N units ($i = 1, 2, \ldots, N$) across T time periods ($t = 1, \ldots, T$), where $T \geq 2$. We assume N is large while T is small. Y_{it} denotes an observed outcome and X_{it} denotes a $K \times 1$ vector of observed time-varying covariates that may include time dummies, observed time-varying confounding variables, and interactions of them. Z_{it} is an $L \times 1$ vector of exogenous variables which can include some or all variables in X_{it}. Its use and required assumptions are discussed later. Each unit is grouped by $D_{it} \in \{0, 1\}$ such that $D_{it} = 1$ and $D_{it} = 0$ indicate treatment and no treatment for unit i at period t, respectively. We suppose that the treatment assignment is time-varying.

Following the potential outcome framework, we suppose that Y_{it} is expressed as

$$Y_{it} = D_{it}Y_{it}(1) + (1 - D_{it})Y_{it}(0) \tag{1}$$

where $Y_{it}(1)$ and $Y_{it}(0)$ denote two potential outcomes under treatment and no treatment for unit i at period t, respectively. We hence observe only one of the potential outcomes for each unit at each time period depending on the treatment assignment D_{it}.

Throughout this paper, we focus on the ATE for the entire population, rather than for a limited subpopulation, at each period t for $t = 1, \cdots, T$:

$$\tau_t^{ate} = E[Y_{it}(1) - Y_{it}(0)].$$

We shall allow the ATE to be time-variant and let our focus be on such time-variant effect.

Given the time-varying covariates X_{it} and the unobserved fixed effects C_i, we suppose that potential outcomes $Y_{it}(1)$ and $Y_{it}(0)$ are expressed as follow:

$$Y_{it}(1) = \alpha^1 + X_{it}'\beta^1 + \gamma^1 C_i + u_{it}^1 \quad \text{for } t = 1, \cdots, T \tag{2}$$

$$Y_{it}(0) = \alpha^0 + X_{it}'\beta^0 + \gamma^0 C_i + u_{it}^0 \quad \text{for } t = 1, \cdots, T \tag{3}$$

where u_{it}^1 and u_{it}^0 are defined as $u_{it}^1 = Y_{it}(1) - E[Y_{it}(1) \mid X_{it}, C_i]$ and $u_{it}^0 = Y_{it}(0) - E[Y_{it}(0) \mid X_{it}, C_i]$, respectively. In the models, we suppose that C_i is represented as scalar unobserved fixed effects with multiplicative coefficients (γ^1 and γ^0).[4] By this expression, the effects of C_i on the two potential outcomes are assumed to be proportionally constant, i.e., $(\gamma^1 C_i)/(\gamma^0 C_i) = \gamma^1/\gamma^0$ is constant (does not depend on i and t). γ^1 and γ^0 may have different values, which implies that unobserved fixed effects may have different influences on the two potential outcomes. Further, in the models, constant terms α^1 and α^0 are separated form unobserved fixed effects terms ($\gamma^1 C_i$ and $\gamma^0 C_i$). Thereby, the constant terms are not involved with the constant proportion restriction of unobserved fixed effects terms. The difference between β^1 and β^0 also allows for the possibility of different influences of the observed time-varying covariates on the two potential outcomes. We can also suppose nonlinear forms of X_{it}, instead of $X_{it}'\beta^1$ and $X_{it}'\beta^0$, in these models.

Under the potential outcome models (2) and (3), τ_t^{ate} is expressed as

$$\tau_t^{ate} = (\alpha^1 - \alpha^0) + E[X_{it}]'(\beta^1 - \beta^0) + (\gamma^1 - \gamma^0)E[C_i]. \tag{4}$$

As seen in Equation (4), the difference between γ^1 and γ^0 causes heterogeneity in treatment effects by unobserved fixed effects. The difference between β^1 and β^0 also causes heterogeneity in treatment effects by observed time-varying covariates. In addition, the time variation of X_{it} causes the ATE time variation.

To formalize the idea of confounding due to the presence of the time-varying observed covariates and unobserved fixed effects, we impose the following assumption.

[4] Throughout this paper, we suppose that $\gamma^1 \neq 0$ and $\gamma^0 \neq 0$.

Assumption 2.1.
$E[Y_{it}(j) \mid D_{it}, X_{i1}, \ldots, X_{iT}, C_i] = E[Y_{it}(j) \mid X_{it}, C_i]$ for all $j = 0, 1$ and $t = 1, \cdots, T$.

This assumption corresponds to the fundamental assumption for the FE estimation and has two meanings. One is that the treatment assignment and the potential outcomes are mean independent at each time period conditional on the observed covariates and unobserved fixed effects. This modifies the standard conditional mean independence assumption for the panel data setting by conditioning on the unobserved fixed effects. The standard conditional mean independence assumption is often controversial, as it assumes that beyond the observed covariates there are no unobserved unit characteristics associated with both the potential outcomes and the treatment assignment.[5] Thus, the conditional mean independence assumption with unobserved fixed effect is attractive to allow for the presence of such unobserved unit characteristics. The other meaning of Assumption 2.1 is that the potential outcomes at each period t do not depend on past and future observed covariates, which means strict exogeneity known in the panel data literature.

We also suppose the following assumption on the support of the distribution of the time-varying treatment assignment.

Assumption 2.2.
For any $j = 0, 1$ and $t = 1, \ldots, T$,
(i) $0 < Pr(D_{it} = 0) < 1$,
(ii) $P(D_{it} = j, 0 < \sum_{t=1}^{T} D_{it} < T) > 0$.

Assumption 2.2 (i) is the usual overlap condition, which is necessary for identification of the ATE at each time period. Assumption 2.2 (ii) implies that at each time period both treatment and no treatment groups include some units who experience both treatment and no treatment at least once across time periods (i.e., units with $\{0 < \sum_{t=1}^{T} D_{it} < T\}$). We call such units "*movers*", borrowing a terminology introduced by Chamberlain (1982). In other words, Assumption 2.2 (ii) guarantees the existence of movers in both the treatment and no treatment groups at each period t. Note that this assumption does not require all

[5] See, for example, Imbens and Wooldridge (2009, Section 5).

units to be movers but requires some units to be movers.

For the proposed identification, we suppose to have an $L \times 1$ vector of exogenous variables Z_{it} that correlate with the unobserved fixed effects conditional on the assigned treatment in the subpopulation of movers. Note again that Z_{it} can include some or all variables in X_{it}. Z_{it} satisfies the following assumption.

Assumption 2.3.
For any $j = 0, 1$, $s = 1, \cdots, T$, and $t = 1, \cdots, T$;
(i) $E[Z_{it} u_{is}^j \mid D_{it} = j, 0 < \sum_{t=1}^T D_{it} < T] = 0$,
(ii) $\text{rank}(\sum_{t=0}^T E[(1, Z_{it}')'(1, C_i) \mid D_{it} = j, 0 < \sum_{t=1}^T D_{it} < T]) = 2$.

There are four remarks about this assumption as follows. First, notice that Assumption 2.3 (i) requires Z_{it} to satisfy the mean strictly exogenous condition. Second, variables in X_{it} are likely to satisfy all the conditions in Assumption 2.3 even when they do not originally correlate with C_i, and, therefore, some or all elements in X_{it} might be included in Z_{it}. Since both variables in X_{it} and variables in C_i affect the treatment assignment, they are likely to correlate with each other conditional on the assigned treatment: which is the reason why variables in X_{it} are likely to satisfy condition (ii) in Assumption 2.3.[6] X_{it} also satisfies the strictly exogenous condition (i) in Assumption 2.3 under Assumption 2.1. Third, exogenous variables that originally correlate with C_i, of course, may be included in Z_{it}. In many empirical studies, there should be some variables that correlate with C_i, regardless of being conditional on D_{it}. For instance, if C_i represents individual unobserved ability, then education should correlate with C_i and, hence, may be included in Z_{it}. Fourth, although Z_{it} has the subscript t, time-invariant exogenous variables also can be included in Z_{it} for any $t = 1, \ldots, T$.

Throughout this paper, we suppose that all the defined random variables are independent and identically distributed (i.i.d.) across units as in the following assumption.

[6] To illustrate with an example, suppose now that D_{it} denotes job training participation, C_i is individual ability, and individual age is included in X_{it}. In this example, if individuals with low age and low ability tend to participate in the training, then, conditional on the training participation, the age correlates with the ability even if they have no correlation without conditioning on the training participation. In this example, the age may be included in Z_{it}.

Assumption 2.4.
$\{\{Y_{it}(0), Y_{it}(1), D_{it}, X_{it}, Z_{it}\}_{t=1}^{T}, C_i\}$ are i.i.d. across i.

This assumption imposes no restrictions for the distribution of the data across time periods unlike Chernozhukov et al. (2013) and Jun, Lee, and Shin (2016).

Remark (Relation to the CRC Panel Data Model)

The observation rule (1) transforms the potential outcome models (2) and (3) into the following observed outcome model:

$$\begin{aligned} Y_{it} &= D_{it}Y_{it}(1) + (1 - D_{it})Y_{it}(0) \\ &= X_{it}'\beta^0 + \tau(C_i)D_{it} + D_{it}X_{it}'(\beta^1 - \beta^0) + \gamma^0 C_i + u_{it} \end{aligned}$$

where $\tau(C_i) = (\gamma^1 - \gamma^0)C_i$ and $u_{it} = D_{it}u_{it}^1 + (1 - D_{it})u_{it}^0$. This model is regarded as the CRC panel data model where $\tau(C_i)$ is the correlated random coefficient. Given that the ATE is expressed as $\tau_t^{ate} = E[\tau(C_i)] + E[X_{it}]'(\beta^1 - \beta^0)$, identification of the APE of D_{it} (i.e., $E[\tau(C_i)]$) is required for identification of τ_t^{ate} as described by Wooldridge (2010, p. 968). However, the generalized within-group approach (e.g., Chamberlain, 1992) identifies the APE only for the subpopulation of movers. As a result, the approach identifies the ATE for the limited subpopulation, not for the entire population. In the following sections, I propose a new approach that identifies and consistently estimates the ATE for the entire population rather than for a limited subpopulation.

3 Identification Outline

This section illustrates the idea of identification of τ_t^{ate} based on the supposed potential outcome models and assumptions described in the previous section. I describe the identification dividing three sequential steps. Through the sequential steps, we consider identification of each parameter α^j and β^j ($j = 0, 1$) and the proportion of unobserved fixed effects terms in the potential outcome models (2) and (3) (γ^1/γ^0). Then, τ_t^{ate} is identified by combing observed data with them to predict and adjust the unobserved potential outcome for each treated and untreated unit. The following outlines the sequential procedure of the identification step-by-step. I formalize the identification result in Proposition 4.1 in the next section. In the remaining of this paper, without loss of generality, we suppose $\gamma^0 = 1$. Thereby, the proportion of unobserved fixed effects terms is represented as γ^1.

First step

In the first step, we consider identification of β^1 and β^0 based on the idea of the within transformation method used in the FE approach. Let denote $\bar{A}_i^j = (1/\sum_{t=0}^{T} 1\{D_{it} = j\}) \sum_{t=1}^{T} 1\{D_{it} = j\} A_{it}$ and $\ddot{A}_{it}^j = A_{it} - \bar{A}_t^j$ for any variable A_{it} and $j = 0, 1$, where $1\{\cdot\}$ is the indicator function. $\bar{A}_t^1$ is the mean of A_{it} for unit i across time periods when he or she is treated. Similarly, $\bar{A}_i^0$ is the mean of A_{it} for unit i across time periods when he or she is untreated. $\ddot{A}_{it}^j$ is difference between A_{it} and $\bar{A}_i^j$. For each $j = 0, 1$ and unit i with $\sum_{t=1}^{T} 1\{D_{it} = j\} = 0$ (i.e., the unit who is never treated or always treated during the observed period), define $\ddot{A}_{it}^j = 0$.[7] This transformation is considered to be a kind of the within transformation for each the treatment and no treatment group.

For the group of units with $D_{it} = j$ $(j = 0, 1)$, we consider the following transformed model

$$\ddot{Y}_{it}^j = \ddot{X}_{it}^{j\prime} \beta^0 + \ddot{u}_{it}^j \quad \text{for } t = 1, \ldots, T.$$

In this model, unobserved fixed effects do not appear since they are differenced out by the within transformation. Then, under some regularity conditions (described in Section 4), β^j $(j = 0, 1)$ is identified through this transformed model as follows:

$$\beta^j = E[\sum_{t=1}^{T} 1\{D_{it} = j\} \cdot \ddot{X}_{it}^j \ddot{X}_{it}^{j\prime}]^{-1} E[\sum_{t=1}^{T} 1\{D_{it} = j\} \cdot \ddot{X}_{it}^j \ddot{Y}_{it}^j]. \tag{5}$$

Second step

In the second step, we consider identification of the proportion of unobserved fixed effects terms γ^1 and transformed constant terms defined bellow. In this step, the identification is

[7]Note that for each $j = 0, 1$ and unit i with $\sum_{t=1}^{T} 1\{D_{it} = j\} = 1$ (i.e., the unit who is untreated or treated only once during the observed period), $\ddot{A}_{it}^j = 0$ also holds.

based on the subpopulation of movers (i.e., units with $\{0 < \sum_{t=1}^{T} D_{it} < T\}$). Given β^1 and β^0 identified in the previous step, on the subpopulation of movers, both of the remaining terms of potential outcome models, which are subtractions of observed covariate terms form potential outcomes, are obtained as follows;

$$\alpha^1 + \gamma^1 C_i + \bar{u}_i^1 = \bar{Y}_i^1 - \bar{X}_i^{1\prime}\beta^1 \text{ and } \alpha^0 + C_i + \bar{u}_i^0 = \bar{Y}_i^0 - \bar{X}_i^{0\prime}\beta^0.$$

The left hand sides of these equations are the sums of the unobserved fixed effects term and the constant term α^j with disturbance term $\bar{u}_i^j$ added ($j = 0, 1$). Note that, for units who are not movers, we obtain only one of the remaining terms ($\alpha^1 + \gamma^1 C_i + \bar{u}_i^1$ and $\alpha^0 + C_i + \bar{u}_i^0$) since such units experience only one of treatment and no treatment over the observed period. In other words, we obtain both of them only for movers. This is the reason why we focus on the subpopulation of movers in this step.

Then, at each period t, we divide the subpopulation of movers into the treatment and no treatment groups: $\{D_{it} = 1, \sum_{t=1}^{T} D_{it} \neq T\}$ and $\{D_{it} = 0, \sum_{t=1}^{T} D_{it} \neq 0\}$. For movers in the group of $\{D_{it} = 1, \sum_{t=1}^{T} D_{it} \neq T\}$, the model for $Y_{it}(1)$ can be re-expressed by replacing C_i with $\alpha^0 + C_i + \bar{u}_i^0$ in the model (2) as follows:

$$Y_{it}(1) = \tilde{\alpha}^1 + X_{it}'\beta^1 + \gamma^1(\alpha^0 + C_i + \bar{u}_i^0) + \tilde{u}_{it}^1 \tag{6}$$

where $\tilde{\alpha}^1 = \alpha^1 - \gamma^1\alpha^0$ and $\tilde{u}_{it}^1 = u_{it}^1 - \gamma^1\bar{u}_i^0$. Similarly, for movers in the group of $\{D_{it} = 0, \sum_{t=1}^{T} D_{it} \neq 0\}$, we obtain the following model for $Y_{it}(0)$ by replacing C_i with $\alpha^1 + \gamma^1 C_i + \bar{u}_i^1$ in the model (3):

$$Y_{it}(0) = \tilde{\alpha}^0 + X_{it}'\beta^0 + \frac{1}{\gamma^1}(\alpha^1 + \gamma^1 C_i + \bar{u}_i^1) + \tilde{u}_{it}^0 \tag{7}$$

where $\tilde{\alpha}^0 = \alpha^0 - \frac{1}{\gamma^1}\alpha^1$ and $\tilde{u}_{it}^0 = u_{it}^0 - \frac{1}{\gamma^1}\bar{u}_i^1$.

For identification of $E[Y_{it}(1)]$ and $E[Y_{it}(0)]$, we then consider to identify $\tilde{\alpha}^1, \tilde{\alpha}^0$, and γ^1 in models (6) and (7). However, due to the presence of disturbance terms ($\bar{u}_{i1}$ and $\bar{u}_{i0}$) added to $\alpha^1 + \gamma^1 C_i$ and $\alpha^0 + C_i$ like measurement errors, identification of the parameters $\tilde{\alpha}^1, \tilde{\alpha}^0$, and γ^1 is suffered from the endogeneity of $\alpha^1 + \gamma^1 C_i + \bar{u}_i^1$ and $\alpha^0 + C_i + \bar{u}_i^0$ without external variables. Recall that we have exogenous variables Z_{it} that satisfy the conditions in Assumptions 2.3, and we use them to solve the identification problem. Since under Assumption 2.3 exogenous variables Z_{it} correlate with C_i conditional on the assigned treatment in the subpopulation of movers, using Z_{it} enables us to identify $\tilde{\alpha}^1, \tilde{\alpha}^0$, and γ^1 built on the idea of the instrumental variables method to deal with measurement error.

Then, given that β^1 and β^0 are identified, the parameters $\tilde{\alpha}^1$, $\tilde{\alpha}^0$, and γ^1 are identified through the following moment conditions:

$$\sum_{t=1}^{T} E\left[(1, Z_{it}')'(Y_{it}(1) - \tilde{\alpha}^1 - X_{it}'\beta^1 - \gamma^1(\alpha^0 + C_i + \bar{u}_{it}^0)) \mid D_{it} = 1, \sum_{t=1}^{T} D_{it} \neq T\right] = 0, \quad (8)$$

$$\sum_{t=1}^{T} E\left[(1, Z_{it}')'(Y_{it}(0) - \tilde{\alpha}^0 - X_{it}'\beta^0 - \frac{1}{\gamma^1}(\alpha^1 + \gamma^1 C_i + \bar{u}_{it}^1)) \mid D_{it} = 0, \sum_{t=1}^{T} D_{it} \neq 0\right] = 0. \quad (9)$$

In these moment conditions, Z_{it} works as instrumental variables for $\alpha^0 + C_i + \bar{u}_{it}^0$ and $\alpha^1 + \gamma^1 C_i + \bar{u}_{it}^1$. Note that identification of $\tilde{\alpha}^1$ comes from the moment condition (8) for the subpopulation of $\{D_{it} = 1, \sum_{t=1}^{T} D_{it} \neq T\}$, identification of $\tilde{\alpha}^0$ comes from the moment condition (9) for the subpopulation of $\{D_{it} = 0, \sum_{t=1}^{T} D_{it} \neq 0\}$, and identification of γ^1 comes from both of the moment conditions for the subpopulation of movers.

Third step

Given all the parameters identified above, $E[Y_{it}(1)]$ and $E[Y_{it}(0)]$ are finally identified as follows:

$$E\left[Y_{it}(1)\right] = E\left[D_{it}Y_{it} + (1 - D_{it})\left\{\tilde{\alpha}^1 + X_{it}'\beta^1 + \gamma^1\left(\alpha^0 + C_i + \bar{u}_{it}^0\right)\right\}\right], \quad (10)$$

$$E\left[Y_{it}(0)\right] = E\left[(1 - D_{it})Y_{it} + D_{it}\left\{\tilde{\alpha}^0 + X_{it}'\beta^0 + \frac{1}{\gamma^1}\left(\alpha^1 + \gamma^1 C_i + \bar{u}_{it}^1\right)\right\}\right]. \quad (11)$$

In Equation (10), since we cannot observe $Y_{it}(1)$ for untreated units (i.e., units with $D_{it} = 0$), we adjust their unobserved potential outcomes $Y_{it}(1)$ by combining X_{it} and $\alpha^0 + C_i + \bar{u}_{it}^0$ with the identified parameters. This is built on the idea of the regression adjustment (see, for example, Wooldridge 2010, Section 21.3.2). In Equation (11), the same method is applied for treated units (i.e., units with $D_{it} = 1$) to adjust their unobserved potential outcomes $Y_{it}(0)$.

Finally, τ_t^{ate} is identified through the regression adjustment as follows

$$\begin{aligned}\tau_t^{ate} =&E[Y_{it}(1)] - E[Y_{it}(0)]\\ =&E[D_{it}\left(Y_{it} - \left\{\tilde{\alpha}^0 + X_{it}'\beta^0 + \frac{1}{\gamma^1}\left(\alpha^1 + \gamma^1 C_i + \bar{u}_{it}^1\right)\right\}\right)\\ &- (1 - D_{it})\left(\left\{\tilde{\alpha}^1 + X_{it}'\beta^1 + \gamma^1\left(\alpha^0 + C_i + \bar{u}_{it}^0\right)\right\} - Y_{it}\right)].\end{aligned} \tag{12}$$

Example (when $T = 2$)

I illustrate the sequential identification approach described above in the case of $T = 2$. In this case, the whole population is divided into the four types of subpopulations: $\{D_1 = D_2 = 1\}, \{D_1 = 1, D_2 = 0\}, \{D_1 = 0, D_2 = 1\}$, and $\{D_2 = D_2 = 0\}$. Table 1 illustrates the four divided subpopulations. The first and the fourth are subpopulations of units who are treated and not treated at any period, respectively. The second is the subpopulation of movers who are treated at period 1 but not treated at period 2. Similarly, the third is the subpopulation of movers who are not treated at period 1 but treated at period 2. The presence of units in the second and third subpopulations is guaranteed in Assumption 2.2, wheres that of the first and fourth subpopulations is guaranteed in Assumption 4.2 imposed in the next section.

Table 1: Four types of subpopulations in the case of $T = 2$

	$D_2 = 1$	$D_2 = 0$
$D_1 = 1$	$\{D_1 = D_2 = 1\}$	$\{D_1 = 1, D_2 = 0\}$
$D_1 = 0$	$\{D_1 = 0, D_2 = 1\}$	$\{D_2 = D_2 = 0\}$

Since units in the subpopulation $\{D_1 = D_2 = 1\}$ are treated at two periods, β^1 is identified from this subpopulation as Equation (5) as described in the first step of the sequential identification. By the same way, β^0 is identified as Equation (5) from the subpopulation $\{D_1 = D_2 = 0\}$.

Next, since units in the subpopulations $\{D_1 = 1, D_2 = 0\}$ and $\{D_1 = 0, D_2 = 1\}$ experience both treatment and no treatment across time periods, $\tilde{\alpha}^1, \tilde{\alpha}^0$, and γ^1 are identified from these subpopulations with Equations (8) and (9) as described in the second step of the sequential identification. Finally, τ_t^{ate} is identified through the regression adjustment (Equation (12)) as described in the third step of the sequential identification.

4 Estimation

This section proposes an estimator of the ATE at each period t building on the sequential identification procedure outlined in the previous section. The proposed estimator is the GMM estimator (Hansen, 1982) constructed from the stack of the moment conditions derived from the sequential identification.[8] Let $W_i = (\{Y_{it}, D_{it}, X_{it}', Z_{it}'\}_{t=0}^{T})'$, $\theta = (\tilde{\alpha}^1, \tilde{\alpha}^0, \beta^{1\prime}, \beta^{0\prime}, \gamma^1, \tau_1^{ate}, \cdots, \tau_T^{ate})'$ denotes the vector of the parameters, and $\Theta \in \mathbb{R}^{2K+3+T}$ denotes its parameter space. In this section, we denote the vector of the true parameters as $\theta_o = (\tilde{\alpha}_o^1, \tilde{\alpha}_o^0, \beta_o^{1\prime}, \beta_o^{0\prime}, \gamma_o^1, \tau_{1,o}^{ate}, \cdots, \tau_{T,o}^{ate})'$.

To build the GMM estimator, we make the following vectors of moment functions:

$$
\begin{aligned}
g_1(W_i, \theta) =& \sum_{t=1}^{T} D_{it} \cdot \ddot{X}_{it}^1 \left(\ddot{Y}_{it}^1 - \ddot{X}_{it}^{1\prime} \beta^1 \right) \\
g_2(W_i, \theta) =& \sum_{t=1}^{T} (1 - D_{it}) \cdot \ddot{X}_{it}^0 \left(\ddot{Y}_{it}^0 - \ddot{X}_{it}^{0\prime} \beta^0 \right) \\
g_3(W_i, \theta) =& \sum_{t=1}^{T} 1 \left\{ D_{it} = 1, \ \sum_{t=1}^{T} D_{it} \neq T \right\} \\
& \times (1, Z_{it}')' \left[Y_{it} - \tilde{\alpha}^1 - X_{it}'\beta^1 - \gamma^1 \left\{ \bar{Y}_i^0 - \bar{X}_i^{0\prime} \beta^0 \right\} \right] \\
g_4(W_i, \theta) =& \sum_{t=1}^{T} 1 \left\{ D_{it} = 0, \ \sum_{t=1}^{T} D_{it} \neq 0 \right\} \\
& \times (1, Z_{it}')' \left[Y_{it} - \tilde{\alpha}^0 - X_{it}'\beta^0 - \frac{1}{\gamma^1} \left\{ \bar{Y}_i^1 - \bar{X}_i^{1\prime} \beta^1 \right\} \right] \\
g_5(W_i, \theta) =& D_{i1} \left(Y_{i1} - \left\{ \tilde{\alpha}^0 + X_{i1}'\beta^0 + \frac{1}{\gamma^1} \left(\alpha^1 + \gamma^1 C_i + \bar{u}_{i1}^1 \right) \right\} \right) \\
& - (1 - D_{i1}) \left(\left\{ \tilde{\alpha}^1 + X_{i1}'\beta^1 + \gamma^1 \left(\alpha^0 + C_i + \bar{u}_{i1}^0 \right) \right\} - Y_{i1} \right) - \tau_1^{ate} \\
& \vdots \\
g_{4+T}(W_i, \theta) =& D_{iT} \left(Y_{iT} - \left\{ \tilde{\alpha}^0 + X_{iT}'\beta^0 + \frac{1}{\gamma^1} \left(\alpha^1 + \gamma^1 C_i + \bar{u}_{iT}^1 \right) \right\} \right)
\end{aligned}
$$

[8] The class of GMM estimators includes sequential estimators, where moment functions from the sequential steps can be stacked into one vector of moment conditions (Newey, 1984).

$$- (1 - D_{iT}) \left(\left\{ \tilde{\alpha}^1 + X_{iT}'\beta^1 + \gamma^1 \left(\alpha^0 + C_i + \bar{u}_{iT}^0 \right) \right\} - Y_{iT} \right) - \tau_T^{ate}.$$

$g_1(W_i, \theta)$ and $g_2(W_i, \theta)$ are the moment functions to identify β^1 and β^0, respectively, that are derived from the first step of the sequential identification procedure described in the previous section. $g_3(W_i, \theta)$ and $g_4(W_i, \theta)$ are the moment functions to identify $\tilde{\alpha}^1, \tilde{\alpha}^0$, and γ^1 that are derived from the second step of the sequential identification procedure. For each $t = 1, \ldots, T$, $g_{4+t}(W_i, \theta)$ is the moment function to identify τ_t^{ate} by the regression adjustment method that is derived from the third step of the sequential identification procedure. Define $g(W_i, \theta)$ to be the vector obtained by stacking all the above vector moment functions into one long vector.

To formalize the discussion of identification in the previous section and provide precise results for the GMM estimator, we require the following additional assumptions.

Assumption 4.1.
Θ is compact.

Assumption 4.2.
For any $j = 0, 1$, $\text{rank}(E[\sum_{t=0}^{T} 1\{D_{it} = j\} \cdot \ddot{X}_{it}^{j} \ddot{X}_{it}^{j\prime}]) = K$.

Assumption 4.3.
Each element of $g(W_i, \theta_o)$ has a finite second moment.

These assumptions are regularity conditions for identification of θ_o and consistency and asymptotic normality of its GMM estimator. Assumption 4.2 requires $E[\sum_{t=0}^{T} 1\{D_{it} = j\} \cdot \ddot{X}_{it}^{j} \ddot{X}_{it}^{j\prime}]$ to be full rank for any $j = 0, 1$. This assumption implicitly requires the presence of units who are treated for at least two periods and units who are not treated for at least two periods.[9] Notice further that, if X_{it} includes time period dummies, then treatment and no treatment groups at each corresponding period must have units who are not treated at the other period and units who are treated at the other period, respectively, to satisfy the condition in Assumption 4.2.

[9]The reason is the following. Since $\ddot{X}_{it}^1 = 0$ and $\ddot{X}_{it}^0 = 0$ hold for units who experience treatment less than twice and units who experience no-treatment less than twice, respectively, the rank condition in Assumption 4.2 is violated if no units experience treatment or no-treatment at least twice.

Given the vector of moment functions, the following proposition formalizes the sequential identification outlined in the previous section.

Proposition 4.1.
Suppose that Assumptions 2.1–4 and 4.2–3 hold for the models of Equations (2) and (3). Then, the only value of $\theta \in \Theta$ that satisfies $E[g(W_i, \theta)] = 0$ is $\theta = \theta_o$.

The proof of this proposition is in the Appendix, which formalizes the discussion in Section 3.

Given Proposition 4.1, the GMM estimation of the true parameter vector θ_o is straightforward. The GMM estimator is

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \sum_{i=1}^{N} g(W_i, \theta)' \Sigma_N^{-1} \sum_{i=1}^{N} g(W_i, \theta) \tag{13}$$

for some sequence of positive definite matrix Σ_N. The following proposition formulates consistency and asymptotic normality results for $\hat{\theta}$.

Proposition 4.2.
Suppose that Assumptions 2.1–4 and 4.1–3 hold for the models of Equations (2) and (3), $\Sigma_N \xrightarrow{p} \Sigma$, and Σ is positive semi-definite. Then, $\hat{\theta} \xrightarrow{p} \theta_o$ and $\sqrt{N}(\hat{\theta} - \theta_o) \xrightarrow{d} N(0, V_o)$ where $V_o = (G_o \Sigma G_o')^{-1} G_o \Sigma \Omega_o \Sigma G_o' (G_o \Sigma G_o')^{-1}$ with $G_o = E\left(\frac{\partial g(W_i, \theta_o)}{\partial \theta'}\right)$ and $\Omega_o = E[g(W_i, \theta_o) g(W_i, \theta_o)']$.

The proof is omitted since it is just application of the asymptotic theory for GMM estimator (see, for example, Theorems 2.6 and 3.4 in Newey and McFadden, 1994). V_o can be consistently estimated with its sample analogue. The asymptotic variance of $\sqrt{N}(\hat{\tau}_t^{ate} - \tau_{t,o}^{ate})$ is obtained as the $(2K+3+t)$-th diagonal of V_o for any $t = 1, \ldots, T$. If Σ_N is a consistent estimator of Ω_o, the resulting estimator is efficient GMM estimator with $V_o = (G_o \Omega_o G_o')^{-1}$. Alternative moment-based estimator with possibly better small-sample properties, such as generalized empirical likelihood, could be used instead of GMM estimator (see, for example, Newey and Smith, 2004).

5 Monte Carlo Simulation

In this section, we perform a simulation experiment to study the behavior of the proposed estimator (13). The data are derived from two kinds of data generating processes (DGPs), DGP1 and DGP2. Both these DGPs are based on binary treatment outcome models and treatment assignment model with $T = 2$. DGP1 and DGP2 are based on different outcome models as follows.

DGP1 for Y_{it} (outcome with same coefficients for unobserved fixed effects):

$$Y_{it} = \begin{cases} 2 + X_{it} + C_i + U_{it}^0 & \text{if} \quad D_{it} = 0 \\ 1 + 2X_{it} + C_i + U_{it}^1 & \text{if} \quad D_{it} = 1 \end{cases}$$

DGP2 for Y_{it} (outcome with different coefficients for unobserved fixed effects):

$$Y_{it} = \begin{cases} 2 + X_{it} + C_i + U_{it}^0 & \text{if} \quad D_{it} = 0 \\ 1 + 2X_{it} + 3C_i + U_{it}^1 & \text{if} \quad D_{it} = 1 \end{cases}$$

DGP1 and DGP2 are based on a same treatment assignment model as follows.

$$D_{i1} = \begin{cases} 0 & \text{if} \quad -1 + X_{i1} - C_i + U_{i1}^D \leqslant 0 \\ 1 & \text{otherwise} \end{cases}$$

$$D_{i2} = \begin{cases} 0 & \text{if} \quad -2 + X_{i2} - C_i + U_{i2}^D \leqslant 0 \\ 1 & \text{otherwise} \end{cases}$$

For both DGP1 and DGP2, observed covariates $(X_{i1}, X_{i2})'$ are drawn from a multivariate normal distribution with mean $(1, 2)'$, standard deviation 1, and pairwise covariance 0.3; unobserved fixed effects C_i is drawn from a normal distribution with mean 1 and standard deviation 1; the disturbance terms are drawn from multivariate normal distributions as follows:

$$(U_1^0, U_1^1, U_2^0, U_2^1)' \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.3 & 0.2 \\ & 1 & 0.2 & 0.3 \\ & & 1 & 0.5 \\ & & & 1 \end{pmatrix}\right),$$

$$\begin{pmatrix} U_1^D \\ U_2^D \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ & 1 \end{pmatrix}\right).$$

Besides the outcome model and the treatment assignment model, an exogenous variable Z_{it} is set as $Z_{it} = C_i + U_{it}^Z$, where $(U_{i1}^Z, U_{i2}^Z)'$ is drawn from a zero mean multivariate normal distribution with standard deviation 1, and pairwise covariance 0.3. Note that this exogenous variable satisfies the conditions in Assumption 2.3.

Several remarks should be noted about the DGPs. First, the difference between DGP1 and DGP2 is in the values of the coefficients for C_i in the outcome models. Between the treated and untreated potential outcomes, DGP1 has same values of the coefficients for C_i while DGP2 has different values of them. Thereby, the treatment effects are homogeneous among C_i under DGP1 while heterogeneous among C_i under DGP2. Second, the two variables Z_{it} and X_{it} can be used as instrumental variables for $\alpha^1 + \gamma^1 C_i + \bar{u}_i^1$ and $\alpha^0 + C_i + \bar{u}_i^0$ that satisfy the conditions in Assumption 2.3. Z_{it} originally correlates with C_i, whereas X_{it} does not originally correlate with C_i, but obtains the correlation by conditioning on the assigned treatment. Third, the threshold values in the DGP of treatment assignment are set to make the numbers of units assigned to the no treatment group larger than those assigned to the treatment group. That meets data characteristics of most empirical studies for program evaluations.

In this simulation, we compare the proposed estimator (13) with the OLS estimator and the FE estimator under DGP1 and DGP2. The focused parameters are ATEs at periods 1 and 2. The true values of the ATEs under DGP1 at periods 1 and 2 are 0 and 1, respectively; the true values of the ATEs under DGP2 at periods 1 and 2 are 2 and 3, respectively. For the proposed estimator, the simulation provides the results under three different sets of the instrumental variables: $\{Z_{it}\}$, $\{X_{it}\}$, and $\{Z_{it}, X_{it}\}$. In this simulation study, the proposed estimator uses the identity matrix as its weighting matrix Σ_N.

Table 2 reports the results of 10,000 simulations with sample sizes $N = 200, 500$, and 1000. Panels I and II in Table 2 report the simulation results for DGP1 and DGP2, respectively. Several findings are worth noting. First, the OLS estimator is severely biased under both DGP1 and DGP2 due to the presence of unobserved fixed effects. Second, the FE estimator is not biased under DGP1; however, it is biased under DGP2 due to the different values of coefficients for unobserved fixed effects. Third, the proposed estimator has little bias when Z_{it} is included in the set of the instrumental variables, whereas it has some bias when X_{it} is solely used as the instrumental variable under both DGP1 and DGP2. This bias is not so small in the case of small sample size. Fourth, with large sample sizes, even in the case that the FE estimator provides consistent estimation (DGP1), the performance of the proposed estimator is not very poor compared with that of the FE estimator. With large sample sizes, RMSEs are not very different between both the estimators when Z_{it} is included in the set of the instrumental variables. Fifth, comparing among the three different sets of the instrumental variables, the proposed estimator has the smallest standard deviations when both Z_{it} and X_{it} are used under DGP2 regardless sample sizes, whereas that is not true under DGP1 with sample sizes 500 and 1000. When X_{it} is solely used as the instrumental variable, the standard deviation of the proposed estimator is relatively large especially under small sample. From this finding, I suggest that, when the sample size is small, empirical researchers do not solely use the covariates already included in the outcome model in the set of the instrumental variables, but include the exogenous variables outside the outcome model in the set of the instrumental variables.

Table 2: Monte Carlo Simulation Results

	True Value	N=200 Bias	N=200 SD	N=200 RMSE	N=500 Bias	N=500 SD	N=500 RMSE	N=1000 Bias	N=1000 SD	N=1000 RMSE
Panel I: Simulation results for DGP1										
ATE for 1st period										
OLS	0	-1.241	0.247	1.265	-1.245	0.157	1.255	-1.245	0.110	1.250
FE	0	-0.001	0.250	0.250	0.000	0.158	0.158	-0.001	0.112	0.112
Proposed Estimator										
(1)	0	0.054	0.442	0.450	0.014	0.204	0.205	0.005	0.141	0.141
(2)	0	0.294	2.359	2.377	0.141	0.535	0.553	0.065	0.279	0.286
(3)	0	0.040	0.362	0.365	0.015	0.214	0.215	0.005	0.149	0.149
ATE for 2nd period										
OLS	1	-1.242	0.248	1.267	-1.244	0.155	1.254	-1.245	0.110	1.250
FE	1	-0.002	0.252	0.252	-0.002	0.160	0.160	-0.001	0.112	0.112
Proposed Estimator										
(1)	1	0.054	0.390	0.393	0.014	0.156	0.156	0.005	0.107	0.107
(2)	1	0.308	2.378	2.398	0.147	0.517	0.538	0.068	0.252	0.261
(3)	1	0.040	0.286	0.289	0.014	0.158	0.159	0.005	0.110	0.110
Panel II: Simulation results for DGP2										
ATE for 1st period										
OLS	2	-3.079	0.447	3.112	-3.084	0.286	3.097	-3.083	0.199	3.090
FE	2	-0.976	0.356	1.038	-0.976	0.227	1.002	-0.976	0.158	0.989
Proposed Estimator										
(1)	2	0.172	1.127	1.140	0.050	0.374	0.377	0.020	0.246	0.247
(2)	2	0.242	1.768	1.784	0.184	0.946	0.964	0.113	0.560	0.572
(3)	2	0.083	0.693	0.698	0.039	0.367	0.369	0.016	0.246	0.246
ATE for 2nd period										
OLS	3	-3.079	0.451	3.112	-3.082	0.284	3.095	-3.085	0.199	3.091
FE	3	-0.977	0.357	1.040	-0.978	0.226	1.003	-0.976	0.157	0.989
Proposed Estimator										
(1)	3	0.172	1.130	1.143	0.048	0.347	0.350	0.020	0.229	0.229
(2)	3	0.255	1.745	1.764	0.194	0.964	0.983	0.122	0.569	0.581
(3)	3	0.084	0.675	0.680	0.036	0.338	0.340	0.016	0.227	0.227

Note: True Value is the true value of the ATE. Bias, SD, and RMSE are the mean bias, standard deviation, and the root mean squared error of the estimates across the simulations, respectively. For the Proposed Estimator, the rows (1), (2), and (3) report the results of the proposed estimations using $\{Z_{it}\}$, $\{X_{it}\}$, and $\{Z_{it}, X_{it}\}$ as the set of instrumental variables, respectively.

6 Empirical Application

As an empirical application, I analyze the effect of mother smoking during pregnancy on her child's birth weight (e.g., Permutt and Hebel, 1989; Evans and Ringel, 1999; Abrevaya, 2006; Abrevaya and Dahl, 2008). Let D_{it} denote the mother smoking indicator where $D_{it} = 1$ indicates that the i-th mother was smoking during pregnancy for her t-th birth and $D_{it} = 0$ indicates no smoking during the pregnancy. The potential outcomes $Y_{it}(1)$ and $Y_{it}(0)$ indicate the child birth weight for the i-th mother's t-th birth if she was smoking and not smoking during the pregnancy, respectively.

The analysis uses the matched panel data set constructed by Abrevaya (2006) from the U.S. Natality Data Sets for 1990–1998. As the original data does not have unique identifiers for mothers, he carefully matches mothers to children, particularly focusing on pairs of the child's state of birth and the mother's state of birth that have a small number of observations, and constructs the matched panel data. I select the "matched panel #3" as it is most conservatively constructed. The same data set is also used by Arellano and Bonhomme (2012) in the CRC panel data model and by Jun, Lee, and Shin (2016) in the nonseparable panel data model. Since the numbers of births in the original data are different among mothers, I focus on mothers who had three children during the observed period. The final sample contains 12,360 mothers of whom 1,349 mothers smoked during their first pregnancy, 1,371 mothers smoked during their second pregnancy, and 1,437 mothers smoked during their third pregnancy.

Using this sample, I estimate the average effect of mother smoking during pregnancy on her child's birth weight at each birth time through the potential outcome models (2) and (3). In the models, X_{it} includes dummy variables indicating birth time, the gender of the child, the age of the mother at the time of birth, dummy variables indicating the existence of prenatal visits, and the "Kessner" index value for the quality of prenatal care (for details, see Abrevaya, 2006). Unobserved fixed effects C_i are supposed to represent mother's lifestyle factor (Jun, Lee, and Shin, 2016, p. 307). I impose a restriction that the time-varying covariates X_{it}, except for dummy variables indicating the number of birth time, have same values of their coefficients as Abrevaya (2006) and Arellano and Bonhomme (2012) do. Under this restriction, we do not consider the interactions between smoking and observed time-varying observed covariates X_{it} except for dummy variables indicating the number of birth time. As a set of instrumental variables Z_{it} that satisfies the conditions in Assumption 2.3, I use the set of the years of education of mother and the age of mother at the time of birth.[10] The years of education should originally correlate with unobserved fixed effects, whereas it is not clear whether the age originally correlates with them or not.

Table 3 presents estimates of common parameters in the transformed potential outcome

[10]Note that the age is included in X_{it} while the years of education is not included because its time variation is little.

models (6) and (7). In this estimation, I use the efficient two-step GMM estimator of (13) as the proposed estimator. All the coefficient estimates have the same signs as those in the results of Abrevaya (2006, Table IV) and Arellano and Bonhomme (2012, Table I). Except for the variable indicating no prenatal visit, all the remaining variables have significant coefficients estimates.

Table 3: Estimates of common parameters in models (6) and (7)

Variable	Estimate	SE
Male	139.48	4.43
Age	226.05	1.26
Age^2	-3.73	0.04
Kessner index = 2	-59.05	9.26
Kessner index = 3	-170.12	20.48
No prenatal visit	-36.68	36.88
First prenatal visit in 2nd trimester	73.55	10.85
First prenatal visit in 3rd trimester	179.66	25.78
Second child for smoker	-37.71	14.14
Second child for nonsmoker	46.71	5.25
Third child for smoker	-122.46	14.83
Third child for nonsmoker	62.24	7.82
γ^1/γ^0	0.61	0.06
$\tilde{\alpha}^1$	-55.35	14.98
$\tilde{\alpha}^0$	77.64	15.34

Note: Robust standard errors are presented in the column of SE.

Table 4 presents estimation results for the average effect of mother smoking using the proposed estimator, which is the efficient two-step GMM estimator of (13), and the FE estimator. The estimates with the proposed estimator show that the average effect of mother smoking increases with birth time while the estimates with the FE estimator do not. The proposed estimator provides higher estimates at the first birth and lower estimates at the second and third births than the FE estimator does. We might suspect that the FE estimator has downward bias for the average effect at the first birth and upward bias for the average effects at the second and third births due to ignoring the presence of heterogeneity in the effect of mother smoking by unobserved mother's characteristics.

Table 4: Estimates of average effect of mother smoking during pregnancy on her child's birth weight

	FE Estimator	Proposed Estimator
Average effect (first birth)	-123.28	-90.67
	(15.49)	(12.94)
Average effect (second birth)	-80.20	-174.75
	(17.07)	(14.85)
Average effect (third birth)	-134.32	-274.89
	(15.81)	(14.88)

Note: Robust standard errors are presented in the parentheses.

7 Conclusion

This paper proposed a new panel data method to identify and estimate time-varying ATE. The method can identify and estimate the ATE even when treatment effects are heterogeneous by unobserved fixed effects. Note again that, in this situation, the usual FE estimator fails to consistently estimate the ATE and the generalized within-group method for the CRC panel model identifies and estimates the ATE only for a limited subpopulation. In contrast, the proposed method can identify and consistently estimate the ATE for the entire population. Therefore, I recommend empirical researchers using panel data to apply the proposed estimator when treatment effects are considered to be heterogeneous by unobserved unit characteristics and their interests are in the ATE for the entire population.

It is also interesting to extend the proposed method to the dynamic treatment model proposed by, for examples, Robins (1986) and Lechner and Miquel (2010). The dynamic treatment model captures dynamic interactions between sequential treatment assignments and potential outcomes. Even though empirical studies for the dynamic treatment models use panel data, there are few studies for controlling for unobserved fixed effects by exploiting panel data. I leave this issue for future research.

Appendix

This section provides a proof of Proposition 4.1.

Proof of Proposition 4.1.

This proof formalizes the discussion in Section 3.

Under Assumptions 2.1, 2.2, 2.4, 4.2, and 4.3, for any $j = 0, 1$, $E[g_j(W_i, \theta)] = 0$ uniquely holds with $\beta_o^j = E[\sum_{t=1}^T I\{D_{it} = j\} \cdot \ddot{X}_{it}^j \ddot{X}_{it}^{j\prime}]^{-1} E[\sum_{t=1}^T I\{D_{it} = j\} \cdot \ddot{X}_{it}^j \ddot{Y}_{it}^j]$.

For $g_3(W_i, \theta)$ and $g_4(W_i, \theta)$ with $\beta^1 = \beta_o^1$ and $\beta^0 = \beta_o^0$, under Assumptions 2.1, 2.2, 2.4, and 4.2–3, their expectations become

$$
\begin{aligned}
&E[g_3(W_i, \theta) \mid \beta^0 = \beta_o^0, \beta^1 = \beta_o^1] \\
=&E[\sum_{t=1}^T 1\left\{D_{it} = 1,\ \sum_{t=1}^T D_{it} \neq T\right\} \cdot (1, Z_{it}')' \left(Y_{it} - \tilde{\alpha}^1 - X_{it}'\beta_o^1 - \gamma^1 \left\{\bar{Y}_i^0 - \bar{X}_i^{0\prime}\beta_o^1\right\}\right)] \\
=&E\left[\sum_{t=1}^T (1, Z_{it}')' \left(Y_{it} - \tilde{\alpha}^1 - X_{it}'\beta_o^1 - \gamma^1 \left\{\alpha^0 + C_i + \bar{u}_i^0\right\}\right) \mid D_{it} = 1, \sum_{t=1}^T D_{it} \neq T\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&E[g_4(W_i, \theta) \mid \beta^0 = \beta_o^0, \beta^1 = \beta_o^1] \\
=&E[\sum_{t=1}^T 1\left\{D_{it} = 0,\ \sum_{t=1}^T D_{it} \neq 0\right\} \cdot (1, Z_{it}')' \left(Y_{it} - \tilde{\alpha}^0 - X_{it}'\beta_o^0 - \frac{1}{\gamma^1} \left\{\bar{Y}_i^1 - \bar{X}_i^{1\prime}\beta_o^1\right\}\right)] \\
=&E\left[\sum_{t=1}^T (1, Z_{it}')' \left(Y_{it} - \tilde{\alpha}^0 - X_{it}'\beta_o^0 - \frac{1}{\gamma^1} \left\{\alpha^1 + \gamma^1 C_i + \bar{u}_i^1\right\}\right) \mid D_{it} = 0, \sum_{t=1}^T D_{it} \neq 0\right].
\end{aligned}
$$

Adding with conditions in Assumption 2.3 for Z_{it}, $E[g_3(W_i, \theta)] = 0$ uniquely holds with $(\tilde{\alpha}_o^1, \beta_o^1, \gamma_o^1)$. Similarly, under the same conditions, $E[g_4(W_i, \theta)] = 0$ uniquely holds with $(\tilde{\alpha}_o^0, \beta_o^0, \gamma_o^1)$.

Then, the following equations hold for each period $t = 1, \ldots, T$;

$$
\begin{aligned}
E[Y_{it}(1)] =&E\left[D_{it}Y_{it} + (1 - D_{it})\left\{\tilde{\alpha_o}^1 + X_{it}'\beta_o^1 + \gamma_o^1 \left(\bar{Y}_i^0 - \bar{X}_i^{0\prime}\beta_o^0\right)\right\}\right], \\
E[Y_{it}(0)] =&E\left[D_{it}\left\{\tilde{\alpha_o}^0 + X_{it}'\beta_o^0 + \frac{1}{\gamma_o^1} \left(\bar{Y}_i^1 - \bar{X}_i^{1\prime}\beta_o^1\right)\right\} + (1 - D_{it})Y_{it}\right].
\end{aligned}
$$

Thus, under Assumptions 2.1, 2.2, and 2.4, $E[g_{4+t}(W_i, \theta)] = 0$ uniquely holds with $\theta = \theta_o$ at each period t.

From all the above, under Assumptions 2.1–4 and 4.2–3, $E[g(W_i, \theta)] = 0$ is uniquely satisfied with $\theta = \theta_o$.

□

References

Abrevaya, J. (2006), "Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach," *Journal of Applied Econometrics*, 21, 489–519.

Abrevaya, J., and Dahl, C. M. (2008), "The Effects of Birth Inputs on Birthweight: Evidence From Quantile Estimation on Panel Data," *Journal of Business & Economic Statistics*, 26, 379–397.

Angrist, J. D., and Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press.

Arellano, M., and Bonhomme, S. (2012), "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *Review of Economic Studies*, 79, 987–1020.

Björklund, A., and Moffitt, R. (1987), "Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69, 42–49.

Card, D. (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64, 957–979.

Chamberlain, G. (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5–46.

——— (1992), "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60, 567–596.

Chernozhukov, V., Fernndez-Val, I., Hahn, J., and Newey, W. (2013), "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 535–580.

Evans, W. N., and Ringel, J. S. (1999), "Can Higher Cigarette Taxes Improve Birth Outcomes?," *Journal of Public Economics*, 72, 135–154.

Gibbons, C. E., Suárez Serrato, J. C., and Urbancic, M. B. (2014), "Broken or Fixed Effects?," Working Paper 20342, National Bureau of Economic Research.

Graham, B. S., and Powell, J. L. (2012), "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models," *Econometrica*, 80, 2105–2152.

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Development in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Jun, S. J., Lee, Y., and Shin, Y. (2016), "Treatment Effects with Unobserved Heterogeneity: A Set Identification Approach," *Journal of Business & Economic Statistics*, 34, 302–311.

LaLonde, R. J. (1986), "Evaluating the Econometric Evaluation of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

Lechner, M. (2011), "The Estimation of Causal Effects by Difference-in-Difference Methods," *Foundations and Trends in Econometrics*, 4, 165–224.

Lechner, M., and Miquel, R. (2010), "Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions," *Empirical Economics*, 39, 111–137.

Newey, W. K. (1984), "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 14, 201–206.

Newey, W. K., and McFadden, D. (1994), “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, vol. 4, eds. J. J. Heckman and E. E. Leamer, Amsterdam, The Netherlands: Elsevier, pp. 2111–2245.

Newey, W. K., and Smith, R. J. (2004), “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72, 219–255.

Permutt, T., and Hebel, J. R. (1989), “Simultaneous-Equation Estimation in a Clinical Trial of the Effect of Smoking on Birth Weight,” *Biometrics*, 45, 619–622.

Robins, J. (1986), “A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods: Application to Control of the Healthy Worker Survivor Effect,” *Mthematical Modelling*, 7, 1393–1512.

Sakaguchi, S. (2016), “Estimation of Time-varying Average Treatment Effects Using Panel Data when Unobserved Fixed Effects Affect Potential Outcomes Differently,” *Economics Letters*, 146, 82–84.

Wooldridge, J. M. (2005), “Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models,” *Review of Economics and Statistics*, 87, 385–390.

——— (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, Cambridge, MA: MIT Press.