ESSAYS ON PSEUDO PANEL DATA AND TREATMENT EFFECTS

By

Fei Jia

A DISSERTATION

Submitted
to Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2017

ABSTRACT

ESSAYS ON PSEUDO PANEL DATA AND TREATMENT EFFECTS

By

Fei Jia

This dissertation is composed of three chapters that study two suitable estimation methods for identifying causal relationships in the presence of (pseudo) panel data. The first and the second chapters are devoted to minimum distance estimation for pseudo panel models, whereas the third chapter is concerned with the estimation of controlled direct effects in causal mediation analyses using panel data.

The first chapter focuses on finite sample properties of minimum distance estimators in pseudo panel models. Previous research shows theoretically that the minimum distance asymptotic theory is a natural fit for pseudo panel models when cohort sizes are large. However, little is known about how minimum distance estimation performs with a realistic sample size. In a carefully designed simulation study that mimics the sampling scheme of repeated cross sections, we compare the optimal minimum distance estimator to the fixed effects estimator which is identical to the minimum distance estimators using identity weighting matrix. The results show that both estimators perform well in realistic finite sample setups. The results also confirm that the optimal minimum distance estimator is generally more efficient than the fixed effect estimator. In particular, we find that cohort-wise heteroskedasticity and varying cohort size are the two typical scenarios that call for the use of optimal weighting. For the fixed effects estimator, we find that the minimum distance inference is more suitable than the naive inference which incorrectly ignores the estimation errors in the pseudo panel of variable cohort means.

The second chapter extends the basic pseudo panel models in the first chapter by adding extra instrumental variables. The additional instruments, if non-redundant, can improve estimation efficiency. To have the efficiency gain result in a general form, we derive it in

a non-separable minimum distance framework developed in this chapter. Along with the efficiency gain result, consistency, asymptotic normality, and optimal weighting theorems are also established. This efficiency gain result echoes the property of generalized methods of moments that more moment conditions do not hurt. After developing the results in the non-separable minimum distance framework, we apply them to the extended pseudo panel models. we show that the minimum distance estimators in the extended pseudo panels are generalized least squares estimators, and the optimal weighting matrix is block diagonal. Because of the last fact, the use of optimal weighting becomes more important than in basic pseudo panels. Simulation evidence confirms the theoretical findings in realistic finite sample setups. For an empirical illustration, we apply the method to estimate returns to education using data from the Current Population Survey in the US.

The third chapter, coauthored with Zhehui Luo and Alla Sikorskii, proposes a flexible plug-in estimator for controlled direct effects in mediation analyses using the potential outcome framework. A controlled direct effect is the direct treatment effect on an outcome when the indirect treatment effect through a mediator is shut off by holding the mediator fixed. The flexible plug-in estimator for controlled direct effects is a parametric g-formula with an additional partially linear assumption on the outcome equation. Compared to simulation based method in the literature, this estimator avoids estimation of conditional densities and numerical evaluation of expectations. We compare the flexible plug-in estimator to the sequential g-formula estimator, and prove theoretically and via simulation that they are numerically equivalent under certain settings. We also discuss a sensitivity analysis to check the robustness of the flexible plug-in estimator to a particular violation of the sequential ignorability assumption. We illustrate the use of the flexible plug-in estimator in a secondary analysis of a random sample of low birthweight and normal birthweight infants to estimate the controlled direct effect of low birth weight on reading scores at age 17 when a behavior problem index is used as the mediator.

ACKNOWLEDGMENTS

The Ph.D. journey at Michigan State University has been a life-changing experience for me. Along this journey, I have been fortunate to meet so many amazing people who help me to gain little advantages every day. I thank Jeffrey Wooldridge who served as the chair of my dissertation committee. His advising throughout my studies at Michigan State University shaped two thirds of this dissertation and my current research. I thank Zhehui Luo who supervised my research assistantship in the Department of Epidemiology and Biostatistics at Michigan State University. The interdisciplinary collaboration with her and Alla Sikorskii makes the the third chapter of this dissertation possible. I also thank my other committee members, Peter Schmidt and Timothy Vogelsang for their help at various stages of my research. Their feedback had an important positive impact on the quality of my work. Finally, I thank Lori Jean Nichols and Margaret Lynch who work in the administrative staff of the department of economics. Their continuous support for five years helped a lot in completing this degree.

The work in the third chapter was supported by National Institute of Mental Health grant RC4MH092737 (Luo) and the data came from the grant R01MH44586 (Breslau).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

FINITE SAMPLE PROPERTIES OF THE MINIMUM DISTANCE ESTIMATOR FOR PSEUDO PANEL DATA

1.1 Introduction

Repeated cross-sectional data is available when a series of different random samples can be obtained from the population over time. The Current Population Survey in the U.S.A, conducted monthly, is an example of such type of data sets. By combining cross sections at consecutive points in time, repeated cross-sectional data gains the replicability over time in absence of genuine panel data. Although we still cannot track each individual over time, we are able to estimate certain panel data models, especially those with fixed individual-specific effects and those with individual dynamics, under appropriate conditions.

The literature that makes possible the estimation of these panel data models with only repeated cross sections dates back to the seminal work by Deaton (1985). Deaton's idea is to divide individuals into cohorts according to certain predetermined characteristics, such as year of birth, and then use the cohort means of all relevant variables to construct a panel at the cohort level. Since the variable cohort means are estimated rather than directly observed, such a constructed panel is often called a pseudo panel. Common panel data approaches such as first difference and fixed effects (FE) estimation are readily applicable because of this panel structure. In this chapter, our focus is on the pseudo panel FE estimator.

Despite the fact that the cohort means are error-ridden estimates, the pseudo panel FE coefficient estimator is generally consistent. The corresponding standard error estimators (the naive standard errors hereafter), however, are potentially problematic for ignoring the estimation errors in the cohort means, whether they are made robust to heteroskedasticity and/or serial correlation. To make the standard errors right, Imbens and Wooldridge (2007) propose a minimum distance (MD) approach for pseudo panel models. With asymptotics

relying on large cohort sizes, this approach is a natural fit for many microeconomic analyses, since for microeconomic data the cohort-wise number of observations is often large, and the number of cohorts and the number of time periods are often small. The MD approach effectively takes account of the estimated cohort means. More importantly, it provides an asymptotically efficient way to utilize all the moment conditions through its weighting procedure. In fact, Imbens and Wooldridge (2007) show that the pseudo panel fixed effect estimator is exactly the MD estimator that puts equal weights on the moment conditions via an identity weighting matrix.

The superiority of the MD approach for pseudo panels relies on large sample theory, but its finite sample properties have not been fully studied. It is possible that the naive FE standard errors, especially those made robust to heteroskedasticity and/or serial correlation, can still achieve acceptable accuracy under certain circumstances. Moreover, although the result on optimal weighting in Imbens and Wooldridge (2007) implies that departures from identity weighting call for optimal weighting, it is unclear what are the typical causes of those departures.

In this chapter, we investigate the finite sample properties of the MD approach for pseudo panels through a carefully designed simulation study. In particular, the attention is paid to the comparison of the optimal MD estimator and the MD estimator with identity weighting matrix. We identify two stylized causes, namely varying cohort sizes and cohort-wise error heteroskedasticity, of departures of the optimal weighting matrix from identity. In presence of these two features, optimal weighting evidently outperforms identity weighting. As for the naive FE inference, we find that it is always inferior to the MD FE inference. Therefore, we should never throw away individual-level data in empirical studies, for they contain useful information that the sample cohort means do not have.

The MD approach is certainly not the only approach to pseudo panels. Deaton (1985), for example, treats the estimated cohort means as a measurement error problem, and proposes a measurement-error corrected ordinary least squares (OLS) estimator. Collado (1997) extends

the analysis to dynamic models, and develops a measurement-error corrected GMM estimator based on the instrument variables (IV) method in Arellano and Bond (1991). Another strand of researches go beyond pseudo panels and dive into individual level. Moffitt (1993) considers both dynamic and binary choice models, and proposes an IV estimator that constructs IV from functions of cohort and/or time. In particular, Moffit points out that aggregating to the cohort level is equivalent to using a full set of cohort, time, and cohort-time dummies as IV. Girma (2000) quasi-differences pairs of individuals in the same cohort to circumvent the problem of missing individual trajectories, and proposes a particular GMM IV method that uses past and present values of the dependent and explanatory variables within the same group. Verbeek and Vella (2005) propose an alternative computationally attractive IV estimator. A more thorough review that also covers important empirical applications can be found in Verbeek (2008).

The rest of the chapter is organized as follows. In section 2 we set up the notations and framework. In section 3 we reports and discusses the results from the simulation study. Section 4 concludes.

1.2 Framework

Deaton (1985) shows the importance of distinguishing between the population model and the sampling scheme. This distinction, as pointed out by Imbens and Wooldridge (2007), "is critical for understanding the nature of the identification problem, and in deciding the appropriate asymptotic analysis". Therefore we follow this convention in this paper . The exposition in this section borrows heavily from Imbens and Wooldridge (2007).

1.2.1 The population models

Consider the population model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t + f_i + u_{it},\ t = 1, \ldots, T. \tag{1.1}$$

in which y_{it} is the dependent random variable, $\mathbf{x}_{it}$ is a $1 \times K$ vector of random covariates with the first entry a constant term, f_i is the unobserved time invariant effect , and u_{it} is the unobserved idiosyncratic error. $\boldsymbol{\beta}$ is the parameter of practical interest. η_t's are the time varying intercepts and are also treated as parameters to estimate since we are considering applications with small T. An alternative representation is to include time dummies in $\mathbf{x}_t$ and then the η_t's are obsorbed in $\boldsymbol{\beta}$. The index i refers to the same individual over time in the population model. Writing the subscript i explicitly helps to indicate whether the quantities are changing only across t, changing only across i, or changing across both, which will become useful later. The model (1.1) imposes the same data generating structure for all T time periods, which assumes a stationary population over time. Later we will see that, by stationary population, we essentially means that the population cohort means of f_i do not change over time.

Following Deaton (1985), we assume the population can be divided into G predetermined group. The group designation must be determined before the samples are drawn, and must be independent of time. Birth year, for example, is one of the most commonly used characteristic to define the group designation. Let g_i be the random variable indicating the group membership of a random draw i. g_i takes values in $\{1, 2, \dots G\}$. Take expectation of (1.1) conditional on group membership, we have

$$E(y_{it}|g_i = g) = E(\mathbf{x}_{it}|g_i = g)\boldsymbol{\beta} + \eta_t + E(f_i|g_i = g) + E(u_{it}|g_i = g),\ t = 1, \dots, T,\ g = 1, \dots, G. \tag{1.2}$$

Define the population cohort means as

$$\begin{aligned} \mu_{gt}^{y} &= E(y_{it}|g_i = g) \\ \boldsymbol{\mu}_{gt}^{\mathbf{x}} &= E(\mathbf{x}_{it}|g_i = g) \\ \alpha_g &= E(f_i|g_i = g) \\ \delta_{gt} &= E(u_{it}|g_i = g) \end{aligned} \tag{1.3}$$

for $g = 1, \dots, G$ and $t = 1, \dots, T$. Note that all the four quantities above are deterministic

population cohort means. Then we can rewrite (1.2) as

$$\mu_{gt}^{y} = \boldsymbol{\mu}_{gt}^{\mathbf{x}}\boldsymbol{\beta} + \eta_t + \alpha_g + \delta_{gt},\ g = 1, \ldots, G,\ t = 1, \ldots, T. \tag{1.4}$$

(1.2) and (1.4) are different notations for the population model at the cohort level. The parameter δ_{gt} can be considered as the effect of the cohort-time cell (g, t) net of the cohort effect α_g and the time effect η_t.

Even if μ_{gt}^{y} and $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ are known, the system of linear equations in (1.4) is not identified if we leave δ_{gt} vary freely. Therefore, we need certain restrictions on δ_{gt}. In a standard panel data model, a weak exogeneity assumption we usually make is the contemporaneous exogeneity of $\mathbf{x}_{it}$ given f_i:

$$E(u_{it}|\mathbf{x}_{it}, f_i) = 0,\ t = 1, \ldots, T. \tag{1.5}$$

This condition is, however, is not required here. A weaker condition that is relevant in the context of (1.1) is

$$E(u_{it}|f_i) = 0,\ t = 1, \ldots, T. \tag{1.6}$$

Note that by iterated expectation, (1.5) implies (1.6). This gives certain flexibility to pseudo panels on the exogeneity of $\mathbf{x}_{it}$, which will be discussed in more details later.

Because f_i summarized all time-invariant unobservables, Imbens and Wooldridge (2007) argue that (1.6) should be true for not only the lump sum f_i but also any time-invariant factors including g_i. In other words, f_i should represent any random variable that does not depend on time. While this thought experiment makes sense, rigorously speaking, it does impose stronger conditions than (1.6).[1] Nevertheless, we keep this treatment in this chapter. In particular, replacing f_i with the group indicator g_i, we obtain

$$E(u_{it}|g_i) = 0,\ t = 1, \ldots, T. \tag{1.7}$$

[1]The sigma algebra generated by f_i is not necessarily a subset of the sigma algebra generated by g_i

Note that $E(u_{it}|g_i)$ is still a random variable. Since g_i takes only finitely many values, an alternative way to write (1.7) is

$$\delta_{gt} = E(u_{it}|g_i = g) = 0,\ g = 1, \ldots, G,\ t = 1, \ldots, T. \tag{1.8}$$

Substitute (1.8) in (1.4), we get

$$\mu^y_{gt} = \boldsymbol{\mu}^{\mathbf{x}}_{gt}\boldsymbol{\beta} + \eta_t + \alpha_g,\ g = 1, \ldots, G,\ t = 1, \ldots, T. \tag{1.9}$$

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}')'$ be the $(K + T + G) \times 1$ column vector of parameters with $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)'$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_G)'$. There are, however, only $T + G - 2$ parameters to estimate. Since $\mathbf{x}_{it}$ includes a constant term, only $(G-1)$ parameters in α_g and $(T-1)$ in η_t are separately identifiable. We impose the normalization $\alpha_1 = 0$ and $\eta_1 = 0$ which is slightly different from the normalization $\sum_{g=1}^{G} \alpha_g = 0$ and $\eta_1 = 0$ in Imbens and Wooldridge (2007). With this treatment, $\alpha_g,\ g = 2, \cdots, G$ and $\eta_t,\ t = 2, \cdots, T$ represent the net effects relative to the first cohort at the first time period. If μ^y_{gt} and $\boldsymbol{\mu}^{\mathbf{x}}_{gt}$ are known, $GT \geq K + T + G - 2$, and the equations in (1.9) are linearly independent, then (1.9) contains enough (maybe over-identified) restrictions to solve for $\boldsymbol{\theta}$.

As pointed out in Imbens and Wooldridge (2007), what (1.7) really imposes is that the cohort-level equations contain only the set of cohort and time effects but not the cohort-time interaction effects. If for any cohort-time cell (g, t) δ_{gt} is nonzero, then there is a misspecification in the population model (1.1). In the extreme case where the true model contains a full set of cohort-time net effects, nothing is identified since the identification of any parameter comes from the variation of its associated variable over cohort and/or time.

Perhaps another representation helps understanding this better. Write the population model with a full set of cohort-time effects as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t + f_i + \delta_{g_i,t} + u_{it},\ t = 1, \ldots, T,$$

where $\delta_{g_i,t} = E(u_{it}|g_i)$, the cohort-time effect of cell (g_i, t), is properly treated as a random

variable. Then (1.7) is exactly.

$$\delta_{g_i,t} = 0,$$

i.e. the population model does not contain a full set of cohort-time effects.

Details about some common estimation strategies given (1.9), such as OLS, FE and FD can be found in Imbens and Wooldridge (2007). They are straightforward after treating the cohort means as known.

1.2.2 Discussion on exogeneity

We argue that a subtle flexibility is gained thanks to the fact that (1.6) is weaker than (1.5). Specifically, the weaker condition (1.6) allows the deviation of $\mathbf{x}_{it}$ from its cohort mean to be non-exogenous with respect to the deviation of u_{it} from its cohort mean. Put it differently, within a given cohort-time cell, $\mathbf{x}_{it}$ and u_{it} are allowed to be correlated. But at the cohort level, the cohort mean of $\mathbf{x}_{it}$ must be exogenous with respect to the cohort mean of u_{it}, if we treat the variation in their cohort means over cohort and time as the source of randomness. In sum, endogeneity at the individual level is allowed, but exogeneity at the cohort level is still required.

The first implication of this is that the allowed dependence between $\mathbf{x}_{it}$ and u_{it} is not arbitrary. $\mathbf{x}_{it}$ can still contain lagged dependent variables, most commonly $y_{i,t-1}$, or explanatory variables that are contemporaneously endogenous, but the dependence cannot be fundamental, meaning that it exists at the cohort level. In our setup, this is guaranteed by two restrictions: (i) the specification in the individual level population model (1.1) is correct, and (ii) the zero cohort mean of u_{it} condition in (1.7) holds. They together translate to the exclusion of a full set of cohort-time effects.

Another implication is that, if G is large enough so that we can rely on large G asymptotics,[2] we do not need the zero cohort mean of u_{it} condition imposed in (1.7) for consistent

[2]Alternatively, we can assume the conditional distribution of $\delta_{\mathfrak{g}t}$ given $\boldsymbol{\mu}^{\mathbf{x}}_{\mathfrak{g}t}$ is normal and use maximum likelihood estimation.

estimation of $\boldsymbol{\beta}$. The condition can be relaxed to some form of exogeneity at the cohort level. Let $\mathfrak{g} = g_i$ to simplify notation, and denote the cohort-level random explanatory variable and error by $\boldsymbol{\mu}^{\mathbf{x}}_{\mathfrak{g}t}$ and $\delta_{\mathfrak{g}t}$, which treats the cohort dimension as random but still leaves the time dimension fixed. Then one form of such exogeneity assumption can be expressed as

$$E(\delta_{\mathfrak{g}t}|\boldsymbol{\mu}^{\mathbf{x}}_{\mathfrak{g}t}) = 0,\ t = 1, \ldots, T. \tag{1.10}$$

Apparently, the condition (1.7) implies (1.10). The analysis in Deaton (1985) goes a bit further to treat the time dimension as random as well, and thus relies on large GT asymptotics, but the idea is essentially the same. Nevertheless, this treatment "seems unnatural for the way pseudo panels are constructed, and the thought experiment about how one might sample more and more groups is convoluted", as pointed out by Imbens and Wooldridge (2007). Therefore, if we do not have large G but only large cohort-time cell size, N_{gt}, MD estimation is the way to go, and we need to impose the stronger zero cohort mean of u_{it} condition (1.7).

The treatment regarding $\boldsymbol{\mu}^{\mathbf{x}}_{\mathfrak{g}t}$ and $\delta_{\mathfrak{g}t}$ above also breaks the barrier between the view of constructing cohort-level equations from the individual level, as represented by Deaton (1985) and Imbens and Wooldridge (2007), and the view of starting the analysis right from the cohort level. Both views make sense and are unified under this treatment. But when starting the analysis from the cohort level, we need to make sure that the assumptions are consistent with the process of construction from the individual level. In particular, attention should be paid to proper asymptotics.

1.2.3 Minimum distance estimation

Given a repeated cross-sectional data set with large cohort sizes, small number of cohorts and small number of time periods, the MD estimator is a natural fit. Because of the large cohort sizes, the cohort means μ^{y}_{gt} and $\boldsymbol{\mu}^{\mathbf{x}}_{gt}$ in (1.9) can be estimated fairly precisely by their sample analogs in each cohort-time cell . The system of equations (1.9) is the link between

the reduced-form parameter $\{(\mu_{gt}^y, \boldsymbol{\mu}_{gt}^{\mathbf{x}}),\ g = 1, \ldots, G,\ t = 1, \ldots, T\}$ and the structural parameter $\boldsymbol{\theta}$. The MD approach is essentially a delta method, recovering structural estimates from reduced-form estimates.

In the next several subsections, we derive the limiting distribution of the sample cohort means, present the minimization problem of MD estimators, and give a closed-form expression of the general MD estimator for pseudo panels. In particular, the optimal MD estimator and the FE estimator as the MD estimator with identity weighting are discussed in detail.

1.2.3.1 Limiting distribution of cohort-time cell means

Specifically, assume we have a random sample on $(\mathbf{x}_{it}, y_{it})$ of size n_t for each t, and we denote them collectively by$\{(x_{it}, y_{it}),\ i = 1, \ldots, n_t\}$. i may refer to different individuals in different time periods. This notation works fine as long as we keep in mind the in each time period we have a new random sample.

For each random draw i, let $\mathbf{r}_i = (r_{it,1}, r_{it,2}, \ldots, r_{it,G})$ be a vector of group indicators such that $r_{it,g} = 1_{\{g_i = g\}}$, where 1_A is the indicator function that takes values in $\{0, 1\}$ and equals 1 only if A is true. In this way we properly treat the group membership of the random draw i as a random vector $\mathbf{r}_i$. With $\mathbf{r}_i$, the sample average of the response variable in cohort-time cell (g, t) can be written as

$$\hat{\mu}_{gt}^y = n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} y_{it} = (n_{gt}/n_t)^{-1} n_t^{-1} \sum_{i=1}^{n_t} r_{it,g} y_{it} \tag{1.11}$$

where $n_{gt} = \sum_{i=1}^{n_t} r_{it,g}$ is properly treated as a random variable.

$\hat{\mu}_{gt}^y$ is generally consistent for μ_{gt}^y. Specifically, let $\rho_g = P(r_{it,g} = 1)$, the fraction of the population in cohort g. We have treated ρ_g as time invariant because we assume the population is stationary. Then

$$\hat{\rho}_{gt} = (n_{gt}/n_t) \xrightarrow{p} \rho_g, \tag{1.12}$$

and thus we have

$$\hat{\mu}_{gt}^{y} = \hat{\rho}_{gt}^{-1} n_t^{-1} \sum_{i=1}^{n_t} r_{it,g} y_{it} \xrightarrow{p} \rho_g^{-1} E(r_{it,g} y_{it}) = \mu_{gt}^{y}.$$

The last equality holds because $E(r_{it,g}y_{it}) = P(r_{it,g} = 1)E(y_{it}|r_{it,g} = 1) = \rho_g \mu_{gt}^{y}$. The same argument also holds for the other cohort means.

Let $\mathbf{s}_{it} = (y_{it}, \mathbf{x}_{it})$, and define $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}} = (\hat{\mu}_{gt}^{y}, \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}})$ as in (1.11). Then the asymptotic distribution of $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}}$ is

$$\sqrt{n_t}(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}\prime} - \boldsymbol{\mu}_{gt}^{\mathbf{s}\prime}) \longrightarrow Normal(\mathbf{0}, \rho_g^{-1} \boldsymbol{\Omega}_{gt}^{\mathbf{s}})$$

where

$$\boldsymbol{\Omega}_{gt}^{\mathbf{s}} = Var(\mathbf{s}_{it}|g)$$

is the $(K+1) \times (K+1)$ variance-covariance matrix for the cohort-time cell (g,t). When later we stack the means across groups and time periods, it is useful to have the result

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}\prime} - \boldsymbol{\mu}_{gt}^{\mathbf{s}\prime}) \longrightarrow Normal(\mathbf{0}, (\rho_g \kappa_t)^{-1} \boldsymbol{\Omega}_{gt}^{\mathbf{s}}) \tag{1.13}$$

where $n = \sum_{t=1}^{T} n_t$ and $\kappa_t = \lim_{n\to\infty}(n_t/n)$ is essentially the fraction of all observations accounted for by cross section t. $\rho_g \kappa_t$ is consistently estimated by n_{gt}/n. A consistent estimator for $\boldsymbol{\Omega}_{gt}^{\mathbf{s}}$ is

$$\hat{\boldsymbol{\Omega}}_{gt}^{\mathbf{s}} = n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} (\mathbf{s}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}})(\mathbf{s}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}})'. \tag{1.14}$$

which is the sample variance-covariance matrix of $\mathbf{s}$ within the cell (g,t).

Let $\boldsymbol{\pi} = (\boldsymbol{\mu}_{11}^{\mathbf{s}}, \boldsymbol{\mu}_{12}^{\mathbf{s}}, \ldots, \boldsymbol{\mu}_{1T}^{\mathbf{s}}, \boldsymbol{\mu}_{21}^{\mathbf{s}} \ldots, \boldsymbol{\mu}_{GT}^{\mathbf{s}})'$, the column vector of all cell means. $\boldsymbol{\pi}$ is a $GT(K+1)$ vector since each $\boldsymbol{\mu}_{gt}^{\mathbf{s}}$ is $K+1$. Define $\hat{\boldsymbol{\pi}}$ by replacing $\boldsymbol{\mu}_{gt}^{\mathbf{s}}$ with $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}}$. Now, $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}}$ are independent across g because we have random sampling for each t. When $\mathbf{x}_{it}$ does not contain lags or leads, $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{s}}$ are independent across t, too. Then, by stacking (1.13) for all (g,t), we have

$$\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \longrightarrow Normal(\mathbf{0}, \boldsymbol{\Omega}), \tag{1.15}$$

where $\mathbf{\Omega}$ is the $GT(K+1) \times GT(K+1)$ block diagonal matrix with the gt-th block $(\rho_g \kappa_t)^{-1} \mathbf{\Omega}^{\mathbf{s}}_{gt}$. Note that $\mathbf{\Omega}$ incorporates both different cell variance-covariance matrices as well as the different frequencies of observations. As we will see in the simulation study, this is exactly the reason why the optimal MD estimator outperforms other MD estimators when there are cohort-wise heteroskedasticity and varying cohort sizes.

1.2.3.2 Minimum distance approach for pseudo panels

Classical MD estimation is useful for obtaining structural estimates from reduced form estimates when a known relationship exists between the structural and reduced form parameters (see, e.g., Wooldridge (2010)). In the pseudo panel setup, the group means in $\boldsymbol{\pi}$ are the reduced form parameters, $\boldsymbol{\theta}$ contains the structural parameters, and the cohort-level equations embody the known relationship between $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.

To facilitate the discussion, we rearrange terms in (1.9) by putting everything on the left hand side of the equality sign. Write the resulting expression as

$$\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{0} \tag{1.16}$$

where $\mathbf{h}(\cdot, \cdot)$ is a $GT \times 1$ vector valued function (recall $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}')'$). The gt-th row of $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is $-\mu^{y}_{gt} + \boldsymbol{\mu}^{\mathbf{x}}_{gt} \boldsymbol{\beta} + \eta_t + \alpha_g$, or equivalently,

$$\boldsymbol{\pi}'_g (-1, \boldsymbol{\beta}')' + \eta_t + \alpha_g \tag{1.17}$$

where $\boldsymbol{\pi}_g$ is the g-th $T \times 1$ block of $\boldsymbol{\pi}$. The parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ do not appear in a separable way directly in $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$, but it can be shown that this is a separable case.

The classical MD estimator is a solution to the minimization problem

$$\min_{\boldsymbol{\theta} \in \mathbf{\Theta}} \; \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})' \, \mathbf{W} \, \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}). \tag{1.18}$$

where $\mathbf{\Theta}$ is the space of $\boldsymbol{\theta}$ and $\mathbf{W}$ is a $GT \times GT$ weighting matrix. $\mathbf{W}$ is needed when the restrictions in (1.16) over-identifies $\boldsymbol{\theta}$ ($GT > K + G + T - 2$). We focus on the over-identified

case because it is usually the case in practice. Chamberlain (Harvard lecture notes) shows that the optimal weighting matrix is the inverse of

$$\mathbf{M} = \nabla_{\pi}\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})\boldsymbol{\Omega}\nabla_{\pi}\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})', \tag{1.19}$$

where $\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is the $GT \times GT(K+1)$ Jacobian of $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\pi}$. Use Kronecker product (notation $\otimes$) and (1.17), we have

$$\nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{I}_{GT} \otimes (-1, \boldsymbol{\beta}')'$$

where $\mathbf{I}_{GT}$ is the $GT \times GT$ identity matrix. This last result is exciting because, with $\boldsymbol{\Omega}$ block diagonal, it implies that (1.19) is a $GT \times GT$ diagonal matrix with the gt-th diagonal entry

$$(\rho_g \kappa_t)^{-1}(-1, \boldsymbol{\beta}')\boldsymbol{\Omega}^{\mathbf{s}}_{gt}(-1, \boldsymbol{\beta}')'. \tag{1.20}$$

But recall that $\boldsymbol{\Omega}^{\mathbf{s}}_{gt} = Var(\mathbf{s}_{it}|g)$, we have

$$\tau^2_{gt} \equiv (-1, \boldsymbol{\beta}')\boldsymbol{\Omega}^{\mathbf{s}}_{gt}(-1, \boldsymbol{\beta}')' = Var(y_{it} - \mathbf{x}_{it}\boldsymbol{\beta}|g),$$

and therefore, a consistent estimator of τ^2_{gt} is

$$\check{\tau}^2_{gt} = n^{-1}_{gt}\sum_{i=1}^{N_t} r_{it,g}(y_{it} - \mathbf{x}_{it}\check{\boldsymbol{\beta}} - \check{\eta}_t - \check{\alpha}_g)^2$$

which is the sample residual variance within cell (g, t). Here $\check{\boldsymbol{\theta}}$ is the initial estimator of $\boldsymbol{\theta}$ obtained by putting $\mathbf{W} = \mathbf{I}_{GT}$. Note that $\check{\boldsymbol{\theta}} = \check{\boldsymbol{\theta}}(\hat{\boldsymbol{\pi}})$. $\check{\boldsymbol{\theta}}$ is exactly the least squares dummy variables (LSDV) estimator of $\boldsymbol{\theta}$ on the pseudo panel. Since $(\rho_g\kappa_t)^{-1}$ is consistently estimated by $(n_{gt}/n)^{-1}$, (1.20) can be consistently estimated by

$$(n_{gt}/n)^{-1}\check{\tau}^2_{gt}. \tag{1.21}$$

.

Denote by $\check{\mathbf{M}}^{-1}$ the estimated optimal weighting matrix, where the gt-th diagonal entry of $\check{\mathbf{M}}^{-1}$ is $(n_{gt}/n)/\check{\tau}^2_{gt}$. Note that $\check{\mathbf{M}} = \mathbf{M}(\check{\boldsymbol{\theta}}) = \mathbf{C}(\check{\boldsymbol{\theta}}(\hat{\boldsymbol{\pi}}))$ is a function of the reduced-form estimate $\hat{\boldsymbol{\pi}}$. Then the minimization problem of the optimal MD estimator is

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})'\,\check{\mathbf{M}}^{-1}\,\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}). \tag{1.22}$$

1.2.3.3 Closed-form MD estimators for pseudo panels

In the pseudo panel setup, $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ is linear in each argument, and the MD estimator of $\boldsymbol{\theta}$ is in closed form. We derive this expression in this section.

Let $\boldsymbol{\mu}_{gt}^{\underline{\mathbf{x}}} = (\boldsymbol{\mu}_{gt}^{\mathbf{x}}, \mathbf{d}_t, \mathbf{c}_g)$ be the $1 \times (K + G + T - 1)$ row vector of regressors, where $\mathbf{d}_t$ is a $1 \times (T-1)$ vector of time dummies and $\mathbf{c}_g$ is a $1 \times G$ vector of group dummies. Let

$$\boldsymbol{\mu}_{\underline{g}}^{\underline{\mathbf{x}}} = \begin{pmatrix} \boldsymbol{\mu}_{\underline{g1}}^{\underline{\mathbf{x}}} \\ \boldsymbol{\mu}_{\underline{g2}}^{\underline{\mathbf{x}}} \\ \vdots \\ \boldsymbol{\mu}_{\underline{gT}}^{\underline{\mathbf{x}}} \end{pmatrix}, \quad T \times (K + G + T - 1),$$

$$\boldsymbol{\mu}^{\underline{\mathbf{x}}} = \begin{pmatrix} \boldsymbol{\mu}_{\underline{1}}^{\underline{\mathbf{x}}} \\ \boldsymbol{\mu}_{\underline{2}}^{\underline{\mathbf{x}}} \\ \vdots \\ \boldsymbol{\mu}_{\underline{G}}^{\underline{\mathbf{x}}} \end{pmatrix}, \quad GT \times (K + G + T - 1).$$

Then $\nabla_{\boldsymbol{\theta}} \mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \boldsymbol{\mu}^{\underline{\mathbf{x}}}$, and the FOC for (1.22) is

$$\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime} \check{\mathbf{M}}^{-1} (\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}} \hat{\boldsymbol{\theta}} - \hat{\mu}^{y}) = \mathbf{0},$$

where $\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}} = (\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}}, \mathbf{d}_t, \mathbf{c}_g)$. Therefore, the optimal MD estimator is

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime} \check{\mathbf{M}}^{-1} \hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}})^{-1} \hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime} \check{\mathbf{M}}^{-1} \hat{\mu}^{y}. \tag{1.23}$$

which looks like a weighted least squares estimator. Following Chamberlain, the estimated asymptotic variance of $\hat{\boldsymbol{\theta}}$ is simply

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime} \check{\mathbf{M}}^{-1} \hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}})^{-1} / n. \tag{1.24}$$

Because $\check{\mathbf{M}}^{-1}$ is the diagonal matrix with entries $(n_{gt}/n)/\check{\tau}_{gt}^2$, it is easy to weight each cell (g, t) by $\sqrt{n_{gt}/n}/\hat{\tau}_{gt}$ and then compute both $\hat{\boldsymbol{\theta}}$ and its asymptotic standard errors via a weighted regression. In STATA, this can be done by specifying aweight $(n_{gt}/n)/\check{\tau}_{gt}^2$.

The FE estimator applied to the pseudo panel of cohort means turns out to be the MD estimator with identity weighting matrix. To see that, we simply replace $\check{\mathbf{M}}^{-1}$ with the identity matrix $\mathbf{I}_{GT}$ in (1.23),

$$\check{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}' \hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}})^{-1} \hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}' \hat{\mu}^{y} \tag{1.25}$$

Strictly speaking, (1.25) is the LSDV estimator on the pseudo panel. But since it gives the same estimates for $\boldsymbol{\beta}$, we also call it the FE estimator. The MD asymptotic variance estimator for $\check{\boldsymbol{\theta}}$ is

$$\widehat{Avar(\check{\boldsymbol{\theta}})} = (\hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}' \hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}})^{-1} (\hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}' \check{\mathbf{M}}^{-1} \hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}) (\hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}' \hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}})^{-1} / n. \tag{1.26}$$

Apparently, this formula is different from the naive FE asymptotic variance estimators to be discussed in the next section, whether they are made robust to heteroskedasticity and/or serial correlation.

1.2.3.4 Discussion on the difference between MD FE and naive FE inference

Unlike the optimal MD asymptotic variance, the MD asymptotic variance for $\check{\boldsymbol{\theta}}$ can not be estimated directly from a weighted regression. In fact, since the corresponding weighting matrix for FE is the identity matrix, the correct weight for each cell is simply no weight (equal weight). Without any weighting, a linear regression gives us the naive asymptotic variance estimator[3]

$$\widehat{Avar_{\mathtt{n}}(\check{\boldsymbol{\theta}})} = (\hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}' \hat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}})^{-1} \check{\sigma}^2 \tag{1.27}$$

where

$$\check{\sigma}^2 = (GT-1)^{-1} \sum_{g,t} \left(\hat{\mu}_{gt}^{y} - \hat{\boldsymbol{\mu}}_{gt}^{\mathbf{x}} \check{\boldsymbol{\beta}} - \check{\eta}_t - \check{\alpha}_g \right)^2.$$

Clearly, (1.26) and (1.27) coincide if $n\check{\mathbf{M}}^{-1}$ equals $\check{\sigma}^2 \mathbf{I}_{GT}$, which is generally not the case.

Making it robust to heteroskedasticity (White (1980)), we get the naive heteroskedasticity-robust asymptotic variance estimator

[3]Proper adjustment of degrees of freedom can also be proposed, which we do not discuss for simplicity. It also applies to the two robust naive variance estimators.

$$\widehat{Avar_{\mathtt{r}}(\check{\boldsymbol{\theta}})} = \left(\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right)^{-1}\left(\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}\left(diag(\hat{\boldsymbol{\mu}}^{\check{u}})\right)^2\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right)\left(\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right)^{-1}$$

where $diag(\hat{\boldsymbol{\mu}}^{\check{u}})$ is the diagonal matrix created by putting the vector $\hat{\boldsymbol{\mu}}^{\check{u}}$ on the principal diagonal. $\hat{\boldsymbol{\mu}}^{\check{u}}$ is the column vector that stacks all cohort-level residuals over g and t, and its $[(g-1)T+t]$-th entry is

$$\hat{\mu}^{\check{u}}_{gt} = n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g}\check{u}_{it}.$$

That is, $\hat{\mu}^{\check{u}}_{gt}$ is the sample cohort mean of the individual-level residuals within cell (g,t). The individual level residual, $\check{u}_{it}$, is defined as

$$\check{u}_{it} = y_{it} - \mathbf{x}_{it}\check{\boldsymbol{\beta}} - (\check{\eta}_t + \check{\alpha}_g).$$

Note that $(\hat{\mu}^{\check{u}}_{gt})^2$ is different from $\check{\tau}^2_{gt}$. The former is the square of the residual cohort mean for cell (g,t), which only contains cohort-level information, where as the latter is the sample variance of the residuals within cell (g,t), which contains individual-level information.

Further making it robust to heteroskedasticity and serial correlation (see, e.g. Wooldridge (2010)), we get the naive cluster-robust asymptotic variance estimator

$$\widehat{Avar_{\mathtt{c}}(\check{\boldsymbol{\theta}})} = \left(\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right)^{-1}\left(\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}diag_G(\hat{\boldsymbol{\mu}}^{\check{u}})diag_G(\hat{\boldsymbol{\mu}}^{\check{u}})'\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right)\left(\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}\prime}\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}}\right)^{-1}$$

where $diag_G(\hat{\boldsymbol{\mu}}^{\check{u}})$ is the block diagonal matrix with the g-th diagonal block $\hat{\boldsymbol{\mu}}^{\check{u}}_g$ for $g = 1,\cdots,G$. The subscript G in the notation $diag_G$ indicates that the block diagonal matrix has G blocks on the diagonal. $\hat{\boldsymbol{\mu}}^{\check{u}}_g$ is a $T\times 1$ vector with the t-th entry $\hat{\mu}^{\check{u}}_{gt}$. Alternatively, the middle term can be written as

$$diag_G(\hat{\boldsymbol{\mu}}^{\check{u}})diag_G(\hat{\boldsymbol{\mu}}^{\check{u}})' = diag(\hat{\boldsymbol{\mu}}^{\check{u}}_1\hat{\boldsymbol{\mu}}^{\check{u}\prime}_1, \hat{\boldsymbol{\mu}}^{\check{u}}_2\hat{\boldsymbol{\mu}}^{\check{u}\prime}_2, \cdots, \hat{\boldsymbol{\mu}}^{\check{u}}_G\hat{\boldsymbol{\mu}}^{\check{u}\prime}_G)$$

Unlike the diagonal matrix $n\check{\mathbf{M}}^{-1}$, this is a block diagonal matrix with $\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}}_g\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}\prime}_g$ on the g-th diagonal block.

To summarize, the three naive FE asymptotic variance estimators can be obtained by replacing $\check{\mathbf{V}}_0 = \check{\mathbf{M}}^{-1}/n$ in (1.26) with $\check{\mathbf{V}}_{\mathtt{n}} = \check{\sigma}^2\mathbf{I}_{GT}$, $\check{\mathbf{V}}_{\mathtt{r}} = \left(diag(\hat{\boldsymbol{\mu}}^{\check{u}})\right)^2$ and $\check{\mathbf{V}}_{\mathtt{c}} =$

$diag_G(\hat{\boldsymbol{\mu}}^{\check{u}})diag_G(\hat{\boldsymbol{\mu}}^{\check{u}})'$, respectively. But the naive FE inference is fundamentally different from the MD inference. The former only relies on cohort-level information, where as the latter abstracts information from the individual level. The robustness of $\widehat{Avar_r}(\check{\boldsymbol{\theta}})$ and $\widehat{Avar_{\mathsf{c}}}(\check{\boldsymbol{\theta}})$ is also with respect to cohort-level heteroskedasticity and/or serial correlation only, i.e. heteroskedasticity and/or serial correlation in μ^u_{gt}, which requires at least large G asymptotics. As illustrated by the simulation study in the next section, the naive FE inference is far less efficient since it discards all individual-level information.

1.3 Simulation and results

We now present the Monte Carlo simulation study that investigates the finite sample properties of the MD approach for pseudo panels. The simulation study focuses on two questions.

First, what are the typical scenarios in which the optimal MD estimator outperforms the FE estimator? From (1.21), we know that if there is cohort-wise heteroskedasticity and/or varying the cell sizes, the optimal MD estimator is expected to outperform the FE estimator. In general, if there is any pattern in the population model that makes the optimal weighting matrix evidently different from the identity matrix, the optimal MD estimator is supposed to perform better. We check if it is the case in the simulation study.

Secondly, can the naive FE inference still provide satisfactory accuracy and if it could, what are these typical scenarios? As discussed in the last section, the naive FE asymptotic variances and the MD FE asymptotic variance are alike in their formulae. On the other hand, the naive FE inference is fundamentally different form the MD FE inference in that the former discards all information at the individual level. The simulation study helps to understand these two seemingly conflicting facts.

As Imbens and Wooldridge (2007) point out, the simulation design should be careful in at least two places.

First, data for each cross section should be drawn from the population independently

across time, and the group identifier should also be randomly drawn. This is accomplished by a two step procedure. In the first step, we draw the population using (1.1). The population cohort sizes are fixed and depending on the design may or may not depend on cohort and/or time.[4] In the second step, we mimic the sampling scheme of repeated cross sections by drawing independent random samples over time. In each period, we draw a tiny portion of the population as the cross-sectional sample for that period.

Second, the underlying model should have full time effects to be realistic. If, as in Verbeek and Vella (2005), we omit the aggregate time effects while let explanatory variables to have means differ by cohort-time cell, the variation in $\mu_{gt}^{\mathbf{x}}$ will be relatively rich and thus we may set up too optimistic a situation for the estimators.

We consider five scenarios. The first is a benchmark scenario in which all things are balanced across cohort-time cells. In the remaining four scenarios, we manipulate four different features of the population model, namely the time effects, the covariate distribution, the cohort-wise heteroskedasticity and the varying cohort sizes, one at a time. In this way, it is easy to isolate the cause. We begin with the benchmark scenario.

1.3.1 Benchmark

In the benchmark scenario, we generate the outcome y_{it} as a linear function of the covariates $(x_{1it} = 1, x_{2it}, x_{3it}, x_{4it})$, the time effect η_t, the individual effect f_i and the idiosyncratic error u_{it} as in

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \eta_t + f_i + u_{it},\ i = 1, \cdots, N_t,\ t = 1, \cdots, T. \quad (1.28)$$

The parameter values used are $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) = (1, 1, 1, 1)$. The time effects are generated by $\eta_t = t - 1$, and the cohort effects are generated by $\alpha_g = g - 1$. Individual

[4] Ideally, we would like a population with infinity many observations so that it is infinitely close to the population distribution defined by (1.1). In reality this is impossible, so we draw a large number of individuals to approximate the population distribution.

fixed effects are generated by adding a random normal disturbance to the cohort effects, i.e. $f_i \sim N(\alpha_g, 1)$. The distribution of idiosyncratic error is given by $u_{it} \sim N(0, 10)$.

To fix ideas, it might be helpful to think of x_{2it}, x_{3it} and x_{4it} as education, experience and marital status, respectively. The outcome y_{it} is the log hourly wage, and there is an individual effect f_i representing some unobserved ability. The three explanatory variables x_{2it}, x_{3it} and x_{4it} are generated as follows

$$\begin{aligned} x_{2it} &\sim N(gt/6, 1), \\ x_{3it} &\sim N(\sin(gt), 1), \\ x_{4it} &\sim Bernouli\left(\frac{1}{1 + exp[1.5 * \sin(gt/2)]}\right). \end{aligned}$$

That is, x_{2it} is a continuous variable with population cohort mean $gt/6$ and with-cell variance 1. x_{3it} is a continuous variable with the population cohort mean $\sin(gt)$ and within-cell variance 1, and x_{4it} is a binary variable equal to 1 with probability $\frac{1}{1+exp(1.5*\sin(gt/2))}$. The key is to let the three variable cohort means have distinct variation over g and t.

We apply the optimal MD estimator and the MD estimator with identity weighting. The latter is numerically equivalent to the FE estimator on the pseudo panel of cohort means. For each estimator, we compute the MD coefficient and standard error estimates. For the MD estimator with identity weighting, we also compute the three naive FE standard errors discussed in the last section.

We consider a small panel with $G = 6$ cohorts and $T = 4$ time periods. The population cell sizes N_{gt} are 2×10^4, 10^5 and 5×10^5 respectively in the three cases considered. After the population panel is generated, we fix it over simulation replications. To mimic the sampling scheme of repeated cross-sectional surveys, we draw .2% of the population in each period. The resulting sample cell sizes n_{gt} are approximately 40, 200 and 1000, respectively. For each case, we consider three different numbers of replications. The results are reported in Table 1.1, Table 1.2 and Table 1.3.

There are several observations worth discussing. First of all, the optimal MD estimator

Table 1.1 Results for benchmark. $G = 6$, $T = 4$, $n_{gt} \approx 40$, sampling rate = .2%; R denotes number of replications; Monte Carlo averages on top, Monte Carlo standard deviations in parentheses.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
$\underline{R = 1000}$	x_2	0.977	0.353	0.345	0.313	0.322	0.979	0.347
		(0.345)	(0.041)	(0.080)	(0.097)	(0.142)	(0.343)	(0.041)
	x_3	0.999	0.199	0.194	0.185	0.201	1.001	0.196
		(0.191)	(0.017)	(0.042)	(0.047)	(0.080)	(0.195)	(0.016)
	x_4	1.001	0.634	0.619	0.584	0.657	0.998	0.622
		(0.631)	(0.063)	(0.139)	(0.163)	(0.249)	(0.633)	(0.061)
	c_2	0.980	0.433	0.424	0.400	0.198	0.977	0.425
		(0.420)	(0.032)	(0.090)	(0.122)	(0.092)	(0.423)	(0.031)
	d_2	1.025	0.412	0.404	0.387	0.469	1.023	0.405
		(0.417)	(0.036)	(0.088)	(0.108)	(0.174)	(0.420)	(0.035)
	$cons$	0.973	0.380	0.373	0.347	0.290	0.978	0.373
		(0.364)	(0.034)	(0.079)	(0.116)	(0.108)	(0.365)	(0.032)
$\underline{R = 5000}$	x_2	0.981	0.351	0.339	0.308	0.320	0.982	0.345
		(0.345)	(0.041)	(0.079)	(0.095)	(0.143)	(0.347)	(0.040)
	x_3	1.003	0.198	0.192	0.182	0.194	1.003	0.195
		(0.193)	(0.017)	(0.042)	(0.046)	(0.075)	(0.195)	(0.017)
	x_4	1.009	0.632	0.611	0.574	0.645	1.006	0.621
		(0.640)	(0.062)	(0.136)	(0.159)	(0.243)	(0.644)	(0.060)
	c_2	0.984	0.432	0.419	0.393	0.194	0.985	0.425
		(0.420)	(0.032)	(0.090)	(0.118)	(0.090)	(0.424)	(0.031)
	d_2	1.020	0.410	0.398	0.380	0.459	1.021	0.403
		(0.414)	(0.036)	(0.087)	(0.105)	(0.169)	(0.416)	(0.035)
	$cons$	0.975	0.379	0.368	0.340	0.286	0.975	0.372
		(0.372)	(0.033)	(0.079)	(0.113)	(0.109)	(0.374)	(0.032)
$\underline{R = 10000}$	x_2	0.984	0.350	0.338	0.306	0.320	0.984	0.344
		(0.344)	(0.041)	(0.079)	(0.095)	(0.141)	(0.346)	(0.040)
	x_3	1.003	0.198	0.191	0.182	0.193	1.004	0.195
		(0.194)	(0.017)	(0.042)	(0.047)	(0.075)	(0.195)	(0.017)
	x_4	1.007	0.632	0.610	0.574	0.646	1.005	0.620
		(0.634)	(0.061)	(0.137)	(0.159)	(0.246)	(0.637)	(0.059)
	c_2	0.984	0.432	0.417	0.392	0.193	0.985	0.424
		(0.422)	(0.031)	(0.090)	(0.119)	(0.089)	(0.426)	(0.030)
	d_2	1.017	0.410	0.397	0.380	0.458	1.018	0.402
		(0.411)	(0.035)	(0.087)	(0.105)	(0.169)	(0.414)	(0.034)
	$cons$	0.979	0.378	0.367	0.339	0.286	0.978	0.372
		(0.373)	(0.032)	(0.079)	(0.113)	(0.108)	(0.375)	(0.031)

Table 1.2 Results for benchmark. $G = 6$, $T = 4$, $n_{gt} \approx 200$, sampling rate = .2%; R denotes number of replications; Monte Carlo averages on top, Monte Carlo standard deviations in parentheses.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
$\underline{R = 1000}$	x_2	0.990	0.161	0.157	0.140	0.143	0.990	0.160
		(0.165)	(0.009)	(0.034)	(0.042)	(0.062)	(0.165)	(0.009)
	x_3	1.008	0.089	0.087	0.082	0.087	1.008	0.089
		(0.088)	(0.003)	(0.018)	(0.021)	(0.033)	(0.087)	(0.003)
	x_4	1.010	0.286	0.279	0.262	0.299	1.011	0.285
		(0.289)	(0.012)	(0.059)	(0.072)	(0.110)	(0.288)	(0.012)
	c_2	0.996	0.191	0.186	0.175	0.088	0.995	0.190
		(0.194)	(0.006)	(0.039)	(0.052)	(0.036)	(0.194)	(0.006)
	d_2	1.000	0.184	0.180	0.171	0.206	1.000	0.183
		(0.181)	(0.007)	(0.038)	(0.047)	(0.075)	(0.182)	(0.007)
	$cons$	0.982	0.168	0.164	0.151	0.128	0.982	0.167
		(0.165)	(0.007)	(0.035)	(0.049)	(0.047)	(0.165)	(0.006)
$\underline{R = 5000}$	x_2	0.991	0.160	0.157	0.140	0.142	0.992	0.160
		(0.161)	(0.009)	(0.033)	(0.042)	(0.063)	(0.161)	(0.009)
	x_3	1.005	0.089	0.087	0.082	0.087	1.005	0.089
		(0.089)	(0.003)	(0.018)	(0.020)	(0.032)	(0.089)	(0.003)
	x_4	1.010	0.286	0.280	0.263	0.298	1.011	0.285
		(0.287)	(0.012)	(0.059)	(0.071)	(0.109)	(0.287)	(0.012)
	c_2	0.991	0.191	0.187	0.174	0.088	0.991	0.190
		(0.192)	(0.006)	(0.039)	(0.051)	(0.036)	(0.192)	(0.006)
	d_2	1.006	0.184	0.180	0.171	0.207	1.006	0.184
		(0.181)	(0.007)	(0.038)	(0.046)	(0.074)	(0.181)	(0.007)
	$cons$	0.983	0.168	0.164	0.150	0.127	0.983	0.167
		(0.168)	(0.006)	(0.034)	(0.049)	(0.046)	(0.169)	(0.006)
$\underline{R = 10000}$	x_2	0.995	0.161	0.157	0.140	0.143	0.995	0.160
		(0.161)	(0.009)	(0.034)	(0.042)	(0.063)	(0.161)	(0.009)
	x_3	1.004	0.089	0.087	0.082	0.088	1.004	0.089
		(0.089)	(0.003)	(0.018)	(0.020)	(0.032)	(0.089)	(0.003)
	x_4	1.012	0.286	0.279	0.263	0.298	1.013	0.285
		(0.286)	(0.012)	(0.059)	(0.071)	(0.110)	(0.286)	(0.012)
	c_2	0.990	0.191	0.186	0.174	0.088	0.990	0.190
		(0.191)	(0.006)	(0.039)	(0.052)	(0.036)	(0.191)	(0.006)
	d_2	1.002	0.184	0.180	0.171	0.206	1.002	0.183
		(0.183)	(0.007)	(0.038)	(0.046)	(0.074)	(0.183)	(0.007)
	$cons$	0.985	0.168	0.164	0.150	0.127	0.985	0.167
		(0.168)	(0.007)	(0.035)	(0.050)	(0.047)	(0.168)	(0.007)

Table 1.3 Results for benchmark. $G = 6$, $T = 4$, $n_{gt} \approx 1000$, sampling rate = .2%; R denotes number of replications; Monte Carlo averages on top, Monte Carlo standard deviations in parentheses.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
$\underline{R = 1000}$	x_2	1.003	0.072	0.070	0.062	0.062	1.003	0.072
		(0.072)	(0.002)	(0.014)	(0.018)	(0.028)	(0.072)	(0.002)
	x_3	0.997	0.040	0.038	0.037	0.039	0.997	0.040
		(0.040)	(0.001)	(0.008)	(0.009)	(0.014)	(0.040)	(0.001)
	x_4	1.001	0.128	0.124	0.117	0.132	1.001	0.128
		(0.130)	(0.002)	(0.026)	(0.031)	(0.049)	(0.130)	(0.002)
	c_2	1.004	0.085	0.082	0.076	0.039	1.004	0.085
		(0.085)	(0.001)	(0.017)	(0.022)	(0.015)	(0.085)	(0.001)
	d_2	1.004	0.082	0.079	0.075	0.090	1.004	0.082
		(0.083)	(0.001)	(0.016)	(0.019)	(0.031)	(0.083)	(0.001)
	$cons$	1.003	0.075	0.072	0.065	0.056	1.003	0.075
		(0.079)	(0.001)	(0.015)	(0.021)	(0.020)	(0.079)	(0.001)
$\underline{R = 5000}$	x_2	1.003	0.072	0.071	0.063	0.064	1.003	0.072
		(0.072)	(0.002)	(0.015)	(0.019)	(0.029)	(0.072)	(0.002)
	x_3	0.998	0.040	0.039	0.037	0.039	0.998	0.040
		(0.040)	(0.001)	(0.008)	(0.009)	(0.015)	(0.040)	(0.001)
	x_4	1.009	0.128	0.126	0.119	0.135	1.009	0.128
		(0.127)	(0.002)	(0.026)	(0.032)	(0.050)	(0.127)	(0.002)
	c_2	0.998	0.085	0.084	0.077	0.040	0.999	0.085
		(0.084)	(0.001)	(0.017)	(0.023)	(0.016)	(0.085)	(0.001)
	d_2	1.002	0.082	0.081	0.077	0.093	1.002	0.082
		(0.081)	(0.001)	(0.017)	(0.021)	(0.033)	(0.081)	(0.001)
	$cons$	1.003	0.075	0.073	0.067	0.057	1.002	0.075
		(0.076)	(0.001)	(0.015)	(0.022)	(0.021)	(0.076)	(0.001)
$\underline{R = 10000}$	x_2	0.999	0.072	0.071	0.063	0.064	0.999	0.072
		(0.073)	(0.002)	(0.015)	(0.019)	(0.028)	(0.073)	(0.002)
	x_3	0.997	0.040	0.039	0.037	0.039	0.997	0.040
		(0.040)	(0.001)	(0.008)	(0.009)	(0.015)	(0.040)	(0.001)
	x_4	1.009	0.128	0.125	0.118	0.134	1.009	0.128
		(0.127)	(0.002)	(0.026)	(0.031)	(0.049)	(0.127)	(0.002)
	c_2	1.002	0.085	0.083	0.077	0.039	1.002	0.085
		(0.084)	(0.001)	(0.017)	(0.023)	(0.015)	(0.084)	(0.001)
	d_2	1.001	0.082	0.080	0.076	0.092	1.001	0.082
		(0.082)	(0.001)	(0.017)	(0.020)	(0.033)	(0.082)	(0.001)
	$cons$	1.001	0.075	0.073	0.067	0.057	1.001	0.075
		(0.075)	(0.001)	(0.015)	(0.022)	(0.020)	(0.075)	(0.001)

has no advantage over the MD estimator with identity weighting in benchmark. This is under the current specification the optimal weighting matrix is an identity matrix, so the MD estimator with identity weighting is the optimal MD estimator.

Second, even in the case where the sample cohort size is about 40, the two MD estimators perform well. The Monte Carlo averages of the coefficient estimates are fairly close to the true parameter values. For each covariate, the Monte Carlo averages of the standard error estimates are also fairly close to the Monte Carlo standard deviations the coefficient estimates. Since the results are fairly stable across the three different numbers of replications, we will report the results for $10,000$ only in later discussions.

Third, the three naive FE standard errors are much more volatile than the MD FE standard error. This observation is consistent with the fact that the naive FE inference relies on cohort-level information and discards all individual-level information. Moreover, there seems to be downward small-sample bias in the naive FE standard errors than in the MD FE standard errors. This is mainly due to the small G setting. It is well known that $\check{\alpha}_g$'s are inconsistent under fixed G, which contaminate the residual estimates and in turn the naive FE standard errors. Another reason is that the degree-of-freedom adjustment used does not take into account the fact that the cohort means are estimated. Although the FE MD standard errors seem also biased downwards, the size of the biases is always smaller across the three tables. Clearly, in the benchmark scenarios the MD FE inference is superior to the naive FE inference.

Fourth, the cluster-robust naive FE standard errors are severely biased downwards for the cohort effect α_2. This observation remains valid for all the scenarios considered in this chapter. The explanation is again the fact that the estimates for the fixed effects obtained via LSDV are essentially based on only T observations, so the cluster-robust naive FE standard errors are inconsistent for fixed T.

Lastly, the performance of all estimators improve universally as the the cohort size increases. For the naive FE standard errors, the reason is that the sample cohort means of the

residuals approaches zero as cohort size increases. To keep the discussion concise, we will report the results for $n_{gt} \approx 200$ in later discussions.

1.3.2 Deterministic aggregate time effects

The aggregate time effects, i.e. the time intercepts η_t's, are treated as parameters in the population. Therefore, to generate the aggregate time effects properly, only deterministic functions of time need to be considered. If randomness is otherwise imposed on η_t, the random disturbance would become part of the idiosyncratic error, which is a separate scenario considered in section 1.3.4.

In the benchmark scenario, the aggregate time effects are $\eta_t = t - 1$ which is linear in t. In this section, we consider two additional deterministic functions of time: quadratic, and natural log

$$\eta_t = (t-1)^2,$$

$$\eta_t = ln(t).$$

The variation in the quadratic function is greater than that in the natural log function.

The results are reported in Table 1.4. The patterns of the results are similar to those in the last section. In fact, the two panels in Table 1.4 are exactly the same as the third panel in Table 1.2 except for the coefficient estimates on the time dummy d_2 in the lower panel where $\eta_t = ln(t)$. Note that the true coefficient on d_2 is $ln(2) \approx .693$ in that case. These results suggest that the aggregate time effect process has little effect on effect on the performance of the estimators. It only changes the true parameter values of the time effects. Of course, the fact the models are correct specified also plays a role. Correct specification implies that both estimators are consistent. As a result, changes in the deterministic process of the aggregate time effect have little effect on the estimated residuals and thus do not matter for inference or the estimation of other coefficients.

Table 1.4 Results for different aggregate time effect processes. $G = 6$, $T = 4$, $n_{gt} \approx 200$, sampling rate = .2%; 10,000 replications; Monte Carlo averages on top, Monte Carlo standard deviations in parentheses.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
$\underline{\eta_t = (t-1)^2}$	x_2	0.995	0.161	0.157	0.140	0.143	0.995	0.160
		(0.161)	(0.009)	(0.034)	(0.042)	(0.063)	(0.161)	(0.009)
	x_3	1.004	0.089	0.087	0.082	0.088	1.004	0.089
		(0.089)	(0.003)	(0.018)	(0.020)	(0.032)	(0.089)	(0.003)
	x_4	1.012	0.286	0.279	0.263	0.298	1.013	0.285
		(0.286)	(0.012)	(0.059)	(0.071)	(0.110)	(0.286)	(0.012)
	c_2	0.990	0.191	0.186	0.174	0.088	0.990	0.190
		(0.191)	(0.006)	(0.039)	(0.052)	(0.036)	(0.191)	(0.006)
	d_2	1.002	0.184	0.180	0.171	0.206	1.002	0.183
		(0.183)	(0.007)	(0.038)	(0.046)	(0.074)	(0.183)	(0.007)
	$cons$	0.985	0.168	0.164	0.150	0.127	0.985	0.167
		(0.168)	(0.007)	(0.035)	(0.050)	(0.047)	(0.168)	(0.007)
$\underline{\eta_t = ln(t)}$	x_2	0.995	0.161	0.157	0.140	0.143	0.995	0.160
		(0.161)	(0.009)	(0.034)	(0.042)	(0.063)	(0.161)	(0.009)
	x_3	1.004	0.089	0.087	0.082	0.088	1.004	0.089
		(0.089)	(0.003)	(0.018)	(0.020)	(0.032)	(0.089)	(0.003)
	x_4	1.012	0.286	0.279	0.263	0.298	1.013	0.285
		(0.286)	(0.012)	(0.059)	(0.071)	(0.110)	(0.286)	(0.012)
	c_2	0.990	0.191	0.186	0.174	0.088	0.990	0.190
		(0.191)	(0.006)	(0.039)	(0.052)	(0.036)	(0.191)	(0.006)
	d_2	0.695	0.184	0.180	0.171	0.206	0.695	0.183
		(0.183)	(0.007)	(0.038)	(0.046)	(0.074)	(0.183)	(0.007)
	$cons$	0.985	0.168	0.164	0.150	0.127	0.985	0.167
		(0.168)	(0.007)	(0.035)	(0.050)	(0.047)	(0.168)	(0.007)

1.3.3 Covariate distributions

To understand how the distributions of the covariates affect estimation, we manipulate the distribution of the covariates in this section. In particular, attention is paid to the covariate x_2.

In addition to the distribution $x_{2it} \sim N(gt/6, 1)$ considered in the benchmark, we look at the following two distributions

$$x_{2it} \sim N((gt)^2/6, 1),$$
$$x_{2it} \sim N(ln(gt)/6, 1).$$

The quadratic product of g and t embodies a greater variation than the product only, and the product in turn embodies a greater variation than its natural log transformation. All the other variables are generated as in the benchmark.

The results are summarized in Table 1.5. The pattern is similar to the last section. The only difference is that the greater variation in the cohort mean of x_{2it} in the first panel makes the estimation of its coefficient easier, whereas in the second panel the weaker variation renders estimation harder. Changes in the distribution of x_{2it} have little effect on the two MD estimators. The explanation is the same as that for the aggregate time effects. Since both estimators are consistent, the variation in the distribution of x_{2it} does not enter the residuals. The optimal weighting matrix is still an identity matrix, so the performance of the two MD estimators are similar.

1.3.4 Cohort-wise heteroskedasticity in the idiosyncratic error

In the benchmark, the idiosyncratic error u_{it} is homoskedastic. In this section, we investigate how cohort-wise heteroskedasticity in u_{it} would affect estimation. We present the results for two case in which the variance of u_{it} depends on (g, t). Specifically, we consider

Table 1.5 Results for distribution of x_2. $G = 6$, $T = 4$, $n_{gt} \approx 200$, sampling rate = .2%; 10,000 replications; Monte Carlo averages on top, Monte Carlo standard deviations in parentheses.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
$\underline{x_2 \sim}$	x_2	1.000	0.004	0.004	0.004	0.004	1.000	0.004
		(0.004)	(0.000)	(0.001)	(0.001)	(0.002)	(0.004)	(0.000)
$\underline{N((gt)^2/6, 1)}$	x_3	1.004	0.086	0.084	0.080	0.083	1.004	0.086
		(0.087)	(0.003)	(0.018)	(0.020)	(0.032)	(0.087)	(0.003)
	x_4	1.011	0.286	0.279	0.263	0.299	1.012	0.285
		(0.286)	(0.012)	(0.059)	(0.071)	(0.111)	(0.286)	(0.012)
	c_2	0.989	0.176	0.172	0.167	0.058	0.988	0.175
		(0.176)	(0.004)	(0.036)	(0.050)	(0.023)	(0.176)	(0.004)
	d_2	1.000	0.157	0.153	0.152	0.189	1.000	0.157
		(0.155)	(0.004)	(0.032)	(0.039)	(0.065)	(0.156)	(0.004)
	$cons$	0.986	0.163	0.159	0.153	0.125	0.986	0.162
		(0.163)	(0.004)	(0.033)	(0.048)	(0.044)	(0.163)	(0.004)
$\underline{x_2 \sim}$	x_2	1.003	1.023	0.954	0.884	0.979	1.003	1.020
		(1.002)	(0.252)	(0.286)	(0.303)	(0.433)	(1.002)	(0.251)
$\underline{N(ln(gt)/6, 1)}$	x_3	1.003	0.090	0.084	0.080	0.086	1.003	0.090
		(0.086)	(0.009)	(0.018)	(0.021)	(0.033)	(0.086)	(0.009)
	x_4	1.014	0.306	0.287	0.268	0.301	1.015	0.305
		(0.293)	(0.031)	(0.063)	(0.072)	(0.112)	(0.294)	(0.031)
	c_2	0.987	0.224	0.210	0.202	0.130	0.987	0.223
		(0.216)	(0.040)	(0.055)	(0.065)	(0.073)	(0.216)	(0.040)
	d_2	0.998	0.201	0.188	0.182	0.213	0.998	0.201
		(0.195)	(0.038)	(0.051)	(0.056)	(0.085)	(0.195)	(0.038)
	$cons$	0.987	0.160	0.150	0.146	0.111	0.987	0.159
		(0.153)	(0.015)	(0.033)	(0.045)	(0.042)	(0.153)	(0.015)

$$u_{it} \sim N(0, 10 + (gt)^2),$$
$$u_{it} \sim N(0, 10 + gt).$$

The degree of heteroskedasticity is greater in the first case. All the other variables are generated in the same way as in the benchmark.

We note here that introducing variation in the distribution of $f_i = \alpha_g + \varepsilon_i^f$ is at most another way to introduce heteroskedasticity. First of all, it is not interesting to vary the deterministic process of the cohort effects α_g's, because they are parameters to estimate. Secondly, it is not interesting to vary the mean of the distribution of ε_i^f, because that would only affect the process of α_g's. Lastly, letting the variance of ε_i^f depend on g is the same as introducing cohort-wise heteroskedasticity in u_{it}. It does not make sense to let the variance of ε_i^f depend on t because f_i is time invariant.

The results in Table 1.6 show that cohort-wise heteroskedasticity has two major effects. First, the optimal MD estimator outperforms the MD estimator with identity weighting, especially in the top panel where $u_{it} \sim N(0, 10 + (gt)^2)$. This is because cohort-wise heteroskedasticity makes the optimal weighting matrix non-identity. Secondly, the strict increase in the variance of u_{it} in either case raises the standard errors of both MD estimators compared to the benchmark. This rise is universal.

1.3.5 Cohort-time cell size

In this section, we let the cohort-time cell size vary by cohort and time. Specifically, we manipulate the sampling rate so that the sample size for cohort g at time t follows approximately the following two processes

1. $n_{gt} \approx (200 + 180 \times 1.5) - 180|g - 3.5| = 470 - 180|g - 3.5|, \qquad g = 1, \ldots, G$
2. $n_{gt} \approx (200 + 50 \times 1.5) - 50|g - (3.5 - (t - 3))| = 275 - 50|g + t - 6.5|, \qquad g = 1, \ldots, G$

Table 1.6 Results for cohort-wise heteroskedasticity in error term. $G = 6$, $T = 4$, $n_{gt} \approx 200$, sampling rate = .2%; 10,000 replications; Monte Carlo averages on top, Monte Carlo standard deviations in parentheses.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
$\underline{u_{it} \sim}$	x_2	0.989	0.651	0.510	0.505	0.509	0.991	0.448
		(0.653)	(0.051)	(0.141)	(0.202)	(0.256)	(0.448)	(0.031)
$\underline{N(0, 10 + (gt)^2)}$	x_3	1.012	0.323	0.283	0.280	0.291	1.012	0.214
		(0.326)	(0.019)	(0.077)	(0.090)	(0.115)	(0.214)	(0.013)
	x_4	1.040	1.101	0.910	0.943	1.098	1.002	0.762
		(1.100)	(0.062)	(0.248)	(0.320)	(0.472)	(0.758)	(0.048)
	c_2	0.969	0.488	0.608	0.495	0.343	0.988	0.321
		(0.489)	(0.040)	(0.164)	(0.159)	(0.163)	(0.316)	(0.023)
	d_2	1.000	0.589	0.585	0.541	0.645	1.007	0.326
		(0.589)	(0.046)	(0.159)	(0.180)	(0.271)	(0.324)	(0.026)
	$cons$	0.976	0.348	0.534	0.416	0.356	0.979	0.265
		(0.349)	(0.030)	(0.144)	(0.149)	(0.138)	(0.263)	(0.016)
$\underline{u_{it} \sim}$	x_2	0.994	0.221	0.209	0.190	0.193	0.993	0.213
		(0.222)	(0.013)	(0.046)	(0.059)	(0.087)	(0.215)	(0.012)
$\underline{N(0, 10 + gt)}$	x_3	1.006	0.122	0.116	0.111	0.117	1.006	0.119
		(0.123)	(0.005)	(0.025)	(0.028)	(0.044)	(0.120)	(0.005)
	x_4	1.017	0.403	0.372	0.361	0.411	1.012	0.390
		(0.403)	(0.018)	(0.081)	(0.102)	(0.158)	(0.391)	(0.018)
	c_2	0.985	0.233	0.249	0.221	0.122	0.988	0.227
		(0.233)	(0.009)	(0.054)	(0.064)	(0.051)	(0.227)	(0.009)
	d_2	1.001	0.236	0.239	0.223	0.268	1.003	0.226
		(0.235)	(0.011)	(0.052)	(0.060)	(0.096)	(0.226)	(0.011)
	$cons$	0.983	0.194	0.218	0.188	0.161	0.983	0.191
		(0.195)	(0.008)	(0.047)	(0.060)	(0.058)	(0.192)	(0.008)

Table 1.7 Cohort-time cell sizes for the two sampling schemes

$n_{gt}:1$					$n_{gt}:2$				
	t					t			
g	1	2	3	4	g	1	2	3	4
1	20	20	20	20	1	285	217	149	81
2	200	200	200	200	2	285	285	217	149
3	380	380	380	380	3	217	285	285	217
4	380	380	380	380	4	149	217	285	285
5	200	200	200	200	5	81	149	217	285
6	20	20	20	20	6	13	81	149	217
col. sum	1200	1200	1200	1200	col. sum	1030	1234	1302	1234
total	4800				total	4800			

In the first case, the cohort size starts from 20 at cohort 1, increases linearly with step 180 up to 380 at cohorts 3 and 4, and then decreases with the same step down to 20 at cohort 6. The idea is to let the cohorts in the middle have more observations. The overall sample size is about 4800. The second case has approximately the same overall sample size, but the middle peak cohorts shifts over time. The highest sample cell size is 285, and the step is 68. The two schemes are shown in Table 1.7. Note that the changes in the two schemes are both quite radical.

The results are summarized in Table 1.8. The impact of varying cell size is similar to that of cohort-wise heteroskedasticity. The optimal MD estimator significantly outperforms the MD estimator with identity weighting in both cases. This is again due to the non-identity weighting matrix caused by the varying cell size.

The naive FE inference cannot provide satisfactory standard error estimates. Depend on the covariate and the robust type, it can either overestimate or underestimate, and the bias is overall large.

1.4 Conclude

Build upon the theoretical analysis in Imbens and Wooldridge (2007), we study the finite sample properties of the MD estimator for pseudo panels in this chapter. In particular, we

Table 1.8 Results for varying cohort size. $G = 6$, $T = 4$; n_{gt} follows the three specifications given in section 1.3.5 and is generated by varying the sampling rate; 10,000 replications; Monte Carlo averages on top, Monte Carlo standard deviations in parentheses.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
$\underline{n_{gt} : 1}$	x_2	0.993	0.444	0.298	0.328	0.285	1.008	0.282
		(0.443)	(0.067)	(0.097)	(0.127)	(0.157)	(0.289)	(0.027)
	x_3	1.004	0.155	0.166	0.148	0.162	1.002	0.095
		(0.155)	(0.017)	(0.049)	(0.045)	(0.068)	(0.095)	(0.006)
	x_4	1.005	0.480	0.529	0.457	0.538	1.020	0.308
		(0.482)	(0.062)	(0.156)	(0.149)	(0.226)	(0.311)	(0.020)
	c_2	0.993	0.472	0.362	0.383	0.177	0.984	0.412
		(0.468)	(0.063)	(0.107)	(0.146)	(0.100)	(0.436)	(0.039)
	d_2	1.005	0.455	0.347	0.370	0.424	0.987	0.263
		(0.453)	(0.063)	(0.105)	(0.140)	(0.207)	(0.266)	(0.021)
	$cons$	0.983	0.461	0.319	0.353	0.232	0.992	0.398
		(0.457)	(0.070)	(0.094)	(0.154)	(0.109)	(0.422)	(0.039)
$\underline{n_{gt} : 2}$	x_2	0.993	0.303	0.196	0.205	0.220	0.998	0.220
		(0.306)	(0.065)	(0.062)	(0.090)	(0.121)	(0.226)	(0.018)
	x_3	1.005	0.099	0.108	0.099	0.108	1.004	0.088
		(0.100)	(0.008)	(0.032)	(0.029)	(0.045)	(0.088)	(0.004)
	x_4	1.008	0.431	0.350	0.342	0.415	1.015	0.308
		(0.438)	(0.086)	(0.106)	(0.126)	(0.186)	(0.312)	(0.017)
	c_2	0.994	0.437	0.236	0.286	0.145	0.983	0.271
		(0.444)	(0.099)	(0.072)	(0.140)	(0.084)	(0.279)	(0.016)
	d_2	1.003	0.418	0.228	0.272	0.354	0.996	0.265
		(0.421)	(0.095)	(0.071)	(0.131)	(0.194)	(0.271)	(0.019)
	$cons$	0.983	0.466	0.209	0.276	0.205	0.992	0.289
		(0.473)	(0.118)	(0.065)	(0.158)	(0.109)	(0.298)	(0.021)

focus on the comparison of the optimal MD estimator and the MD estimator with identity weighting matrix. The latter is of interest because it coincides with the FE estimator applied to the pseudo panel of cohort means. We find that in cases where there is significant heteroskedasticity by cohort-time cells, or in cases where the cohort-time cell size varies, the optimal MD estimator significantly outperform the MD estimator with identity weighting in that the former's standard errors are smaller. This finding is consistent with the large cohort size asymptotics under the MD estimation framework, as the optimal MD estimator achieves the smallest asymptotic variance.

We also compare the MD FE inference to the naive FE inference. We find that in cases where the optimal weighting matrix is close to an identity matrix, the naive FE standard errors are barely satisfactory. But when the optimal weighting matrix is far from identity, the naive FE standard errors are not acceptable without doubt. In any case, the MD FE inference is always more efficient than the naive FE inference. This finding is consistent with the fact that the FE inference relies on large number of cohorts and it discards all individual-level information. In a setup with small number of cohorts and time periods, the naive FE inference cannot work well.

The simulation setup in this analysis considers sample cohort sizes in hundreds, and the results are already promising provided that the variation in covariate cohort means are rich enough. In practice, sample cohort sizes of repeated cross sections can easily exceed thousands. Therefore, the results in this chapter should bring confidence to the application of the MD approach to pseudo panels.

In future studies, we could extend the analysis to dynamic models where we have lagged dependent or explanatory variables. For robustness check, we should allow correlation covariates and individual fixed effects. Moreover, given the weak exogeneity condition (1.7), we could also allow covariates that are endogenous at the individual level but not at the cohort level. Results from these extensions can provide more practical implications.

BIBLIOGRAPHY

BIBLIOGRAPHY

Arellano, Manuel, and Stephen Bond. 1991. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations." *The review of economic studies*, 58(2): 277–297.

Collado, M Dolores. 1997. "Estimating dynamic models from time series of independent cross-sections." *Journal of Econometrics*, 82(1): 37–62.

Deaton, Angus. 1985. "Panel data from time series of cross-sections." *Journal of econometrics*, 30(1): 109–126.

Girma, Sourafel. 2000. "A quasi-differencing approach to dynamic modelling from a time series of independent cross-sections." *Journal of Econometrics*, 98(2): 365–383.

Imbens, Guido, and Jeffrey M Wooldridge. 2007. *What's new in econometrics?* NBER.

Moffitt, Robert. 1993. "Identification and estimation of dynamic models with a time series of repeated cross-sections." *Journal of Econometrics*, 59(1): 99–123.

Verbeek, Marno. 2008. "Pseudo-panels and repeated cross-sections." In *The Econometrics of Panel Data*. 369–383. Springer.

Verbeek, Marno, and Francis Vella. 2005. "Estimating dynamic models from repeated cross-sections." *Journal of econometrics*, 127(1): 83–102.

White, Halbert. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica: Journal of the Econometric Society*, 817–838.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* . 2nd ed., Boston MA:MIT Press.

CHAPTER 2

EXPLORING ADDITIONAL MOMENT CONDITIONS IN NON-SEPARABLE MINIMUM DISTANCE ESTIMATION WITH AN APPLICATION TO PSEUDO PANELS

2.1 Introduction

Minimum distance (MD) estimation is a useful approach to recover structural estimates from reduced form estimates when there exists a known relationship between the structural and reduced form parameters. The known relationship is often in the form of structural equations, moment conditions,[1] or restrictions, which are terminology used interchangeably hereafter. When applying MD, researchers may encounter situations in which they need to introduce additional moment conditions into estimation. This could happen, for example, when some new instrument variables (IVs) become available as the research proceeds. An important question to ask in such a situation is whether we can always improve asymptotic efficiency by using all the moment conditions than using just part of them. In this chapter, we provide an affirmative answer to this question. We show that in MD estimation it never hurts to have more moment conditions. In particular, when the additional moment conditions are non-redundant, adding them to estimation strictly improves efficiency. This efficiency gain result echoes the similar property for generalized method of moments (GMM) in Breusch et al. (1999).

The motivation for deriving this efficiency gain result comes from the need in pseudo panel models to incorporate external IVs. A pseudo panel model can estimate an underlying unobserved effect panel data model with only repeated cross sections. The idea, which dates back to Deaton (1985), is to divide the population into a number of groups by certain predetermined group membership such as age cohorts. Then the group averages of the

[1] More precisely, the moment conditions in MD are conditional moment conditions.

variables can be used to construct a panel at the group level. Since the group averages are error ridden estimates, Deaton suggests to treat the estimation as a measurement error problem. In this chapter, we adopt the MD perspective proposed by Imbens and Wooldridge (2007). Within the MD framework, the group averages of the variables are the reduced form estimates, and the group averages of the panel data model are the structural equations linking the reduced form to the structural parameters. When new IVs become available, the additional set of structural equations induced by the IVs can be easily added to the estimation. Clearly, the aim of having more structural equations is to improve estimation efficiency. However, there is no such theory in MD estimation telling us whether efficiency gain can be achieved. Therefore, we attempt to derive such a result in this chapter to fill this gap.

We derive the result within a so called non-separable minimum distance (NMD) framework developed in this chapter. The framework is a special case of the "high level" MD framework in Newey and McFadden (1994).[2] The key difference between NMD and the high level MD framework is that NMD models the reduced form parameters explicitly. This feature makes the NMD framework convenient to use when our exact purpose is to recover structural estimates from reduced form estimates. The qualifier "non-separable" highlights NMD's capability to deal with structural equations that are non-separable in the structural and reduced form parameters. Note, however, that the separable framework, i.e. the Classical Minimum Distance (CMD) framework, is still covered as a special case.

We establish consistency and asymptotic normality within the NMD framework. We also derive the optimal weighting matrix for the over-identified case in which the the number of structural equations is greater than that of the structural parameters. The optimal weighting matrix turns out to be the asymptotic variance of the rescaled structural equations, which gives an intuitive explanation of the weighting procedure. That is, the optimal weighting

[2]In effect, the MD framework in Newey and McFadden (1994) is so general that both generalized method of moments (GMM) and classical minimum distance (CMD) are its special cases.

matrix readjusts the relative importance of the conditions according to their own volatility as well as their correlation with each other. Building on these basic results, we then give the main efficiency result discussed at the beginning of the chapter.

After the general results are established in the NMD framework, we apply them back to the case of pseudo panels with external IVs. We show that a pseudo panel NMD estimator with an arbitrary weighting matrix is a generalized least squares (GLS) estimator. The inverse of the optimal weighting matrix corresponds to the usual unconditional variance-covariance matrix in GLS estimation. As a result of the added structural equations, the optimal weighting matrix becomes block diagonal. This result generalizes the finding in Imbens and Wooldridge (2007) that the optimal weighting matrix is diagonal in the case without external IVs. The inclusion of extra IVs in pseudo panel models also highlights a typical case where the optimal weighting matrix should be used over the naive identity matrix. In the first chapter, we have shown that varying cohort sizes and cohort-wise heteroskedasticity in idiosyncratic errors are two typical causes of a non-identity yet diagonal optimal weighting matrix. When IVs are added, the optimal weighting matrix is usually block diagonal since within-cohort dependence between structural equations generally exists. As a result, it is more likely to achieve efficiency gain by using the optimal weighing matrix.

A related question is whether we can estimate pseudo panel models naively by applying fixed effect on the sample cohort means and then making the inference robust to heteroskedasticity and/or serial correlation. In this chapter, we show that the naive fixed effect coefficient estimator is still valid because it coincides with the NMD estimator using the identity weighting matrix. But the naive inference, whether made robust or not, is invalid because it is different from the correct NMD inference. The fundamental reason of the difference is that the naive inference only uses the cohort averages and ignores any individual level information. In terms of asymptotic theory, the naive inference requires the number of cohorts tend to infinity and the number of time periods remain fixed (see, e.g., Arellano (1987); Wooldridge (2010); Hansen (2007a)), or both tend to infinity (Kezdi (2003); Hansen

(2007*b*)). In pseudo panel models, however, we often have large cohort sizes but fixed numbers of cohorts and time periods. This fact makes the MD framework a natural fit to the pseudo panel models.

As mentioned in Verbeek (2008), repeated cross sections have several advantages over panel data sets. Because it is usually easier and less costly to collect random samples than panel data, the sample sizes of repeated cross sections are often much larger than common panel data sets. Moreover, repeated cross sections are naturally immune to attrition which is a common issue for panel data. Therefore, the availability of the NMD approach to pseudo panels potentially opens many new research opportunities in cases where unobserved individual fixed effects are a concern.

The rest of the chapter is organized as follows. In section 2, we lay out the NMD framework. The consistency and asymptotic normality the NMD estimator, the optimal weighting matrix, and the property that more moment conditions do not hurt are discussed. In section 3, we apply the NMD framework to pseudo panel models with additional instruments. There are also two special subsections in which we discuss the GLS perspective and the naive variance estimators. Section 4 contains a simulation study of the pseudo panel NMD estimators. The last section concludes.

2.2 The NMD framework

Minimum distance is essentially a delta method - it recovers structural estimates from reduced form estimates when there exists a known set of structural equations that links the structural and reduced form parameters. Formally, let $\mathbf{\Pi} \times \mathbf{\Theta}$ be an subset of $\mathbb{R}^P \times \mathbb{R}^K$, which is the product space for the reduced form parameter $\boldsymbol{\pi}$ and the structural parameter $\boldsymbol{\theta}$. Let $\mathbf{h} : \mathbf{\Pi} \times \mathbf{\Theta} \to \mathbb{R}^J$ be an vector-valued function satisfying

$$\mathbf{h}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = \mathbf{0} \tag{2.1}$$

for some true parameter value $(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \in \boldsymbol{\Pi} \times \boldsymbol{\Theta}$. Hereafter, $\mathbf{h}$ is referred to as the structural function, and eq. (2.1) is the set of structural equations. Suppose there is an estimator $\hat{\boldsymbol{\pi}} \xrightarrow{p} \boldsymbol{\pi}_0$. Then an NMD estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ is defined as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{argmin}\ \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})' \hat{\mathbf{W}}\, \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}). \tag{2.2}$$

where $\hat{\mathbf{W}}$ is a J-dimensional positive semi-definite matrix and $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$.

2.2.1 Consistency

A consistency result for NMD is summarized in the following theorem (similar to Theorem 2.6 in Newey and McFadden (1994)):

Theorem 1. *Suppose that $\hat{\boldsymbol{\pi}} \xrightarrow{p} \boldsymbol{\pi}_0$, $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, and (i) (Identification) $\mathbf{W}$ is positive semi-definite and $\mathbf{W}\mathbf{h}(\boldsymbol{\pi}_0, \boldsymbol{\theta}) = \mathbf{0}$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; (ii) (Boundedness) $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, which is compact; (iii) (Continuity) $h_j(\boldsymbol{\pi}, \boldsymbol{\theta})$ is continuous on $\boldsymbol{\Pi}$ and on $\boldsymbol{\Theta}$, for $j = 1, \cdots, J$; (iv) (Uniform convergence) $sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |h_j(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}) - h_j(\boldsymbol{\pi}_0, \boldsymbol{\theta})| \xrightarrow{p} 0$, for $j = 1, \cdots, J$. Then $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.*

Proof. See Appendix. □

2.2.2 Asymptotic normality

Theorem 3.2 in Newey and McFadden (1994) requires $\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{\Omega})$, which demands effort to verify when $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ takes on some general functional form. If in addition continuous differentiability of $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\pi}$ is assumed, a Taylor expansion of $\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ around $\boldsymbol{\pi}_0$ can be used to verify that $\sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{\Omega})$ holds. The verification step however could be saved with the establishment of the following theorem.

Theorem 2. *Suppose that $\hat{\boldsymbol{\theta}}$ satisfies (2.2), $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$ where $\mathbf{W}$ is positive semi-definite, and*

(i) $\boldsymbol{\pi}_0 \in interior(\boldsymbol{\Pi})$ and $\boldsymbol{\theta}_0 \in interior(\boldsymbol{\Theta})$;

(ii) $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ *is continuously differentiable with respect to* $\boldsymbol{\theta}$ *in a neighborhood* $\mathcal{N}(\boldsymbol{\theta}_0)$ *of* $\boldsymbol{\theta}_0$*, and* $\mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$ *is continuously differentiable with respect to* $\boldsymbol{\pi}$ *in a neighborhood* $\mathcal{N}(\boldsymbol{\pi}_0)$ *of* $\boldsymbol{\pi}_0$*;*

(iii) $\sqrt{n}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$*;*

(iv) For $\mathbf{L}(\boldsymbol{\pi}, \boldsymbol{\theta}) \equiv \nabla_{\boldsymbol{\theta}} \mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$*,* $sup_{\boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{\theta}_0)} \|\mathbf{L}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}) - \mathbf{L}(\boldsymbol{\pi}_0, \boldsymbol{\theta})\| \xrightarrow{p} 0$*, and for* $\mathbf{B}(\boldsymbol{\pi}, \boldsymbol{\theta}) \coloneqq \nabla_{\boldsymbol{\pi}} \mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta})$*,* $sup_{\boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{\theta}_0)} \|\mathbf{B}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}) - \mathbf{B}(\boldsymbol{\pi}_0, \boldsymbol{\theta})\| \xrightarrow{p} 0$*;*

(v) $\mathbf{L}'\mathbf{W}\mathbf{L}$ *is nonsingular, where* $\mathbf{L} \equiv \mathbf{L}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$*.*

Let $\mathbf{B} \equiv \mathbf{B}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$*. Then*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N(\mathbf{0}, \left(\mathbf{L}'\mathbf{W}\mathbf{L}\right)^{-1} \mathbf{L}'\mathbf{W}\mathbf{B}\boldsymbol{\Omega}\mathbf{B}'\mathbf{W}\mathbf{L} \left(\mathbf{L}'\mathbf{W}\mathbf{L}\right)^{-1}). \tag{2.3}$$

Proof. See Appendix. □

With the presence of the added smoothness assumption with respect to $\boldsymbol{\pi}$, the theorem above provides a more constructive and straightforward version of the "high level" theorem in Newey and McFadden (1994). To obtain the asymptotic variance of the MD estimator, all we need is to find the two partial derivatives of $\mathbf{h}$ and plugging them in eq. (2.3).

2.2.3 Optimal weighting matrix

The asymptotic variance in (2.3) depends on the probability limit $\mathbf{W}$ of the weighting matrix $\hat{\mathbf{W}}$. When $\mathbf{W} = \mathbf{M}^{-1}$ where

$$\mathbf{M} = \mathbf{B}\boldsymbol{\Omega}\mathbf{B}' \tag{2.4}$$

the asymptotic variance simplifies to

$$Avar\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\right) = (\mathbf{L}'(\mathbf{B}\boldsymbol{\Omega}\mathbf{B}')^{-1}\mathbf{L})^{-1}. \tag{2.5}$$

As shown in the following theorem, the inverse of $\mathbf{B}\boldsymbol{\Omega}\mathbf{B}'$ is the optimal weighting matrix since (2.5) is the "smallest" asymptotic variance that can be obtained by optimizing over all possible nonsingular weighting matrices.

Theorem 3. *Suppose* $\mathbf{M} = \mathbf{B}\boldsymbol{\Omega}\mathbf{B}'$ *is nonsingular. Then an NMD estimator with* $\hat{\mathbf{W}} \xrightarrow{p} \boldsymbol{W} = \mathbf{M}^{-1}$ *is asymptotically efficient in the class of NMD estimators based on the same set of structural eqations.*

Proof. See Appendix. □

The intuition for using an optimal weighting matrix is straightforward. Asymptotically, it is not about over-identification. Rather, it is because the the conditions in $\sqrt{n}\mathbf{h}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = \mathbf{0}$ are asymptotically random. More accurate conditions exhibit less volatility, and the conditions are potentially correlated. To use all the conditions optimally, more weights should be given to less volatile conditions, and the correlation between conditions should also be accounted for.

The best characterization of the relative volatility of all conditions is the the asymptotic variance-covariance matrix of the rescaled conditions. It turns out that $\mathbf{M}$ is exactly that variance-covariance matrix . Specifically, the first part of condition (ii) in Theorem 2 and a Taylor expansion imply that

$$\begin{aligned} \sqrt{n}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) &= \mathbf{B} \cdot \sqrt{n}\left(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0\right) + o_p(1), \\ &\xrightarrow{d} N(\mathbf{0}, \mathbf{B}\boldsymbol{\Omega}\mathbf{B}'). \end{aligned} \tag{2.6}$$

The optimal weighting operation is essentially a standardization that assigns more loadings to less volatile conditions and untangles the correlation between conditions. It standardizes the asymptotic variance-covariance matrix to an identity matrix. Admittedly, that the inverse of the optimal weighting matrix is the asymptotic variance is a known result which can be found in, e.g., Newey and McFadden (1994), and the idea of standardization by volatility can also be found in the generalized method of moments (GMM) and generalized least squares (GLS) literature. However, this intuitive explanation is often overlooked when it comes to MD estimation.

In the application to pseudo panel, the intuition is even clearer, for $\mathbf{M}$ is exactly the variance-covariance matrix of individual level residuals.

2.2.4 Estimation

Given a consistent estimator $\hat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}_0$, the NMD estimator using the identity weighting matrix, i.e.

$$\check{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{argmin}\ \mathbf{h}(\hat{\boldsymbol{\pi}},\boldsymbol{\theta})'\mathbf{h}(\hat{\boldsymbol{\pi}},\boldsymbol{\theta}),$$

can be used as an initial estimator for $\boldsymbol{\theta}_0$. Consistency of $\check{\boldsymbol{\theta}}$ follows from Theorem 1. By continuity of the partial derivatives, the plug-in estimator $\check{\mathbf{B}} = \nabla_\pi \mathbf{h}(\hat{\boldsymbol{\pi}},\check{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{B}$. Then, given a consistent estimator $\hat{\boldsymbol{\Omega}}$ for $\boldsymbol{\Omega}$, $\hat{\mathbf{M}} \equiv \check{\mathbf{B}}\hat{\boldsymbol{\Omega}}\check{\mathbf{B}}'$ is a consistent estimator for $\mathbf{M}$, and an asymptotically efficient for $\boldsymbol{\theta}_0$ can be obtained by

$$\hat{\boldsymbol{\theta}}^{opt} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{argmin}\ \mathbf{h}(\hat{\boldsymbol{\pi}},\boldsymbol{\theta})'\hat{\mathbf{M}}^{-1}\mathbf{h}(\hat{\boldsymbol{\pi}},\boldsymbol{\theta}).$$

The corresponding consistent estimator for the asymptotic variance-covariance matrix is given by

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}^{opt}) = (\hat{\mathbf{L}}'(\hat{\mathbf{B}}\hat{\boldsymbol{\Omega}}\hat{\mathbf{B}}')^{-1}\hat{\mathbf{L}})^{-1}/n$$

where $\hat{\mathbf{B}} \equiv \nabla_\pi \mathbf{h}(\hat{\boldsymbol{\pi}},\check{\boldsymbol{\theta}})$ and $\hat{\mathbf{L}} \equiv \nabla_\pi \mathbf{h}(\hat{\boldsymbol{\pi}},\check{\boldsymbol{\theta}})$.

The estimator defined above iterates only once. Multiple iterations are also allowed. They are, however, asymptotically equivalent.

2.2.5 More conditions do not hurt

Partition the restrictions in (2.1) into two parts:

$$\mathbf{h}(\boldsymbol{\pi}_0,\boldsymbol{\theta}_0) = \begin{bmatrix} \mathbf{h}_1(\boldsymbol{\pi}_0,\boldsymbol{\theta}_0) \\ \mathbf{h}_2(\boldsymbol{\pi}_0,\boldsymbol{\theta}_0) \end{bmatrix} = \mathbf{0}, \tag{2.7}$$

where $\mathbf{h}_1$ is $J_1 \times 1$, $\mathbf{h}_2$ is $J_2 \times 1$, and $J_1 + J_2 = J$. Let

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{argmin}\ \mathbf{h}_1(\hat{\boldsymbol{\pi}},\boldsymbol{\theta})'\mathbf{M}_{1,1}^{-1}\mathbf{h}_1(\hat{\boldsymbol{\pi}},\boldsymbol{\theta})$$

with $\mathbf{M}_{1,1} = \mathbf{B}_1\mathbf{\Omega}\mathbf{B}_1'$ for $\mathbf{L}_1 \equiv \mathbf{L}_1(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \equiv \nabla_{\boldsymbol{\theta}}\mathbf{h}_1(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$, $\mathbf{B}_1 \equiv \mathbf{B}_1(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \equiv \nabla_{\boldsymbol{\pi}}\mathbf{h}_1(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$. Then

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, [\mathbf{L}_1'(\mathbf{B}_1\mathbf{\Omega}\mathbf{B}_1')^{-1}\mathbf{L}_1]^{-1}).$$

On the other hand, if all restrictions are used, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{opt} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N(\mathbf{0}, [\mathbf{L}'(\mathbf{B}\mathbf{\Omega}\mathbf{B}')^{-1}\mathbf{L}]^{-1}).$$

The following theorem shows that asymptotically $\hat{\boldsymbol{\theta}}$ is at least as efficient as $\tilde{\boldsymbol{\theta}}$. The theorem as well as the proof is similar to its GMM counterpart in Breusch et al. (1999).

Theorem 4. *Let* $\mathbf{L}_i \equiv \mathbf{L}_i(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \equiv \nabla_{\boldsymbol{\theta}}\mathbf{h}_i(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$ *and* $\mathbf{B}_i \equiv \mathbf{B}_i(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \equiv \nabla_{\boldsymbol{\pi}}\mathbf{h}_i(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0)$ *for* $i = 1, 2$. *Let* $\mathbf{M}_{i,j} = \mathbf{B}_i\mathbf{\Omega}\mathbf{B}_j'$ *for* $i = 1, 2$ *and* $j = 1, 2$. *Assume* $\mathbf{B}\mathbf{\Omega}\mathbf{B}'$ *and* $\mathbf{B}_1\mathbf{\Omega}\mathbf{B}_1'$ *are both nonsingular. Let* $\mathbf{F} = \mathbf{M}_{2,2} - \mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}$. *Then*

$$\begin{aligned}&\mathbf{L}'(\mathbf{B}\mathbf{\Omega}\mathbf{B}')^{-1}\mathbf{L} - \mathbf{L}_1'(\mathbf{B}_1\mathbf{\Omega}\mathbf{B}_1')^{-1}\mathbf{L}_1 \\ =& \left(\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 - \mathbf{L}_2\right)' \mathbf{F}^{-1} \left(\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 - \mathbf{L}_2\right)\end{aligned}$$

and thus is positive semi-definite.

Proof. See Appendix. □

The condition $\mathbf{h}_2(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = \mathbf{0}$ is redundant if $\mathbf{L}_2 = \mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1$, i.e.

$$\mathbf{L}_2 = \mathbf{B}_2\mathbf{\Omega}\mathbf{B}_1'(\mathbf{B}_1\mathbf{\Omega}\mathbf{B}_1')^{-1}\mathbf{L}_1. \tag{2.8}$$

We can think of $\mathbf{\Phi} = (\mathbf{B}_1\mathbf{\Omega}\mathbf{B}_1')^{-1}\mathbf{B}_1\mathbf{\Omega}\mathbf{B}_2'$ as the coefficient matrix from the GLS of $\mathbf{B}_2'$ on $\mathbf{B}_1$ with weight $\mathbf{\Omega}$. Then $\mathbf{h}_2(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = \mathbf{0}$ is redundant if $\mathbf{L}_2$ is a linear transformation of $\mathbf{L}_1$ with the transformation matrix $\mathbf{\Phi}'$.

Eq. (2.8) is similar to condition (C) of Theorem 1 in Breusch et al. (1999). We can also derive a condition similar to condition (B) in that theorem to have a more intuitive explanation of the redundancy condition. Specifically, define

$$\sqrt{n}\mathbf{r}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) \equiv \sqrt{n}\mathbf{h}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) - \mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\sqrt{n}\mathbf{h}_1(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0). \tag{2.9}$$

By eq. (2.6), $\mathbf{M}_{1,1}$ is the asymptotic variance of $\sqrt{n}\mathbf{h}_1(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$, and $\mathbf{M}_{2,1}$ is the asymptotic covariance of $\mathbf{h}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ and $\mathbf{h}_1(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$. Therefore, asymptotically, $\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\sqrt{n}\mathbf{h}_1(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ is the linear projection of $\sqrt{n}\mathbf{h}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ on $\sqrt{n}\mathbf{h}_1(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$, and $\sqrt{n}\mathbf{r}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ is the residual in this linear projection. It follows that a redundancy condition that is equivalent to but more intuitive than eq. (2.8) is

$$\nabla_{\boldsymbol{\theta}} \left[\sqrt{n}\mathbf{r}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)\right] \equiv \nabla_{\boldsymbol{\theta}} \left[\sqrt{n}\mathbf{h}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) - \mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\sqrt{n}\mathbf{h}_1(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)\right] = \mathbf{0}.$$

That is, the condition for $\mathbf{h}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) = \mathbf{0}$ to be redundant is that $\mathbf{r}_2(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ is marginally uninformative for $\boldsymbol{\theta}_0$.

2.3 Pseudo panels with additional IVs

In the case of pseudo panels with additional IVs, the restrictions defined by (2.1) are not additively separable in $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ as in the CMD case. Therefore it serves as a good example to illustrate the NMD framework. Moreover, the first-order condition takes the form of the normal equation of a GLS estimation. Therefore, the optimal NMD estimator in this case turns out to be a GLS estimator using the optimal weighting matrix as the unconditional variance-covariance matrix.

The adoption of the MD perspective in pseudo panel models provides a new way to deal with errors in variables. In the seminal work of Deaton (1985), this issue is treated as a measurement error problem. By specifying the measurement error structure, Deaton proposes a measurement-error corrected estimator. Collado (1997) follows the measurement error thinking and extends Deaton's method to a more general measurement-error corrected GMM estimator. In the MD framework, group averages are treated as estimates for the reduced form parameters. Since the group sizes are usually large for repeated cross sections, the MD framework is a natural fit for pseudo panel models.

In the following subsections we go through the derivation of the particular contents of $\mathbf{h}$, $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, $\mathbf{L}$, $\mathbf{B}$ and $\mathbf{M}$ in the pseudo panel case, discuss estimation, and summarize the

asymptotics.

2.3.1 Population model and structural equations

Formally, for outcome y_{it}, covariate $\mathbf{x}_{it}$ $(1 \times K)$, coefficient $\boldsymbol{\beta}$ $(K \times 1)$, time effect η_t, fixed effect f_i, and idiosyncratic error u_{it}, consider the following model for a generic individual in the population

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \eta_t + f_i + u_{it},\ t = 1, \ldots, T. \tag{2.10}$$

f_i and u_{it} are unobserved. The first entry of $\mathbf{x}_{it}$ is unity for notation convenience. Essentially, we are thinking of the population as a genuine panel data set from which different samples are drawn each period. The same treatment is also adopted in Verbeek and Vella (2005).

Let $\mathbf{z}_{it} = (1, z_{2it}, \cdots, z_{Pit})$ be a $1 \times P$ row vector of instrumental variables satisfying

$$E(\mathbf{z}'_{it} u_{it} | f_i) = \mathbf{0}, \tag{2.11}$$

$$Cov(z_{pit}, f_i | g_i) = 0, \quad p = 1, 2, \cdots, P, \tag{2.12}$$

where, for convenience, z_{1it} is also set to 1.

In a standard panel, the conditional exogeneity of $\mathbf{x}_{it}$ given f_i is usually assumed:

$$E(u_{it} | \mathbf{x}_{it}, f_i) = 0,\ t = 1, \ldots, T. \tag{2.13}$$

This condition is not required here. A weaker condition that suffices is

$$E(u_{it} | f_i) = 0,\ t = 1, \ldots, T. \tag{2.14}$$

Note that by iterated expectation, (2.13) implies (2.14). Because f_i aggregates all time-constant unobservables, we should think of (2.13) and (2.14) as being true for not only the lump sum f_i but also any time constant factors. In particular, replacing f_i with the group indicator g_i (i.e. applying iterated expectation) leads to

$$E(u_{it} | g_i) = 0,\ t = 1, \ldots, T. \tag{2.15}$$

Let $E(\cdot|g)$ be the shorthanded notation for $E(\cdot|g_i = g)$, and let $\alpha_g = E(f_i|g)$ be the group fixed effect for group g. By (2.12) and the fact that

$$E(z_{pit} \cdot f_i|g) = Cov(z_{pit}, f_i|g) + E(z_{pit}|g) \cdot E(f_i|g),\ p = 1, \cdots, P,$$

the structural model follows as

$$\begin{aligned} E(\mathbf{z}'_{it} y_{it}|g) =& E(\mathbf{z}'_{it}\mathbf{x}_{it}|g)\boldsymbol{\beta} + E(\mathbf{z}'_{it}|g)\eta_t + E(\mathbf{z}'_{it}|g)\alpha_g, \\ & for\ t = 1, \ldots, T;\ g = 1, \cdots, G. \end{aligned} \tag{2.16}$$

Thanks to $z_{1it} = 1$, the first row in eq. (2.16) represents the cohort level equations without instruments, which is the basic case studied in Imbens and Wooldridge (2007).

The exogeneity condition (2.15) might appear non-substantial at the first glance, because it seems we can always make $E(u_{it}|g_i) = 0$ holds by subtracting $E(u_{it}|g_i)$ from u_{it} and redefine the deviation as u_{it}. But this subtraction operation is equivalent to the inclusion of a full set of cohort-time effects in the structural model (2.16). Perhaps the following equivalent representation of (2.15) makes the explanation clearer

$$\delta_{gt} = E(u_{it}|g_i = g) = 0,\ g = 1, \ldots, G,\ t = 1, \ldots, T. \tag{2.17}$$

If eq. (2.17) (or equivalently (2.15)) is not imposed, the GT parameters δ_{gt} for $g = 1, \ldots, G,\ t = 1, \ldots, T$ will enter the structural model (2.16) as the full set of cohort-time effects. Including the full set of cohort-time effects is equivalent to not imposing (2.17) (or (2.15)). Therefore, the key assumption disguised by eq. (2.15) together with the specification in (2.10) is that the structural model (2.16) requires only the set of group and time effects (η_t and α_g) but not the full set of cohort-time effects (δ_{gt}). If any such cohort-time effect is required, then, as pointed out in Imbens and Wooldridge (2007), one way to think about the misspecification is that some $\delta_{gt} = E(u_{it}|g_i = g)$ is not zero.

Note that the structural model with the full set of cohort-time effects is always correctly specified, but it is not interesting because the variation in the covariate cohort means is

absorbed by δ_{gt}. As a result, such a the model is only identified up to the GT cohort-time effects.

A technical point here is that, due to the setup $x_{1it} = z_{1it} = 1$, there are only $(G-1)$ parameters in α_g and $(T-1)$ in η_t to estimate. Imbens and Wooldridge (2007) make the normalization $\sum_{g=1}^{G} \alpha_g = 0$ and $\eta_1 = 0$. This chapter however proceeds with $\alpha_1 = 0$ and $\eta_1 = 0$. The purpose of this slightly different normalization is to cope with the estimation convention that the dummies for the first cohort and the first time period are always dropped. As a result of the dropout, the sum $(\beta_1 + \alpha_1 + \eta_1)$ is identified, but β_1, α_1 and η_1 are not separately identifiable. The remaining estimated group and time effects are the relative effects $(\alpha_g - \alpha_1)$ for $g = 2, \cdots, G$ and $(\eta_t - \eta_1)$ for $t = 2, \cdots, T$. Setting $\alpha_1 = \eta_1 = 0$ then conveniently simplifies $(\beta_1 + \alpha_1 + \eta_1)$, $(\alpha_g - \alpha_1)$ and $(\eta_t - \eta_1)$ to β_1, α_g and η_t.

2.3.2 Useful notations

Some notations are useful later. Let $\mu_{gt}^{x} = E(x_{it}|g)$ denote the population mean of a generic variable x_{it} conditional on $g_i = g$. For a vector (e.g. $\mathbf{x}_{it}$) or a matrix (e.g. $\mathbf{z}_{it}'\mathbf{x}_{it}$) variable, bold symbols like $\boldsymbol{\mu}_{gt}^{\mathbf{x}}$ or $\boldsymbol{\mu}_{gt}^{\mathbf{z}'\mathbf{x}}$ will be used. In this notation, eq. (2.16) can be written as

$$\begin{aligned} \mathbf{0} = -\boldsymbol{\mu}_{gt}^{\mathbf{z}'y} + \boldsymbol{\mu}_{gt}^{\mathbf{z}'\mathbf{x}}\boldsymbol{\beta} + \boldsymbol{\mu}_{gt}^{\mathbf{z}'}(\eta_t + \alpha_g), \\ for\ t = 1, \ldots, T;\ g = 1, \cdots, G. \end{aligned} \tag{2.18}$$

Also, for a generic variable x_{it} and $j = (g-1)T + t$, let $\boldsymbol{\mu}^{x}$ denote the column "vector" with $\boldsymbol{\mu}_{gt}^{x}$ the jth row block. Depending on the dimension of x_{it}, $\boldsymbol{\mu}^{x}$ can be either a column vector or a matrix.

Let $\mathbf{v}_{it} = (y_{it}, \mathbf{x}_{it})$ and $\mathbf{s}_{it} = \mathbf{z}_{it} \otimes \mathbf{v}_{it}$ with $\otimes$ denotes Kronecker product. $\mathbf{s}_{it}$ is a long row vector. Assume the variance-covariance matrix of $\mathbf{s}_{it}$ exists and is denoted by

$$\boldsymbol{\Omega}_{gt}^{\mathbf{s}} = Var(\mathbf{s}_{it}|g). \tag{2.19}$$

An explicit formula for $\boldsymbol{\Omega}_{gt}^{\mathbf{s}}$ is given in Appendix.

Then the reduced form parameter $\boldsymbol{\pi}$ in this pseudo panel case can be expressed as

$$\boldsymbol{\pi} = \boldsymbol{\mu}^{\mathbf{s}'} = (\boldsymbol{\mu}^{\mathbf{s}}_{11}, \boldsymbol{\mu}^{\mathbf{s}}_{12}, \cdots, \boldsymbol{\mu}^{\mathbf{s}}_{GT})'.$$

The structural parameter is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\eta}', \boldsymbol{\alpha}')'$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_G)'$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)'$.

Moreover, the right hand side of eq. (2.18) says the j-th row block of the $\mathbf{h}$ function is

$$\mathbf{h}_j(\boldsymbol{\pi}, \boldsymbol{\theta}) = -\boldsymbol{\mu}^{\mathbf{z}'y}_{gt} + \boldsymbol{\mu}^{\mathbf{z}'\mathbf{x}}_{gt}\beta + \boldsymbol{\mu}^{\mathbf{z}'}_{gt}(\eta_t + \alpha_g) \tag{2.20}$$

with $j = (g-1)T + t$. Note that each $\mathbf{h}_j(\boldsymbol{\pi}, \boldsymbol{\theta})$ is $P \times 1$. Let $\underline{\mathbf{x}}_{it} = (\mathbf{x}_{it}, \mathbf{d}, \mathbf{c})$ with $\mathbf{d}$ the vector of time dummies and $\mathbf{c}$ the vector of group dummies. Then a second useful expression for $\mathbf{h}_j$ is

$$\mathbf{h}_j(\boldsymbol{\pi}, \boldsymbol{\theta}) = -\boldsymbol{\mu}^{\mathbf{z}'y}_{gt} + \boldsymbol{\mu}^{\mathbf{z}'\underline{\mathbf{x}}}_{gt}\boldsymbol{\theta}. \tag{2.21}$$

Later we will see that the two expressions (2.20) and (2.21) are convenient for calculating partial derivatives of $\mathbf{h}$.

2.3.3 The partial derivatives L and B and the inverse optimal weighting matrix M

By eq. (2.21), it is trivial that

$$\mathbf{L} = \nabla_\theta \mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \boldsymbol{\mu}^{\mathbf{z}'\underline{\mathbf{x}}} \tag{2.22}$$

where, as defined in the last section, $\boldsymbol{\mu}^{\mathbf{z}'\underline{\mathbf{x}}}$ is the matrix with $\boldsymbol{\mu}^{\mathbf{z}'\underline{\mathbf{x}}}_{gt}$ the j-th row block for $j = (g-1)T + t$.

On the other hand, recall eq. (2.20). For $z_{1it} = 1$, $\boldsymbol{\mu}^{z_1}_{gt} = 1$, define $\nabla_{\boldsymbol{\mu}^{z_1}_{gt}}[\boldsymbol{\mu}^{z_1}_{gt}(\eta_t + \alpha_g)] = (\eta_t + \alpha_g)$. Define $\boldsymbol{\beta}_{gt}$ by replacing the first entry of $\boldsymbol{\beta}$, i.e. β_1, with $(\beta_1 + \eta_t + \alpha_g)$, and let $\mathbf{b}_{gt}(\boldsymbol{\theta}) = \mathbf{I}_P \otimes (-1, \boldsymbol{\beta}'_{gt})$ with $\mathbf{I}_P$ the P-dimensional identity matrix. Some algebra (see Appendix) then shows that

$$\nabla_{\boldsymbol{\pi}_{\tilde{g}\tilde{t}}} \mathbf{h}_j(\boldsymbol{\pi}, \boldsymbol{\theta}) \;=\; \begin{cases} \mathbf{b}_{gt}(\boldsymbol{\theta}), & if\ \tilde{g} = g\ and\ \tilde{t} = t, \\ \mathbf{0}, & otherwise. \end{cases}$$

Define another block diagonal matrix $\mathbf{b}(\boldsymbol{\theta})$ by putting $\mathbf{b}_{gt}(\boldsymbol{\theta})$ on its gt-th diagonal block. Then

$$\mathbf{B} = \nabla_{\boldsymbol{\pi}} \mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \mathbf{b}(\boldsymbol{\theta}). \tag{2.23}$$

With the general formula in eq. (2.4) and the particular contents in eq. (2.19) and (2.23), the inverse of the optimal weighting matrix, $\mathbf{M}$, is given by

$$\mathbf{M} = \mathbf{b}(\boldsymbol{\theta}) \boldsymbol{\Omega}^{\mathbf{s}} \mathbf{b}(\boldsymbol{\theta})'. \tag{2.24}$$

In the Appendix, we show that an expansion of the right hand side of eq. (2.24) leads to

$$\begin{aligned} \mathbf{M} = diag_{GT}[(\rho_1 \kappa_1)^{-1} \mathbf{b}_{11}(\boldsymbol{\theta}) \boldsymbol{\Omega}^{\mathbf{s}}_{11} \mathbf{b}_{11}(\boldsymbol{\theta})', (\rho_1 \kappa_2)^{-1} \mathbf{b}_{12}(\boldsymbol{\theta}) \boldsymbol{\Omega}^{\mathbf{s}}_{12} \mathbf{b}_{12}(\boldsymbol{\theta})', \cdots \\ \cdots, (\rho_G \kappa_T)^{-1} \mathbf{b}_{GT}(\boldsymbol{\theta}) \boldsymbol{\Omega}^{\mathbf{s}}_{GT} \mathbf{b}_{GT}(\boldsymbol{\theta})']. \end{aligned} \tag{2.25}$$

That is, $\mathbf{M}$ is a block diagonal matrix with $(\rho_g \kappa_t)^{-1} \mathbf{b}_{gt}(\boldsymbol{\theta}) \boldsymbol{\Omega}^{\mathbf{s}}_{gt} \mathbf{b}'_{gt}(\boldsymbol{\theta})$ on the gt-th diagonal block.

We also show in the Appendix that $\mathbf{b}_{gt}(\boldsymbol{\theta}) \boldsymbol{\Omega}^{\mathbf{s}}_{gt} \mathbf{b}'_{gt}(\boldsymbol{\theta})$ is actually the variance-covariance matrix of the composite errors within cell (g, t)

$$\mathbf{b}_{gt}(\boldsymbol{\theta}) \boldsymbol{\Omega}^{\mathbf{s}}_{gt} \mathbf{b}'_{gt}(\boldsymbol{\theta}) = \boldsymbol{\Xi}_{gt} \equiv Var[\mathbf{z}'_{it} y_{it} - \mathbf{z}'_{it} \mathbf{x}_{it} \boldsymbol{\beta} - \mathbf{z}'_{it}(\eta_t + \alpha_g) | g]. \tag{2.26}$$

Therefore another useful expression for $\mathbf{M}$ is

$$\mathbf{M} = diag_{GT} \left[(\rho_1 \kappa_1)^{-1} \boldsymbol{\Xi}_{11}, (\rho_1 \kappa_2)^{-1} \boldsymbol{\Xi}_{12}, \cdots, (\rho_G \kappa_T)^{-1} \boldsymbol{\Xi}_{GT} \right]. \tag{2.27}$$

2.3.4 Estimation

Assume we have T repeated cross-sectional random samples denoted by

$$\{(y_{it}, \mathbf{x}_{it}, \mathbf{z}_{it}, g_{it}),\ i = 1, \cdots, n_t;\ t = 1, \cdots T\}$$

where n_t is the number of observations for cross section t. Note that in each time period we have a new random sample, so in general the same index i refers to different individuals in different time periods, and thus g_{it} sees a subscript t added.

2.3.4.1 Asymptotics of $\hat{\boldsymbol{\pi}}$

Let 1_A be the indicator function equal to 1 if A is true and equal to 0 otherwise. Let $\mathbf{r}_{it} = (r_{it,1}, r_{it,2}, \ldots, r_{it,G})$ be a vector of group indicators with $r_{it,g} = 1\{g_{it} = g\}$ where $1\{\cdot\}$ is the indicator function equal to one if the event in $\{\cdot\}$ is ture. In this way the group membership of the random draw i at time t is properly treated as a random variable. It follows that the number of observations in cell (g,t) is also a random variable given by $n_{gt} = \sum_{i=1}^{n_t} r_{it,g}$.

Let $\hat{\mu}^x_{gt}$ denote the sample average within cell (g,t) for a generic variable x_{it}. Let $\rho_g = P(r_{it,g} = 1)$ be the fraction of the population in cohort g and assume $\hat{\rho}_{gt} = n_{gt}/n_t \xrightarrow{p} \rho_g$. Let $\kappa_t = lim_{n\to\infty} n_t/n$ be the fraction of all observations accounted for by cross section t. By (essentially) the central limit theorem, for $g = 1, \cdots, G$ and $t = 1, \cdots, T$,

$$\sqrt{n_t}(\hat{\boldsymbol{\mu}}^{\mathbf{s}\,\prime}_{gt} - \boldsymbol{\mu}^{\mathbf{s}\,\prime}_{gt}) \xrightarrow{d} Normal(\mathbf{0}, (\rho_g\kappa_t)^{-1}\boldsymbol{\Omega}^{\mathbf{s}}_{gt}).$$

Furthermore, let $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\mu}}^{\mathbf{s}}_{11}, \hat{\boldsymbol{\mu}}^{\mathbf{s}}_{12}, \cdots, \hat{\boldsymbol{\mu}}^{\mathbf{s}}_{GT})'$ and $\boldsymbol{\pi}_0 = (\boldsymbol{\mu}^{\mathbf{s}}_{gt}, \boldsymbol{\mu}^{\mathbf{s}}_{gt}, \cdots, \boldsymbol{\mu}^{\mathbf{s}}_{gt})'$. Then the results above can be stacked in

$$\sqrt{n}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}^{\mathbf{s}})$$

where $\boldsymbol{\Omega}^{\mathbf{s}}$ is a block diagonal matrix with $(\rho_g\kappa_t)^{-1}\boldsymbol{\Omega}^{\mathbf{s}}_{gt}$ on the gt-th diagonal block.

2.3.4.2 Estimation of L

Eq. (2.22) suggests that a straightforward estimator for $\mathbf{L}$ is

$$\hat{\mathbf{L}} = \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}. \tag{2.28}$$

$\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}$ is the sample analog of $\boldsymbol{\mu}^{\mathbf{z}'\underline{\mathbf{x}}}$. Recall that $\underline{\mathbf{x}}_{it} = (\mathbf{x}_{it}, \mathbf{d}, \mathbf{c})$ and that $\mathbf{x}_{it}$ contains a constant term. Then $\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}$ is the matrix of the sample cohort means of the explanatory variables, the instruments, and their interactions. Its dimension is $GPT \times (K + G + T - 2)$.

2.3.4.3 The general estimator $\hat{\theta}$ and the FE estimator $\check{\theta}$

There exists an analytical solution to eq. (2.2) in the current setup, which turns out to be a GLS estimator.

Specifically, given eq.(2.21) and (2.28), the first-order condition to eq. (2.2) in the current setup can be written as[3]

$$(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{W}}\left(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}^{\mathbf{z}'y}\right) = \mathbf{0}.$$

Assume $(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{W}}\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}$ is nonsingular, then the general pseudo panel NMD estimator with a weighting matrix $\hat{\mathbf{W}}$ is given by

$$\hat{\boldsymbol{\theta}} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{W}}\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{W}}\hat{\boldsymbol{\mu}}^{\mathbf{z}'y} \tag{2.29}$$

Clearly, (2.29) is of the form of a GLS estimator where $\hat{\mathbf{W}}$ serves as the inverse of the "unconditional variance-covariance matrix of the error term", and the cohort means $\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}$ and $\hat{\boldsymbol{\mu}}^{\mathbf{z}'y}$ are the matrix of right-hand-side variables and left-hand-side variable, receptively.

In particular, replacing $\hat{\mathbf{W}}$ with the identity matrix gives the fixed effect estimator

$$\check{\boldsymbol{\theta}} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}^{\mathbf{z}'y}. \tag{2.30}$$

The standard case without instrument in Imbens and Wooldridge (2007) corresponds to the case $P = 1$, i.e. deleting the letter $\mathbf{z}$ in eq. (2.30).

2.3.4.4 Estimation of B, M and $\hat{\theta}^{opt}$

With $\check{\boldsymbol{\theta}}$ as an initial estimator, an estimator for $\mathbf{B}$ follows from eq. (2.23) by substituting $\boldsymbol{\theta}$ with $\check{\boldsymbol{\theta}}$ which leads to

$$\hat{\mathbf{B}} = \mathbf{b}(\check{\boldsymbol{\theta}}).$$

An obvious estimator for the variance-covariance matrix of $\mathbf{s}$ defined in (2.19) is

$$\hat{\boldsymbol{\Omega}}^{\mathbf{s}}_{gt} = n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g}(\mathbf{s}_{it} - \hat{\boldsymbol{\mu}}^{\mathbf{s}}_{gt})'(\mathbf{s}_{it} - \hat{\boldsymbol{\mu}}^{\mathbf{s}}_{gt}).$$

[3]Note that $(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'$ is the transpose of $\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}$ and is not the same as $\hat{\boldsymbol{\mu}}^{\underline{\mathbf{x}}'\mathbf{z}}$.

Then an estimator for $\mathbf{\Omega^s}$, $\hat{\mathbf{\Omega}}^{\mathbf{s}}$, can be defined as the block diagonal matrix with the gt-th diagonal block $(n_{gt}/n)^{-1}\hat{\mathbf{\Omega}}^{\mathbf{s}}_{gt}$, i.e.,

$$\hat{\mathbf{\Omega}}^{\mathbf{s}} = diag_{GT}((n_{11}/n)^{-1}\hat{\mathbf{\Omega}}^{\mathbf{s}}_{11}, (n_{12}/n)^{-1}\hat{\mathbf{\Omega}}^{\mathbf{s}}_{12}, \cdots, (n_{GT}/n)^{-1}\hat{\mathbf{\Omega}}^{\mathbf{s}}_{GT}).$$

Given $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Omega}}^{\mathbf{s}}$, the following estimator for the inverse of the optimal weighting matrix follows from eq. (2.24)

$$\hat{\mathbf{M}} = \mathbf{b}(\check{\boldsymbol{\theta}})\hat{\mathbf{\Omega}}^{\mathbf{s}}\mathbf{b}(\check{\boldsymbol{\theta}})'. \tag{2.31}$$

Eq. (2.31), however, may involve big matrices in calculation when the number of covariates and/or instruments is large (the dimension of $\mathbf{s}$ increase quickly with multiple instruments). Fortunately, eq. (2.27) provides an alternative but numerically equivalent way to estimate $\mathbf{M}$ - all we need is an estimator for $\mathbf{\Xi}_{gt}$. By eq. (2.26), $\mathbf{\Xi}_{gt}$ can be conveniently estimated by $\hat{\mathbf{\Xi}}_{gt}$, the sample variance-covariance matrix of the residuals in cell (g,t) to be defined as follows.

First, using the fixed effect estimator $\check{\boldsymbol{\theta}}$ to obtain the individual residual

$$\check{u}_{it} \;=\; y_{it} - \mathbf{x}_{it}\check{\boldsymbol{\beta}} - (\check{\eta}_t + \check{\alpha}_g). \tag{2.32}$$

The cohort residual is then defined as

$$\hat{\mu}^{z_p\check{u}}_{gt} \;=\; n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g} z_{pit}\check{u}_{it}. \tag{2.33}$$

For $p, q = 1, \cdots, P$, let $\hat{\tau}_{pq}$ (drop subscript g,t from τ for simplicity) denote the entry on row p, column q of $\hat{\mathbf{\Xi}}_{gt}$. Then $\hat{\tau}_{pq}$ is given by

$$\hat{\tau}_{pq} = n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g}\left(z_{pit}\check{u}_{it} - \hat{\mu}^{z_p\check{u}}_{gt}\right)\left(z_{qit}\check{u}_{it} - \hat{\mu}^{z_q\check{u}}_{gt}\right). \tag{2.34}$$

Finally, $\hat{\mathbf{\Xi}}_{gt}$ is defined as the matrix with the pq-th entry $\hat{\tau}_{pq}$

$$\hat{\mathbf{\Xi}}_{gt} = (\hat{\tau}_{pq}).$$

Given $\hat{\boldsymbol{\Xi}}_{gt}$, the second method to estimate $\mathbf{M}$ is via

$$\hat{\mathbf{M}} = diag_{GT}[(n_{11}/n)^{-1}\hat{\boldsymbol{\Xi}}_{11}, (n_{12}/n)^{-1}\hat{\boldsymbol{\Xi}}_{12}, \cdots, (n_{GT}/n)^{-1}\hat{\boldsymbol{\Xi}}_{GT}] \tag{2.35}$$

which is the block diagonal matrix with the gt-th diagonal block $(n_{gt}/n)^{-1}\hat{\boldsymbol{\Xi}}_{gt}$.

The numerical equivalence of the two estimators for $\mathbf{M}$ is summarized in the following theorem.

Theorem 5. *The two ways of computing* $\hat{\mathbf{M}}$ *defined in eq. (2.31) and (2.35) are numerically equivalent.*

Proof. See Appendix. □

When $p = q = 1$, $\hat{\tau}_{11} = n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g}\left(\check{u}_{it} - \check{\mu}_{gt}^{u}\right)^2$ which is of the same form as the $\hat{\tau}^2$ defined in Imbens and Wooldridge (2007),[4] and $\hat{\mathbf{M}}$ becomes a diagonal matrix that coincides with the matrix $\hat{\mathbf{C}}$ in Imbens and Wooldridge (2007).

With $\hat{\mathbf{M}}$ in hand, by replacing $\hat{\mathbf{W}}$ with $\hat{\mathbf{M}}^{-1}$ in eq. (2.29), the optimal pseudo panel NMD estimator is obtained as

$$\hat{\boldsymbol{\theta}}^{opt} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{M}}^{-1}\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{M}}^{-1}\hat{\boldsymbol{\mu}}^{\mathbf{z}'y}. \tag{2.36}$$

The above formula in its appearance is similar to a GLS estimator on the cohort level data. But the weighting matrix is not the usual one used by a feasible GLS because $\hat{\mathbf{M}}$ is computed from individual level data. For more detail about the connection and difference of $\hat{\boldsymbol{\theta}}^{opt}$ to GLS, see the next section.

[4]The formula in Imbens and Wooldridge (2007) needs the correction of demeaning. Because in STATA, the command for calculating the sample variance automatically demeans the residuals.

2.3.4.5 Estimation of the asymptotic variances of $\hat{\boldsymbol{\theta}}$, $\check{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{opt}$

With all the pieces worked out, and by Theorem 2, the asymptotic variance estimator for $\hat{\boldsymbol{\theta}}$ is

$$\widehat{Avar(\hat{\boldsymbol{\theta}})} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{W}}\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}\left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{W}}\hat{\mathbf{M}}\hat{\mathbf{W}}\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)\left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{W}}\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}/n.$$

For $\check{\boldsymbol{\theta}}$, it is

$$\widehat{Avar(\check{\boldsymbol{\theta}})} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}\left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{M}}\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)\left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}/n. \tag{2.37}$$

Finally, for $\hat{\boldsymbol{\theta}}^{opt}$, it is

$$\widehat{Avar(\hat{\boldsymbol{\theta}}^{opt})} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\mathbf{M}}^{-1}\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}/n.$$

With the presence of additional IVs, dependence between restrictions are introduced since each cohort repeats itself several times in the restrictions. The optimal weighing matrix is more likely to be non-diagonal (it is block diagonal with block $(n_{gt}/n)^{-1}\hat{\boldsymbol{\Xi}}_{gt}$). In fact, some algebra (see Appendix) shows that another expression for $\boldsymbol{\Xi}_{gt}$ is

$$\boldsymbol{\Xi}_{gt} = E\left[(\varepsilon_i^f + u_{it})^2\mathbf{z}_{it}'\mathbf{z}_{it}|g\right]. \tag{2.38}$$

where

$$\varepsilon_i^f \equiv f_i - \alpha_g \tag{2.39}$$

is the deviation of individual effect from its cohort mean. Without further assumptions regarding the correlation between the quadratic terms $(\varepsilon_i^f + u_{it})^2$ and $\mathbf{z}_{it}'\mathbf{z}_{it}$, and the correlation among the IVs in $\mathbf{z}_{it}$, $\boldsymbol{\Xi}_{gt}$ is generally non-diagonal. As a result, the use of optimal weighting matrix becomes more important with the presence of additional IVs.

2.3.5 The GLS perspective

To better understand the relationship between $\hat{\boldsymbol{\theta}}^{opt}$ and its relation to GLS, define the individual composite error as

$$e_{it} \equiv y_{it} - \mathbf{x}_{it}\boldsymbol{\beta} - (\eta_t + \alpha_g) = \varepsilon_i^f + u_{it}.$$

The residual $\check{u}_{it}$ given in (2.32) is obviously an consistent estimator for e_{it}. With e_{it}, the vector of individual composite errors in cohort g is $\mathbf{z}'_{it}e_{it}$, and an alternative expression for $\mathbf{\Xi}_{gt}$ in (2.26) is

$$\mathbf{\Xi}_{gt} = Var[\mathbf{z}'_{it}e_{it}|g]. \tag{2.40}$$

For a given g, define the cohort composite error as

$$\hat{\mu}^{zpe}_{gt} = n^{-1}_{gt}\sum_{i=1}^{n_t} r_{it,g}z_{pit}e_{it}. \tag{2.41}$$

$\hat{\boldsymbol{\mu}}^{\mathbf{z}'e}_{gt}$ is similarly defined and represents the vector of cohort composite errors in cell (g,t). The variance-covariance matrix of $\hat{\boldsymbol{\mu}}^{\mathbf{z}'e}_{gt}$ conditional on g is given by

$$Var[\hat{\boldsymbol{\mu}}^{\mathbf{z}'e}_{gt}|g] = n^{-1}_{gt}\mathbf{\Xi}_{gt}. \tag{2.42}$$

From the MD perspective, n_{gt} is large, and $n^{-1}_{gt}\mathbf{\Xi}_{gt} \to 0$ as $n_{gt} \to \infty$. It thus does not make sense to model and estimate the "cohort composite errors" because they degenerate to 0 asymptotically.

The usual feasible GLS on the pseudo panel of cohort means ignores individual level data and relies on much stringent assumptions on the unconditional variance-covariance structure of the cohort composite error. In particular, the underlying asymptotics rely on large G. The GLS estimator in eq. (2.36) is apparently not the usual feasible GLS. Rather, it is an GLS imposing the following block diagonal variance-covariance structure of all the cohort composite errors

$$diag_{GT}[n^{-1}_{11}\mathbf{\Xi}_{11}, n^{-1}_{12}\mathbf{\Xi}_{12}, \cdots, n^{-1}_{GT}\mathbf{\Xi}_{GT}] \tag{2.43}$$

Eq. (2.43) contains $GTP(P+1)/2$ parameters, and thus is never feasible if only the GTP cohort means are observed. But if the individual level data are available, eq. (2.43) can be well estimated by

$$diag_{GT}[n^{-1}_{11}\hat{\mathbf{\Xi}}_{11}, n^{-1}_{12}\hat{\mathbf{\Xi}}_{12}, n^{-1}_{GT}\hat{\mathbf{\Xi}}_{GT}] = n^{-1}\hat{\mathbf{M}}. \tag{2.44}$$

Using $n^{-1}\hat{\mathbf{M}}$ in a GLS formula leads to eq. (2.36); the n^{-1} cancels off. From the GLS perspective, the weighting by $n^{-1}\hat{\mathbf{M}}^{-1}$ standardizes the sample cohort composite errors so that they become close to uncorrelated and homoskedastic.

It is worth noting that eq. (2.43) is not the unconditional variance-covariance matrix of $\hat{\boldsymbol{\mu}}^{\mathbf{z}'e}$. On each diagonal block is eq. (2.42), the conditional variance-covariance matrix of $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'e}$ given g.

From the MD perspective, there is no asymptotic variance-covariance matrix for the sample cohort composite errors because $Var[\hat{\boldsymbol{\mu}}^{\mathbf{z}'e}] \to \mathbf{0}$ as $n_{gt} \to \infty$; so is $n^{-1}\hat{\mathbf{M}}^{-1} \to 0$ as $n_{gt} \to \infty$. Rather, what matters is the following set of relocated and rescaled estimated structural equations,

$$\begin{aligned}
\mathbf{0} =& -\sqrt{n}(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'y} - \boldsymbol{\mu}_{gt}^{\mathbf{z}'y}) + \sqrt{n}(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'\mathbf{x}} - \boldsymbol{\mu}_{gt}^{\mathbf{z}'\mathbf{x}})\beta + \sqrt{n}(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'} - \boldsymbol{\mu}_{gt}^{\mathbf{z}'})(\eta_t + \alpha_g), && (2.45)\\
=& -\sqrt{n}\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'e}, && (2.46)\\
& for\ t = 1, \ldots, T;\ g = 1, \cdots, G.
\end{aligned}$$

The above equation is obtained by manipulating eq. (2.18). Asymptotically, eq. 2.46 converges to GTP random restrictions. The asymptotic variance of $\sqrt{n}\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'e}$ is exactly $\mathbf{M}$. Therefore, to use all the random conditions efficiently, the random restrictions need to be weighted by the square root of $\mathbf{M}^{-1}$. In estimation, $\mathbf{M}^{-1}$ is replaced by $\hat{\mathbf{M}}^{-1}$, but their function is equivalent asymptotically. Given fixed G, T and P, and $n_{gt} \to \infty$, the use of $\hat{\mathbf{M}}^{-1}$ is totally legit since $\hat{\mathbf{M}}^{-1}$ does not converge to 0. The weight $\hat{\mathbf{M}}^{-1}$ adjusts the relative importance of each sample restriction according to its level of accuracy. The level of accuracy for the gt-th sample restriction is measured by $(n_{gt}/n)^{-1}\hat{\boldsymbol{\Xi}}_{gt}$.

2.3.6 Naive variance estimators for $\check{\boldsymbol{\theta}}$

Because $\check{\boldsymbol{\theta}}$ is the fixed effect estimator on the pseudo panel of the sample cohort means, it is also convenient to compute the usual asymptotic variance estimators for a fixed effect

estimator. These naive estimators, however, are generally incorrect because they only make use of the sample cohort means.

Before listing the formulae for the naive variance estimators, we summarize several possible reasons in repeated random cross sections that may ruin their validity. We cite the reasons that apply to the breakdown of each estimator in later discussion.

1. $\hat{\mu}_{gt}^{zp\check{u}}$ and $\hat{\mu}_{gt}^{zq\check{u}}$ are generally correlated (dependence over p for fixed g and t)
2. the variance of $\hat{\mu}_{gt}^{zp\check{u}}$, as well as the covariance of $\hat{\mu}_{gt}^{zp\check{u}}$ and $\hat{\mu}_{gt}^{zq\check{u}}$, depends on $\mathbf{z}_{it}$ (heteroskedasticity)
3. $\hat{\mu}_{gt}^{zp\check{u}}$ depends on g because of either $\mathbf{z}_{it}$ or even u_{it} itself depends on g (non-identical distribution over g)

Among the three items, the last one is the most crucial because all the naive variance estimators discussed below rely on large G.

We consider three naive asymptotic variance estimators. Their formulae in a standard model can be found in Wooldridge (2010) as well as other textbooks. The first is the non-robust variance estimator for which the consistency relies on a scalar (proportional to an identity matrix) variance-covariance structure of the cohort composite errors. To obtain this formula, recall the definition of $\hat{\mu}_{gt}^{zp\check{u}}$ in eq. (2.33). Define the mean squared error for the pseudo panel as

$$\check{\sigma}^2 \quad = \quad (GTP - K - G - T + 2)^{-1} \sum_{g,t,p} (\hat{\mu}_{gt}^{zp\check{u}})^2.$$

Then the naive non-robust variance estimator can be written as

$$\widehat{Avar_{\mathtt{n}}}(\check{\boldsymbol{\theta}}) = \check{\sigma}^2 \left(\sum_{g,t,p} \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' z_p} \hat{\boldsymbol{\mu}}_{gt}^{z_p \underline{\mathbf{x}}} \right)^{-1}.$$

The subscript $\mathtt{n}$ in typewriter font stands for "non-robust". Its validity hings on i.i.d. sampling over (g, t, p) and homoskedasticity of $\hat{\mu}_{gt}^{zp\check{u}}$, neither of which holds in a pseudo panel of sample cohort means due to all three reason listed.

The second is the naive heteroskedasticity-robust variance estimator whose formula is given by

$$\widehat{Avar_{\mathtt{r}}(\check{\boldsymbol{\theta}})} = \left(\sum_{g,t,p} \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gt}^{zp\underline{\mathbf{x}}}\right)^{-1} \left(\sum_{g,t,p} (\hat{\mu}_{gt}^{zp\check{u}})^2 \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gt}^{zp\underline{\mathbf{x}}}\right) \left(\sum_{g,t,p} \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gt}^{zp\underline{\mathbf{x}}}\right)^{-1}.$$

The subscript $\mathtt{r}$ stands for "robust". The estimator is robust to heteroskedasticity in the cohort composite error $\hat{\mu}_{gt}^{zpe}$. But its validity still relies on i.i.d. sampling over (g,t,p) which does not hold due to reasons 1 and 3 mentioned above.

The third is the naive cluster-robust variance estimator and its formula is

$$\widehat{Avar_{\mathtt{c}}(\check{\boldsymbol{\theta}})} = \left(\sum_{g,t,p} \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gt}^{zp\underline{\mathbf{x}}}\right)^{-1} \left(\sum_{g,t,r,p,q} \hat{\mu}_{gt}^{zp\check{u}} \hat{\mu}_{gr}^{\check{u}zq} \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gr}^{zq\underline{\mathbf{x}}}\right) \left(\sum_{g,t,p} \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gt}^{zp\underline{\mathbf{x}}}\right)^{-1}.$$

The middle term can also be written as

$$\sum_{g,t,r,p,q} \hat{\mu}_{gt}^{zp\check{u}} \hat{\mu}_{gr}^{\check{u}zq} \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gr}^{zq\underline{\mathbf{x}}} = \sum_{g,t,p} (\hat{\mu}_{gt}^{zp\check{u}})^2 \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gt}^{zp\underline{\mathbf{x}}} + \sum_{g,t\neq r,p\neq q} \hat{\mu}_{gt}^{zp\check{u}} \hat{\mu}_{gr}^{\check{u}zq} \hat{\boldsymbol{\mu}}_{gt}^{\underline{\mathbf{x}}' zp} \hat{\boldsymbol{\mu}}_{gr}^{zq\underline{\mathbf{x}}}$$

where the first sum is exactly the middle term in the naive heteroskedasticity-robust variance estimator. The naive cluster-robust variance estimator is robust to arbitrary heteroskedasticity and serial correlation in the cohort composite errors $\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'e}$. But its validity relies on i.i.d. sampling over g which may not hold due to reason 3 listed above.

Some other equivalent representations of the three naive variance estimators are informative of their link to $\widehat{Avar(\check{\boldsymbol{\theta}})}$. Write the three naive estimators as

$$\widehat{Avar_{\mathtt{n}}(\check{\boldsymbol{\theta}})} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1} \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' (\check{\sigma}^2 \mathbf{I}) \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right) \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1},$$

$$\widehat{Avar_{\mathtt{r}}(\check{\boldsymbol{\theta}})} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1} \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' \left(diag(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}})\right)^2 \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right) \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1},$$

$$\widehat{Avar_{\mathtt{c}}(\check{\boldsymbol{\theta}})} = \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1} \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' diag_G(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}}) diag_G(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}})' \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right) \left((\hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}})' \hat{\boldsymbol{\mu}}^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1},$$

where $diag(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}})$ is the square, diagonal matrix created by putting the vector $\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}}$ on the principal diagonal, and $diag_G(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}})$ is the block diagonal matrix with the gth diagonal block $\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}}$ for $g = 1, \cdots, G$. Then clearly, the three naive variance estimators can be obtained

by replacing $\hat{\mathbf{M}}/n$ in eq. (2.37) with $(\check{\sigma}^2\mathbf{I})$, $\left(diag(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}})\right)^2$ or $diag_G(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}})diag_G(\hat{\boldsymbol{\mu}}^{\mathbf{z}'\check{u}})'$, respectively.

Yet another set of equivalent representations provide some insights on the large G perspective of the naive estimators. Specifically, write

$$\begin{aligned}
\widehat{Avar_{\mathtt{n}}(\check{\boldsymbol{\theta}})} &= \left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}\left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'(\check{\sigma}^2\mathbf{I}_{TP})\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)\left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}, \\
\widehat{Avar_{\mathtt{r}}(\check{\boldsymbol{\theta}})} &= \left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}\left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\left(diag(\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}})\right)^2\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)\left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}, \\
\widehat{Avar_{\mathtt{c}}(\check{\boldsymbol{\theta}})} &= \left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}\left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}}(\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)\left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}.
\end{aligned} \tag{2.47}$$

where $diag(\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}})$ is the square, diagonal matrix created by putting the vector $\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}}$ on the principal diagonal. In essence, the three naive variance estimators differ in estimating (treat g as random)

$$E\left[(\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'e}(\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'e})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right],$$

i.e. the middle term of the sandwich-form. But they all need i.i.d. sampling over g, which is not satisfied in the MD framework due to reason 3 listed above. The estimation errors in the cohort means are also ignored.

The last point we want to make is about the relationship between $\widehat{Avar_{\mathtt{c}}(\check{\boldsymbol{\theta}})}$ and $\widehat{Avar(\check{\boldsymbol{\theta}})}$. First, rewrite

$$\widehat{Avar(\check{\boldsymbol{\theta}})} = \left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}\left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'(n^{-1}\hat{\mathbf{M}}_g)\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)\left(\sum_g (\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}})'\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\underline{\mathbf{x}}}\right)^{-1}. \tag{2.48}$$

It is then clear that $\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}}(\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}})'$ and $n^{-1}\hat{\mathbf{M}}_g$ are the the only difference between $\widehat{Avar_{\mathtt{c}}(\check{\boldsymbol{\theta}})}$ and $\widehat{Avar(\check{\boldsymbol{\theta}})}$. Notice that $\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}}(\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'\check{u}})' \approx diag(\hat{\boldsymbol{\mu}}_{g1}^{\mathbf{z}'\check{u}}(\hat{\boldsymbol{\mu}}_{g1}^{\mathbf{z}'\check{u}})', \cdots, \hat{\boldsymbol{\mu}}_{gT}^{\mathbf{z}'\check{u}}(\hat{\boldsymbol{\mu}}_{gT}^{\mathbf{z}'\check{u}})')$ and that $n^{-1}\hat{\mathbf{M}}_g = diag(n_{g1}^{-1}\hat{\boldsymbol{\Xi}}_{g1}, \cdots, n_{gT}^{-1}\hat{\boldsymbol{\Xi}}_{gT})$. Moreover, notice that $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'\check{u}}(\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'\check{u}})' = (\hat{\mu}_{gt}^{zp\check{u}}\hat{\mu}_{gt}^{zq\check{u}})_{p,q}$ and that

$n_{gt}^{-1}\hat{\boldsymbol{\Xi}}_{gt} = n_{gt}^{-1}(\hat{\tau}_{pq})_{p,q}$. Therefore, the comparison boils down to the difference between

$$\hat{\mu}_{gt}^{zp\check{u}}\hat{\mu}_{gt}^{zq\check{u}} = \left(n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g} z_{pit}\check{u}_{it}\right)\left(n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g} z_{qit}\check{u}_{it}\right)$$

[5]and

$$n_{gt}^{-1}\hat{\tau}_{pq} = n_{gt}^{-2}\sum_{i=1}^{n_t} r_{it,g}\left(z_{pit}\check{u}_{it} - \hat{\mu}_{gt}^{zp\check{u}}\right)\left(z_{qit}\check{u}_{it} - \hat{\mu}_{gt}^{zq\check{u}}\right).$$

That is, $\widehat{Avar_{\mathsf{c}}}(\check{\boldsymbol{\theta}})$ uses $\hat{\mu}_{gt}^{zp\check{u}}\hat{\mu}_{gt}^{zq\check{u}}$ to approximate the covariance between the cohort composite errors $\hat{\mu}_{gt}^{zpe}$ and $\hat{\mu}_{gt}^{zqe}$, which uses only cohort-level information, and is not an estimator for $Cov(\hat{\mu}_{gt}^{zpe}, \hat{\mu}_{gt}^{zqe}|g)$ because $\hat{\mu}_{gt}^{zp\check{u}}\hat{\mu}_{gt}^{zq\check{u}}$ is observed only once for given g, t, p. On the other hand, $\widehat{Avar}(\check{\boldsymbol{\theta}})$ uses $\hat{\tau}_{pq}$ to estimate the covariance between the individual composite errors $z_p e$ and $z_q e$, which uses individual-level information, and is indeed an estimator for $Cov(z_p e, z_q e|g)$ because $\hat{\tau}_{pq}$ averages over n_{gt} observations. The additional n_{gt}^{-1} then transform it to a legitimate estimator for $Cov(\hat{\mu}_{gt}^{zpe}, \hat{\mu}_{gt}^{zqe}|g)$. Apparently, $n_{gt}^{-1}\hat{\tau}_{pq}$ is a better estimator for $Cov(\hat{\mu}_{gt}^{zpe}, \hat{\mu}_{gt}^{zqe}|g)$ than $\hat{\mu}_{gt}^{zp\check{u}}\hat{\mu}_{gt}^{zq\check{u}}$.

What conclusion do we get from this comparison? First of all, $\widehat{Avar_{\mathsf{c}}}(\check{\boldsymbol{\theta}})$ can only make sense if we have random sample over g, because eq. (2.47) averages over g. Second, a relatively large number of groups is also needed for the large G asymptotics to work. In the just-identified case, there is no $\widehat{Avar_{\mathsf{c}}}(\check{\boldsymbol{\theta}})$ because the residuals are all 0. Third, $\widehat{Avar_{\mathsf{c}}}(\check{\boldsymbol{\theta}})$ also needs fixed n_{gt}, otherwise the cohort composite error $\hat{\boldsymbol{\mu}}_{gt}^{\mathbf{z}'e}$ degenerates to 0. This is however not too much a problem because in a sample n_{gt} is always finite.

2.4 Simulation

This section contains a simulation study for the optimal NMD estimator and the NMD estimator with identity matrix (i.e. the FE estimator) in the pseudo panel case with instru-

[5]Note that we do not need the formula above to calculate $\hat{\mu}_{gt}^{zp\check{u}}\hat{\mu}_{gt}^{zq\check{u}}$; $\hat{\mu}_{gt}^{zp\check{u}}\hat{\mu}_{gt}^{zq\check{u}}$ can be obtained from calculating the cohort-level residuals. The formula is to provide an insight of its relationship to $n_{gt}^{-1}\hat{\tau}_{pq}$.

ments. The major purposes of this simulation study are (i) to illustrate that the formulae derived in the last section work when the model are correctly specified, and (ii) to show that valid instruments improve estimation efficiency. We also look at naive ways of computing the standard errors that only make use of the cohort level data. Their performance is compared to the NMD standard errors, and explanations for the difference are provided.

2.4.1 Simulation design

Throughout the simulation study, the outcome y_{it} is generated as a linear function of the covariates $(x_{1it} = 1, x_{2it}, x_{3it}, x_{4it})$, the time effect η_t, the individual effect f_i, and the idiosyncratic error u_{it}:

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_4 x_{4it} + \eta_t + f_i + u_{it},\ i = 1, \cdots, N_t,\ t = 1, \cdots, T. \quad (2.49)$$

The parameter values used are $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) = (1, 1, 1, 1)$. The time effects are generated by $\eta_t = t - 1$, and the cohort effects are generated by $\alpha_g = g - 1$. Individual fixed effects are generated by adding a random normal disturbance to the cohort effects, i.e. $f_i \sim N(\alpha_g, 1)$. To fix ideas, it might be helpful to think of x_{2it}, x_{3it} and x_{4it} as education, experience and marital status, respectively. The outcome y_{it} is the log hourly wage, and there is an individual effect f_i representing some unobserved ability.

We focus on estimating the coefficient of x_{2it} for which the distribution is given later. The distributions of the two auxiliary variables x_{3it} and x_{4it} are given by

$$\begin{aligned} x_{3it} &\sim N(\sin(gt), 1), \\ x_{4it} &\sim Bernouli\left(\frac{1}{1 + exp[1.5\sin(gt/2)]}\right). \end{aligned}$$

That is, x_{3it} is a continuous variable with population cohort mean $\sin(gt)$ and within cell variance 1; x_{4it} is a binary variable equal to 1 with probability $\frac{1}{1+exp(1.5*\sin(gt/2))}$. Since the individual-level disturbance to x_{3it} and x_{4it} are independently generated, they are always valid IVs. A time-invariant external instrument is generated as $z_i \sim N(0, 1)$.

We investigate a small pseudo panel ($G = 6$, $T = 4$) and a middle sized one ($G = 30$, $T = 20$). In the small pseudo panel, the idiosyncratic error u_{it} follows $N(0, 10)$, and the following 5 cases for x_{2it} are considered

1. $x_{2it} \sim N(gt/6, 1)$,
2. $x_{2it} \sim N(gt/6, 1) + f_i$,
3. $x_{2it} \sim N(gt/6, 1) + z_i$,
4. $x_{2it} \sim N(gt/6, 1) + z_i + f_i$,
5. $x_{2it} \sim N(gt/2, 1) + z_i + f_i$.

The standard deviation for μ_{gt}^{x2} over (g, t) is about 1. Note that x_{2it} is a valid IV in cases 1 through 3, but not valid in cases 4 and 5.

In the middle sized pseudo panel, u_{it} follows $N(0, 100)$ which has a bigger variance than in the small pseudo panel. The five cases considered for x_{2it} are

1. $x_{2it} \sim N(gt/150, 1)$,
2. $x_{2it} \sim N(gt/150, 1) + f_i$,
3. $x_{2it} \sim N(gt/150, 1) + z_i$,
4. $x_{2it} \sim N(gt/150, 1) + z_i + f_i$,
5. $x_{2it} \sim N(gt/50, 1) + z_i + f_i$,

The standard deviation for μ_{gt}^{x2} over (g, t) is about 23. The variance-covariance as well as correlation matrix of $(\mu_{gt}^{x2}, \mu_{gt}^{x3}, \mu_{gt}^{x4})$ are given in Table 2.1.

Case 4 in each setup is the case of major interest. Cases 1 through 3 are used to isolate the effect of adding f_i or z_i as part of x_{2it}. Case 5 checks the effect of a larger variation in the cohort mean of x_{2it}.

Table 2.1 Variance-covariance and correlation matrix of $(\mu_{gt}^{x2}, \mu_{gt}^{x3}, \mu_{gt}^{x4})$; correlation coefficients in parentheses. $\mu_{gt}^{x3} = \sin(gt)$, $\mu_{gt}^{x4} = (1 + exp[1.5 * \sin(gt/2)])^{-1}$.

	$G = 6,\ T = 4;$ $\mu_{gt}^{x2} = gt/6$				$G = 30,\ T = 20;$ $\mu_{gt}^{x2} = gt/150$		
	μ_{gt}^{x2}	μ_{gt}^{x3}	μ_{gt}^{x4}		μ_{gt}^{x2}	μ_{gt}^{x3}	μ_{gt}^{x4}
μ_{gt}^{x2}	1.078			μ_{gt}^{x2}	0.834		
	(1)				(1)		
μ_{gt}^{x3}	-0.142	0.488		μ_{gt}^{x3}	-0.013	0.507	
	(-0.195)	(1)			(-0.020)	(1)	
μ_{gt}^{x4}	0.107	0.018	0.054	μ_{gt}^{x4}	0.009	-0.001	0.056
	(0.444)	(0.109)	(1)		(0.040)	(-0.006)	(1)
	$G = 6,\ T = 4;$ $\mu_{gt}^{x2} = gt/2$				$G = 30,\ T = 20;$ $\mu_{gt}^{x2} = gt/50$		
	μ_{gt}^{x2}	μ_{gt}^{x3}	μ_{gt}^{x4}		μ_{gt}^{x2}	μ_{gt}^{x3}	μ_{gt}^{x4}
μ_{gt}^{x2}	9.701			μ_{gt}^{x2}	7.508		
	(1)				(1)		
μ_{gt}^{x3}	-0.425	0.488		μ_{gt}^{x3}	-0.039	0.507	
	(-0.195)	(1)			(-0.020)	(1)	
μ_{gt}^{x4}	0.322	0.018	0.054	μ_{gt}^{x4}	0.026	-0.001	0.056
	(0.444)	(0.109)	(1)		(0.040)	(-0.006)	(1)

In the ideal situation, we would like to draw a population of size infinity so that the cohort level population equations hold exactly. Take case 4 as an example, that would mean the following set of equations holds exactly

$$E(y_{it}|g) = \beta_1 + \beta_2 \frac{gt}{150} + \beta_3 \sin(gt) + \beta_4 \frac{1}{1 + exp[1.5 \sin(gt/2)]} + (t-1) + (g-1), \quad (2.50)$$
$$g = 1, \cdots, G;\ t = 1, \cdots, T.$$

But drawing an infinite number of observations is obviously infeasible. Therefore, we choose a relatively large number as the population size. The true distribution of the resulting population is of course its empirical distribution, but we could think of it as an approximation

of the population defined by eq. (2.49). Eq. (2.50) also holds only approximately, but the difference should be negligible for the purpose of this simulation study.

In the current setup, the population cohort sizes are set equally to $N_{gt} = 10^5$ for all g and t. That means a population panel of $N = 2.4 \times 10^6$ individual-time points for $G = 6$ and $T = 4$, and $N = 6 \times 10^7$ for $G = 30$ and $T = 20$.

After the population is generated, we fix it over simulations. In each replication, we draw repeated random cross sections from this fixed population. To have an idea on how the sample size affects the estimates, we consider two different sampling rates, 0.2% and 1%, which result in the sample cohort sizes $n_{gt} = 200$ and 1000, respectively.[6]

The simulation design above is careful in the two places emphasized by Imbens and Wooldridge (2007). First, data for each section is drawn from the population independently across time, and because of the random sampling in each period, the group identifier is also randomly drawn. Second, eq. (2.49) has full time effects which is more realistic than Verbeek and Vella (2005) that omits the aggregate time effects, for the variation in $\mu_{gt}^{\mathbf{x}}$ here is net of the time effects.

For $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\alpha})$ in eq. (2.49), we consider the NMD estimator with identity matrix ($\check{\boldsymbol{\theta}}$) and its standard error (s.e.), and the optimal NMD estimator ($\hat{\boldsymbol{\theta}}$) and its s.e.. Because $\check{\boldsymbol{\theta}}$ is the fixed effect estimator on the pseudo panel of sample cohort means, three naive s.e. estimators, namely the non-robust s.e., the heteroskedasticity-robust s.e. and the cluster-robust s.e., are also computed. They are the usual s.e. estimators routinely computed for the fixed effect estimator in a true panel, but are naive in a pseudo panel because they treat the sample cohort means as observations carrying no errors and completely ignore the individual-level data.

Besides the basic NMD uses no IV, each of z, x_2, x_3 and x_4 is used one at a time as the

[6]A relatively higher sampling rate might introduce too much overlap among the repeated cross-sectional samples. Therefore, we also consider the setup $N_t = 1.5 * 10^7$ with sampling rate 0.2%. The result shows that there is no essential difference from the setup $N_t = 3 * 10^3$ with sampling rate 1%.

additional IV. The NMD using all 4 variables as the additional IVs is also estimated.

2.4.2 Simulation results for the small pseudo panel

At the center of the simulation study is Case 4, as cases 1 through 3 are its simplified cases to pin down the effect of the correlation of x_{2it} with z_i and f_i, and Case 5 is a variation of Case 4 that increases the variation in the cohort mean of x_2. Therefore we focus on analyzing Case 4 in this section. The Monte Carlo simulation results for case 4 from 1000 replications for the coefficient and s.e. estimators of x_2 are presented in Table 2.3. Two sample cohort sizes, $n_{gt} = 200$ and 1000, are considered. For each considered quantity, the Monte Carlo average and standard deviation over the 1000 replications are reported, with the standard deviation in parentheses. The estimators with no IV, z as IV, and x_2 as IV are picked because they provide most of the insights. The results on the same quantities in Case 3 are reported in Table 2.2. Detailed results are in Tables B.1 through B.20 in Appendix B.

Several observations stand out from Table 2.3. First of all, the NMD coefficient estimators work well in all cases except when the invalid IV x_2 is used. Both $\check{\beta}_2$ and $\hat{\beta}_2$ are close to the true value in columns 1, 2, 4 and 5. As the sample cohort size n_{gt} gets bigger, the slight biases in $\check{\beta}_2$ and $\hat{\beta}_2$ get even smaller, and their Monte Carlo standard deviations also shrink. Second, the NMD s.e. estimators also work well, even when x_2 is used as the IV. The Monte Carlo averages of $\widehat{se(\check{\beta}_2)}$ and $\widehat{se(\hat{\beta}_2)}$ are close to the standard deviations of $\check{\beta}_2$ and $\hat{\beta}_2$ throughout all columns, and having a bigger cohort size, as expected, reduces $\widehat{se(\check{\beta}_2)}$ and $\widehat{se(\hat{\beta}_2)}$ universally. Third, using a valid and relevant IV improves efficiency, but the validity of IV is crucial. Compared to using no IV (column 1 and 4), using z as IV (columns 2 and 5) leads to reduced Monte Carlo averages of $\widehat{se(\check{\beta}_2)}$ and $\widehat{se(\hat{\beta}_2)}$ and smaller finite sample bias in $\check{\beta}_2$ and $\hat{\beta}_2$. The usage of the invalid IV x_2 (columns 3 and 6), however, introduces persistent biases in $\check{\beta}_2$ and $\hat{\beta}_2$ that do not vanish as cohort size gets larger. Note that x_2 is not a valid IV because it is correlated with f_i, which violates the condition in (2.12).

A comparison with the results in Table 2.2 confirms that the correlation between x_{2it} and

Table 2.2 Finite sample properties of various estimators of β_2 and its standard error, $G = 6$, $T = 4$. Case 3. $x_{2it} \sim N(gt/6, 1) + z_i$

	$n_{gt} = 200$			$n_{gt} = 1000$		
	none	z	x_2	none	z	x_2
$\check{\beta}_2$	.9909	.9980	.9965	.9937	.9982	.9964
	(.1623)	(.0430)	(.0421)	(.0714)	(.0191)	(.0191)
$\widehat{se(\check{\beta}_2)}$	.1590	.0463	.0436	.0720	.0206	.0192
	(.0117)	(.0012)	(.0020)	(.0023)	(.0002)	(.0004)
$\widehat{se_n(\check{\beta}_2)}$	.1552	.0457	.0488	.0698	.0205	.0218
	(.0349)	(.0054)	(.0092)	(.0144)	(.0025)	(.0039)
$\widehat{se_r(\check{\beta}_2)}$	.1393	.0508	.0490	.0627	.0229	.0217
	(.0425)	(.0073)	(.0094)	(.0181)	(.0034)	(.0040)
$\widehat{se_c(\check{\beta}_2)}$	.1423	.0496	.0431	.0639	.0223	.0189
	(.0603)	(.0159)	(.0154)	(.0283)	(.0073)	(.0069)
$\hat{\beta}_2$	.9911	.9979	.9981	.9936	.9982	.9987
	(.1623)	(.0437)	(.0324)	(.0713)	(.0191)	(.0143)
$\widehat{se(\hat{\beta}_2)}$	.1585	.0452	.0327	.0720	.0205	.0148
	(.0117)	(.0012)	(.0007)	(.0023)	(.0002)	(.0001)

f_i invalidates x_{2it} as IV. In absence of the correlation between x_{2it} and f_i, x_{2it} is exogenous and becomes a valid IV for itself. As a result, no obvious bias is observed in $\check{\beta}_2$ and $\hat{\beta}_2$ when x_2 is used as IV in Table 2.2. In effect, x_2 is a better IV than z, since Table 2.2 shows that $\widehat{se(\check{\beta}_2)}$ and $\widehat{se(\hat{\beta}_2)}$ become smaller on average when the IV is changed from z to x_2. This makes sense because no IV is more relevant to a variable than the variable itself.

When the IV is changed from z to x_2 in Table 2.2, a larger reduction is observed in $\widehat{se(\hat{\beta}_2)}$ than in $\widehat{se(\check{\beta}_2)}$. This observation highlights a typical situation to use the optimal weighting matrix - when the IV brings in within-cell heteroskedasticity and correlation. Specifically, when z is the IV, we show in the Appendix that

$$\mathbf{\Xi}_{gt} = \sigma_e^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{2.51}$$

where $\sigma_e^2 = E\left[(\varepsilon_i^f + u_{it})^2 | g\right]$. This implies that the optimal weighting matrix is proportional

Table 2.3 Finite sample properties of various estimators of β_2 and its standard error, $G = 6$, $T = 4$. Case 4. $x_{2it} \sim N(gt/6, 1) + z_i + f_i$

	$n_{gt} = 200$			$n_{gt} = 1000$		
	none	z	x_2	none	z	x_2
$\check{\beta}_2$	1.0153	1.0048	1.2218	.9989	.9996	1.2166
	(.1599)	(.0431)	(.0947)	(.0716)	(.0191)	(.0423)
$\widehat{se(\check{\beta}_2)}$	.1575	.0462	.0951	.0719	.0206	.0421
	(.0143)	(.0014)	(.0071)	(.0029)	(.0003)	(.0014)
$\widehat{se_n(\check{\beta}_2)}$	.1537	.0455	.0728	.0697	.0205	.0405
	(.0356)	(.0054)	(.0188)	(.0145)	(.0025)	(.0082)
$\widehat{se_r(\check{\beta}_2)}$	.1388	.0506	.0752	.0627	.0229	.0458
	(.0427)	(.0074)	(.0191)	(.0181)	(.0034)	(.0106)
$\widehat{se_c(\check{\beta}_2)}$	.1440	.0494	.0947	.0642	.0223	.0711
	(.0631)	(.0159)	(.0420)	(.0284)	(.0073)	(.0254)
$\hat{\beta}_2$	1.0155	1.0048	1.3194	.9989	.9996	1.3220
	(.1598)	(.0437)	(.0266)	(.0715)	(.0191)	(.0120)
$\widehat{se(\hat{\beta}_2)}$	.1569	.0451	.0266	.0719	.0205	.0120
	(.0142)	(.0014)	(.0006)	(.0029)	(.0003)	(.0001)

to an identity matrix, which explains why the averages of $\widehat{se(\hat{\beta}_2)}$ and $\widehat{se(\check{\beta}_2)}$ are close to each other and to the standard deviations of $\check{\beta}_2$ and $\hat{\beta}_2$ in Table 2.2 when z is the IV. On the other hand, when x_2 is used as the IV, we show that

$$\mathbf{\Xi}_{gt} = \sigma_e^2 \begin{pmatrix} 1 & \frac{gt}{150} \\ \frac{gt}{150} & 2 + \left(\frac{gt}{150}\right)^2 \end{pmatrix}, \tag{2.52}$$

which has within-cell heteroskedasticity and correlation. The resulting optimal weighting matrix is distinct from an identity matrix.

The results on the naive s.e. estimators, $\widehat{se_n(\check{\beta}_2)}$, $\widehat{se_r(\check{\beta}_2)}$ and $\widehat{se_c(\check{\beta}_2)}$ are also consistent with the theory. We leave the discussion to the next subsection because the pattern is more obvious when G is greater.

Table 2.4 Finite sample properties of various estimators of β_2 and its standard error, $G = 30$, $T = 20$. Case 3. $x_{2it} \sim N(gt/150, 1) + z_i$

	$n_{gt} = 200$			$n_{gt} = 1000$		
	none	z	x_2	none	z	x_2
$\check{\beta}_2$	1.0134	1.0022	1.0012	.9980	.9996	.9996
	(.0839)	(.0277)	(.0248)	(.0399)	(.0121)	(.0106)
$\widehat{se(\check{\beta}_2)}$	.0842	.0277	.0239	.0387	.0123	.0105
	(.0011)	(.0001)	(.0002)	(.0002)	(.0000)	(.0000)
$\widehat{se_n(\check{\beta}_2)}$	.0842	.0274	.0286	.0388	.0123	.0129
	(.0028)	(.0006)	(.0009)	(.0012)	(.0003)	(.0004)
$\widehat{se_r(\check{\beta}_2)}$	.0838	.0281	.0269	.0387	.0125	.0119
	(.0045)	(.0008)	(.0009)	(.0021)	(.0003)	(.0004)
$\widehat{se_c(\check{\beta}_2)}$	.0841	.0282	.0241	.0385	.0124	.0105
	(.0143)	(.0037)	(.0035)	(.0067)	(.0016)	(.0016)
$\hat{\beta}_2$	1.0134	1.0020	1.0016	.9979	.9995	.9993
	(.0843)	(.0279)	(.0209)	(.0400)	(.0121)	(.0087)
$\widehat{se(\hat{\beta}_2)}$	.0837	.0269	.0196	.0387	.0123	.0089
	(.0011)	(.0001)	(.0001)	(.0002)	(.0000)	(.0000)

2.4.3 Simulation results for the middle sized pseudo panel

The results in Table 2.5 and 2.4 for the middle sized pseudo panel basically tell the same story as the small pseudo panel. We focus on two points that stand out. These two points are less clear, although also exist, in the small pseudo panel.

First, the results on the naive s.e. estimators, $\widehat{se_n(\check{\beta}_2)}$, $\widehat{se_r(\check{\beta}_2)}$ and $\widehat{se_c(\check{\beta}_2)}$ are consistent with the theory. This is best seen from the last column in Table 2.4. Moving down the list $\widehat{se_n(\check{\beta}_2)}$, $\widehat{se_r(\check{\beta}_2)}$ and $\widehat{se_c(\check{\beta}_2)}$, the bias in the Monte Carlo averages gradually declines. The Monte Carlo average of $\widehat{se_c(\check{\beta}_2)}$ rounded four decimal places is even identical to that of $\widehat{se(\check{\beta}_2)}$. The reason is that, when x_2 is used as IV,

$$Var[\mu\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'e}|g] = diag(n_{g1}^{-1}\Xi_{g1}, \cdots, n_{gT}^{-1}\boldsymbol{\Xi}_{gT})$$

is indeed block diagonal by eq. (2.52). Among the three naive variance estimators, only the

Table 2.5 Finite sample properties of various estimators of β_2 and its standard error, $G = 30$, $T = 20$. Case 4. $x_{2it} \sim N(gt/150, 1) + z_i + f_i$

	$n_{gt} = 200$			$n_{gt} = 1000$		
	none	z	x_2	none	z	x_2
$\check{\beta}_2$	1.0496	1.0106	1.1137	1.0061	1.0013	1.0285
	(.0830)	(.0276)	(.1376)	(.0395)	(.0120)	(.0642)
$\widehat{se(\check{\beta}_2)}$	.0826	.0276	.1339	.0386	.0123	.0633
	(.0012)	(.0002)	(.0026)	(.0003)	(.0000)	(.0006)
$\widehat{se_n(\check{\beta}_2)}$	.0826	.0273	.0783	.0387	.0123	.0382
	(.0028)	(.0006)	(.0037)	(.0012)	(.0003)	(.0016)
$\widehat{se_r(\check{\beta}_2)}$	.0822	.0280	.1249	.0385	.0125	.0594
	(.0045)	(.0008)	(.0111)	(.0020)	(.0003)	(.0051)
$\widehat{se_c(\check{\beta}_2)}$	.0824	.0281	.1222	.0384	.0124	.0585
	(.0139)	(.0037)	(.0302)	(.0066)	(.0016)	(.0151)
$\hat{\beta}_2$	1.0493	1.0104	1.3186	1.0060	1.0013	1.3199
	(.0835)	(.0279)	(.0175)	(.0395)	(.0120)	(.0071)
$\widehat{se(\hat{\beta}_2)}$	.0821	.0268	.0162	.0385	.0122	.0073
	(.0012)	(.0002)	(.0001)	(.0003)	(.0000)	(.0000)

cluster-robust version correctly accounts for the variance-covariance structure of the cohort composite error. The heteroskedasticity-robust version only captures the heteroskedasticity but not the within-cluster correlation. The non-robust version accounts for neither.

As a comparison, in the 4th and 5th columns of Table 2.4, the Monte Carlo averages of $\widehat{se_n(\check{\beta}_2)}$, $\widehat{se_r(\check{\beta}_2)}$ and $\widehat{se_c(\check{\beta}_2)}$ are all close to that of $\widehat{se(\check{\beta}_2)}$. This is because $Var[\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'e}|g]$ is proportional to an identity matrix when none or z is used as IV. As a result, all three versions of the naive variance estimators are correct in their modeling of $Var[\hat{\boldsymbol{\mu}}_g^{\mathbf{z}'e}|g]$. Moreover, from $\widehat{se_n(\check{\beta}_2)}$ through $\widehat{se_r(\check{\beta}_2)}$ to $\widehat{se_c(\check{\beta}_2)}$, the s.e. estimators become less and less efficient, indicated by greater and greater Monte Carlo standard deviations. This is also consistent with their well-known relative efficiency property. Of course, $\widehat{se(\check{\beta}_2)}$ is much more efficient than any of the naive estimators, for $\widehat{se(\check{\beta}_2)}$ makes use of the extra information from the individual-level data.

Secondly, the first column in Table 2.5 shows noticeable bias in $\check{\beta}_2$ and $\hat{\beta}_2$. It is finite sample bias because as n_{gt} gets larger, the bias shrinks quickly. A comparison with the first column in Table 2.4 confirms that the correlation between x_{2it} and f_i contributes to a large part of the bias.

2.5 Concluding remarks

This chapter develops a general NMD framework that imposes (partial) differentiability on the structural equations. The differentiability conditions are stronger than the MD framework in Newey and McFadden (1994), but the resulting framework is more convenient to work with in application. Consistency and asymptotic normality are established, as well as the optimal weighting matrix expressed as functions of the partial derivatives of the structural equations. A theorem that echoes the GMM property that more moment conditions do not hurt is given. The general framework is then applied to the special case of pseudo panel. Simulation results are consistent with the theory.

The property that having more moment conditions could improve efficiency is first noticed in the exercise of adding external instruments in pseudo panel MD estimation, which is an extension to the work on pseudo panel by Imbens and Wooldridge (2007). We would like to establish this property in a more general setup, hence the NMD framework at the beginning of this chapter is motivated. Having both the general framework and the case of pseudo panel as an example in the same chapter helps the understanding of the general concepts. In particular, we find that the inverse optimal weighting matrix is exactly the variance-covariance of the relocated and rescaled structural equations in the pseudo panel application, which provides straightforward intuition for why the optimal weighting matrix works. Essentially, the optimal weighting matrix down-weights the structural equations that are volatile and give more weights to those that are less volatile, and correlation between structural equations are also accounted for.

The NMD estimation in pseudo panel correctly relies on large n_{gt} but fixed G, T asymptotics. Naive methods like FE on the cohort means are found to be the NMD estimators using the identity weighting matrix. But the naive s.e. estimators, including the usual s.e., the s.e. robust to heteroskedasticity and the cluster-robust s.e., rely on at least large G asymptotics. Even when G is moderate or large, depending on how complicated the IVs are, the usual s.e. and the one only robust to heteroskedasticity may not capture the correct variance-covariance structure. The cluster-robust s.e. though has the potential to work for large G because it is fully robust. But since it ignores the individual level data completely, it is always less efficient than the NMD s.e. estimator using identity weighting matrix. The optimal NMD is always the most efficient among these candidates. we conclude that when there are extra IVs to explore, it is important to use optimal weighting.

The comments in Imbens and Wooldridge (2007) regarding flexible specifications provide several ideas we would like to investigate in future research. First, we intend to extend the application to dynamic models, i.e., to add lagged dependent variables in the list of explanatory variables. This is an issue that has been studied by Moffitt (1993); Collado (1997); Girma (2000); Verbeek and Vella (2005); McKenzie (2004) among others. In dynamic models, the advantage of having the general NMD framework stands out. There is no need to tailor the framework in any way, since we can still define the vector of reduced form parameters as before. Because cohort means of the dependent variable do not appear redundantly in the reduce-form parameters, their asymptotics are well defined. The cohort-level equations are also of the same form as before; the only difference is that the equations for the first several periods need to be dropped because of the lags. Second, we intend to add unit-specific trend in the unobserved heterogeneity as in the random growth model of Heckman and Hotz (1989). Third, an even more flexible extension is to let the factor loads on the unobserved heterogeneity be time-varying. These extensions should be easily handled by the NMD framework. Lastly, we are also interested in an empirical application of the method. Currently, we am working on applying the pseudo panel method to estimate returns

to education using data from the U.S. Current Population Survey.

APPENDICES

APPENDIX A

PROOFS AND ALGEBRA

A.1 Proof of consistency

Proof. Prove by verifying verifying (i)-(iv) of Theorem 2.1 in Newey and McFadden (1994) □

A.2 Proof of asymptotic normality

A sketch of the idea first. By the first part of (ii), a mean value expansion of each component of $\mathbf{h}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ leads to $\mathbf{h}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) = \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) + \mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ where $\bar{\boldsymbol{\theta}}$ is a vector of mean values. A similar expansion, $\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) = \mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$, follows by the second part of (ii) and $\mathbf{h}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = 0$. Substituting the two expansions in the first-order condition and solving gives $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -[\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})'\hat{\mathbf{W}}\mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})]^{-1} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})'\hat{\mathbf{W}}\mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) \cdot \sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$. By the first part of (iv), $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and continuity of $\mathbf{L}(\boldsymbol{\pi}, \boldsymbol{\theta})$ on $\mathscr{N}(\boldsymbol{\theta}_0)$ in (ii), we have that, with probability approaching one, $\left\|\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) - \mathbf{L}\right\| \leq \left\|\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) - \mathbf{L}(\boldsymbol{\pi}_0, \hat{\boldsymbol{\theta}})\right\| + \left\|\mathbf{L}(\boldsymbol{\pi}_0, \hat{\boldsymbol{\theta}}) - \mathbf{L}\right\| \leq sup_{\boldsymbol{\theta} \in \mathscr{N}(\boldsymbol{\theta}_0)} \|\mathbf{L}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}) - \mathbf{L}(\boldsymbol{\pi}_0, \boldsymbol{\theta})\| + \left\|\mathbf{L}(\boldsymbol{\pi}_0, \hat{\boldsymbol{\theta}}) - \mathbf{L}\right\| \xrightarrow{p} 0$. Similar, the convergence $\left\|\mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}}) - \mathbf{L}\right\| \xrightarrow{p} 0$ follows by $\bar{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, and $\|\mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) - \mathbf{B}\| \xrightarrow{p} 0$ by $\bar{\boldsymbol{\pi}} \xrightarrow{p} \boldsymbol{\pi}_0$ and continuity of $\mathbf{B}(\boldsymbol{\pi}, \boldsymbol{\theta})$ on $\mathscr{N}(\boldsymbol{\theta}_0)$. Condition (iv) guarantees that $\left(\mathbf{L}'\mathbf{W}\mathbf{L}\right)^{-1}$ exists, and thus, with probability approaching one, the existence of $[\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})'\hat{\mathbf{W}}\mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})]^{-1}$. The conclusion then follows by (iii), Slutsky's theorem and the asymptotic equivalence theorem.

A full proof with technical details is given below.

Proof. By (i), without of generality $\mathscr{N}(\boldsymbol{\theta}_0)$ ($\mathscr{N}(\boldsymbol{\pi}_0)$) can be assumed to be a convex, open set contained in $\boldsymbol{\Theta}$ ($\boldsymbol{\Pi}$). Then $\mathscr{N}(\boldsymbol{\theta}_0)$ ($\mathscr{N}(\boldsymbol{\pi}_0)$) is also connected since $\boldsymbol{\Theta} \in \mathbb{R}^P$ ($\boldsymbol{\Pi} \in \mathbb{R}^K$).

Let 1_A denote the indicator function for an event A. Let $A_1 = \{\hat{\boldsymbol{\theta}} \in \mathscr{N}(\boldsymbol{\theta}_0)\}$ and $A_2 = \{\hat{\boldsymbol{\pi}} \in \mathscr{N}(\boldsymbol{\pi}_0)\}$, and . Note that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ ($\hat{\boldsymbol{\pi}} \xrightarrow{p} \boldsymbol{\pi}_0$) implies $1_{A_1} \xrightarrow{p} 1$ ($1_{A_2} \xrightarrow{p} 1$).

(1) By the first part of condition (ii) and the first order condition for a minimum, $1_{A_1} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{h}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) = \mathbf{0}$. The multiplication by 1_{A_1} is needed because by (ii) $\mathbf{L}(\boldsymbol{\pi}, \boldsymbol{\theta})$ only exists on $\mathscr{N}(\boldsymbol{\theta}_0)$. $\hat{\boldsymbol{\theta}}$ is pointwise defined by $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{argmin}\ \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})' \hat{\mathbf{W}}\, \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta})$, not by the first order condition $\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{h}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) = \mathbf{0}$. For some realization of $\hat{\boldsymbol{\pi}}$, the corresponding $\hat{\boldsymbol{\theta}}$ may not lie in $\mathscr{N}(\boldsymbol{\theta}_0)$.

(2) Since $\mathscr{N}(\boldsymbol{\theta}_0)$ is connected, by condition (ii) and mean value expansion theorem, $1_{A_1} \cdot h_j(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) = 1_{A_1} \cdot h_j(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) + 1_{A_1} \cdot \mathbf{L}_j(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}}_j)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ for $j = 1, \cdots, J$, where $\bar{\boldsymbol{\theta}}_j$ is a random variable equal to the mean value if $1_{A_1} = 1$ and equal to $\boldsymbol{\theta}_0$ otherwise. Again, this complication is needed because $\hat{\boldsymbol{\theta}}$ is not necessarily in $\mathscr{N}(\boldsymbol{\theta}_0)$. Clearly, $\bar{\boldsymbol{\theta}}_j \xrightarrow{p} \boldsymbol{\theta}_0$ as $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Collect all the J mean values in the matrix $\bar{\boldsymbol{\theta}}$, and let $\mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})$ be the matrix with j-th row $\mathbf{L}_j(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}}_j)$. Then those expansions can be written collectively as $1_{A_1} \cdot \mathbf{h}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) = 1_{A_1} \cdot \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) + 1_{A_1} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Substituting in $\mathbf{0} = 1_{A_1} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{h}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})$ leads to $\mathbf{0} = 1_{A_1} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) + 1_{A_1} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

(3) By a similar reasoning and the fact $\mathbf{h}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = 0$, write $1_{A_2} \cdot \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) = 1_{A_2} \cdot \mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$, where $\bar{\boldsymbol{\pi}}_j$, the j-th column of the matrix $\bar{\boldsymbol{\pi}}$, equals to a mean value the if $1_{A_2} = 1$ and equal to $\boldsymbol{\pi}_0$ otherwise, and $\mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)$ is the matrix with the j-th row $\mathbf{B}_j(\bar{\boldsymbol{\pi}}_j, \boldsymbol{\theta}_0)$. $\bar{\boldsymbol{\pi}}_j \xrightarrow{p} \boldsymbol{\pi}_0$ as $\bar{\boldsymbol{\pi}}_j \xrightarrow{p} \boldsymbol{\pi}_0$. Also, $1_{A_2} \xrightarrow{p} 1$, and $\mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{B}$ by the second part of condition (iv). Substituting again gives $\mathbf{0} = 1_{A_1 \cap A_2} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) \mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) + 1_{A_1 \cap A_2} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

(4) Let $A_3 = \{\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})$ *is nonsingular*$\}$. Let $\bar{\mathbf{V}}$ be a random variable equal to $\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}})$ if $1_{A_3} = 1$ and equal to the K-dimensional identity matrix otherwise. By the first part of condition (iv), $\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{L}$ and $\mathbf{L}(\hat{\boldsymbol{\pi}}, \bar{\boldsymbol{\theta}}) \xrightarrow{p} \mathbf{L}$. Then by condition (v) and $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, $1_{A_3} \xrightarrow{p} 1$ and $\bar{\mathbf{V}} \xrightarrow{p} \mathbf{L}'\mathbf{W}\mathbf{L}$. Substituting for another time gives $\mathbf{0} = 1_{A_1 \cap A_2 \cap A_3} \cdot \mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) \mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) + 1_{A_1 \cap A_2 \cap A_3} \cdot \bar{\mathbf{V}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

Now, let $A_0 = A_1 \cap A_2 \cap A_3$. Note that $1_{A_0} = 1_{A_1} \cdot 1_{A_2} \cdot 1_{A_3} \xrightarrow{p} 1$. Multiplying by $\sqrt{n}$ and

solving gives $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -1_{A_0} \cdot \bar{\mathbf{V}}^{-1}\mathbf{L}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})'\hat{\mathbf{W}}\mathbf{h}(\hat{\boldsymbol{\pi}}, \boldsymbol{\theta}_0)\mathbf{B}(\bar{\boldsymbol{\pi}}, \boldsymbol{\theta}_0) \cdot \sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) - (1_{A_0} - 1)\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. The conclusion follows by Slutsky's theorem and the asymptotic equivalence theorem. □

A.3 Proof of optimal weighting matrix

Proof. For an arbitrary $\mathbf{W}$,

$$\left(\mathbf{L}'\mathbf{W}\mathbf{L}\right)^{-1}\mathbf{L}'\mathbf{W}\mathbf{B}\boldsymbol{\Omega}\mathbf{B}'\mathbf{W}\mathbf{L}\left(\mathbf{L}'\mathbf{W}\mathbf{L}\right)^{-1} - (\mathbf{L}'(\mathbf{B}\boldsymbol{\Omega}\mathbf{B}')^{-1}\mathbf{L})^{-1} = \mathbf{D}'(\mathbf{B}\boldsymbol{\Omega}\mathbf{B}')^{-1}\mathbf{D}$$

is positive semi-definite, where

$$\mathbf{D} = \left(\mathbf{L}'\mathbf{W}\mathbf{L}\right)^{-1}\mathbf{L}'\mathbf{W} - [\mathbf{L}'\left(\mathbf{L}'\mathbf{W}\mathbf{L}\right)^{-1}\mathbf{L}]^{-1}\mathbf{L}'(\mathbf{B}\boldsymbol{\Omega}\mathbf{B}')^{-1}.$$

□

A.4 Proof that extra conditions do not hurt

Proof. Notice that

$$\mathbf{B} = \nabla_{\boldsymbol{\pi}}\mathbf{h}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = \begin{bmatrix} \nabla_{\boldsymbol{\pi}}\mathbf{h}_1(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \\ \nabla_{\boldsymbol{\pi}}\mathbf{h}_2(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}.$$

Then

$$\mathbf{M} = \mathbf{B}\boldsymbol{\Omega}\mathbf{B}' = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \boldsymbol{\Omega} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}'$$
$$= \begin{bmatrix} \mathbf{B}_1\boldsymbol{\Omega}\mathbf{B}_1' & \mathbf{B}_1\boldsymbol{\Omega}\mathbf{B}_2' \\ \mathbf{B}_2\boldsymbol{\Omega}\mathbf{B}_1' & \mathbf{B}_2\boldsymbol{\Omega}\mathbf{B}_2' \end{bmatrix} = \begin{bmatrix} \mathbf{M}_{1,1} & \mathbf{M}_{1,2} \\ \mathbf{M}_{2,1} & \mathbf{M}_{2,2} \end{bmatrix}.$$

Also notice that

$$\mathbf{L} = \nabla_{\boldsymbol{\theta}}\mathbf{h}(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = \begin{bmatrix} \nabla_{\boldsymbol{\theta}}\mathbf{h}_1(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \\ \nabla_{\boldsymbol{\theta}}\mathbf{h}_2(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) \end{bmatrix} = \begin{bmatrix} \mathbf{L}_1 \\ \mathbf{L}_2 \end{bmatrix}.$$

Now, define $\mathbf{F} = \mathbf{M}_{2,2} - \mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}$. $\mathbf{F}$ is the Schur complement of $\mathbf{M}_{1,1}$ in $\mathbf{M}$. Then

$$
\begin{aligned}
&\mathbf{L}'\mathbf{M}^{-1}\mathbf{L} \\
=&\begin{bmatrix}\mathbf{L}_1\\ \mathbf{L}_2\end{bmatrix}' \begin{bmatrix}\mathbf{M}_{1,1} & \mathbf{M}_{1,2}\\ \mathbf{M}_{2,1} & \mathbf{M}_{2,2}\end{bmatrix}^{-1} \begin{bmatrix}\mathbf{L}_1\\ \mathbf{L}_2\end{bmatrix} \\
=&\begin{bmatrix}\mathbf{L}_1\\ \mathbf{L}_2\end{bmatrix}' \begin{bmatrix}\mathbf{M}_{1,1}^{-1}+\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1} & -\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\\ -\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1} & \mathbf{F}^{-1}\end{bmatrix} \begin{bmatrix}\mathbf{L}_1\\ \mathbf{L}_2\end{bmatrix} \\
=&\Big[\mathbf{L}_1'\left(\mathbf{M}_{1,1}^{-1}+\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\right) - \mathbf{L}_2'\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1},\ ... \\
&... -\mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}+\mathbf{L}_2'\mathbf{F}^{-1}\Big] \begin{bmatrix}\mathbf{L}_1\\ \mathbf{L}_2\end{bmatrix} \\
=&\mathbf{L}_1'\left(\mathbf{M}_{1,1}^{-1}+\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\right)\mathbf{L}_1 - \mathbf{L}_2'\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 \\
&- \mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\mathbf{L}_2 + \mathbf{L}_2'\mathbf{F}^{-1}\mathbf{L}_2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbf{L}'\mathbf{M}^{-1}\mathbf{L} - \mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 \\
=&\mathbf{L}_1'\left(\mathbf{M}_{1,1}^{-1}+\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\right)\mathbf{L}_1 - \mathbf{L}_2'\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 \\
&- \mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\mathbf{L}_2 + \mathbf{L}_2'\mathbf{F}^{-1}\mathbf{L}_2 - \mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{L}_1. \\
=&\mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 - \mathbf{L}_2'\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 \\
&- \mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2}\mathbf{F}^{-1}\mathbf{L}_2 + \mathbf{L}_2'\mathbf{F}^{-1}\mathbf{L}_2 \\
=&\left(\mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2} - \mathbf{L}_2'\right)\mathbf{F}^{-1}\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 \\
&- \left(\mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2} - \mathbf{L}_2'\right)\mathbf{F}^{-1}\mathbf{L}_2 \\
=&\left(\mathbf{L}_1'\mathbf{M}_{1,1}^{-1}\mathbf{M}_{1,2} - \mathbf{L}_2'\right)\mathbf{F}^{-1}\left(\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 - \mathbf{L}_2\right) \\
=&\left(\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 - \mathbf{L}_2\right)'\mathbf{F}^{-1}\left(\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 - \mathbf{L}_2\right)
\end{aligned}
$$

where the fact $\mathbf{M}_{1,2} = \mathbf{M}_{2,1}'$ is used in the last equality. Clearly, the last expression is in quadratic form and thus is positive semi-definite. The condition for redundancy of

$\mathbf{h}_2(\boldsymbol{\pi}_0, \boldsymbol{\theta}_0) = 0$ is

$$\mathbf{M}_{2,1}\mathbf{M}_{1,1}^{-1}\mathbf{L}_1 - \mathbf{L}_2 = \mathbf{0},$$

or

$$\mathbf{B}_2\boldsymbol{\Omega}\mathbf{B}_1'(\mathbf{B}_1\boldsymbol{\Omega}\mathbf{B}_1')^{-1}\mathbf{L}_1 - \mathbf{L}_2 = 0.$$

□

A.5 Useful expressions for $\mathbf{s}_{it}$ and $\boldsymbol{\Omega}_{gt}^{\mathbf{s}}$

For $\mathbf{v}_{it} = (y_{it}, \mathbf{x}_{it})$ and $\mathbf{s}_{it} = \mathbf{z}_{it} \otimes \mathbf{v}_{it}$, an explicit expression for $\mathbf{s}_{it}$ is

$$\mathbf{s}_{it} = (\mathbf{v}_{it}, z_{2it}\mathbf{v}_{it}, \cdots, z_{Pit}\mathbf{v}_{it}).$$

Hence,

$$\boldsymbol{\Omega}_{gt}^{\mathbf{s}} = \begin{bmatrix} Var(\mathbf{v}_{it}|g) & Cov(\mathbf{v}_{it}, z_{2it}\mathbf{v}_{it}|g) & \cdots & Cov(\mathbf{v}_{it}, z_{Pit}\mathbf{v}_{it}|g) \\ Cov(z_{2it}\mathbf{v}_{it}, \mathbf{v}_{it}|g) & Var(z_{2it}\mathbf{v}_{it}|g) & \cdots & Cov(z_{2it}\mathbf{v}_{it}, z_{Pit}\mathbf{v}_{it}|g) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(z_{Pit}\mathbf{v}_{it}, \mathbf{v}_{it}|g) & Cov(z_{2it}\mathbf{v}_{it}, z_{2it}\mathbf{v}_{it}|g) & \cdots & Var(z_{Pit}\mathbf{v}_{it}|g) \end{bmatrix}.$$

For $\hat{\boldsymbol{\mu}}_{gt}^{z_p\mathbf{v}} = n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g} z_{pit}\mathbf{v}_{it}$, an estimator for $Cov(z_{pit}\mathbf{v}_{it}, z_{qit}\mathbf{v}_{it}|g)$ is

$$\hat{\Gamma}_{pq,gt} = n_{gt}^{-1}\sum_{i=1}^{n_t} r_{it,g}(z_{pit}\mathbf{v}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{z_p\mathbf{v}})'(z_{qit}\mathbf{v}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{z_q\mathbf{v}}).$$

It is also informative to write $\mathbf{s}_{it}$ as

$$\mathbf{s}_{it} = (y_{it}, \mathbf{x}_{it}, z_{2it}y_{it}, z_{2it}\mathbf{x}_{it}, \cdots, z_{Pit}y_{it}, z_{Pit}\mathbf{x}_{it}).$$

Because $\mathbf{x}_{it}$ includes unity, $\mathbf{s}_{it}$ contains all $y_{it}, \mathbf{x}_{it}, \mathbf{z}_{it}$ and the interactions of $\mathbf{z}_{it}$ with $(y_{it}, \mathbf{x}_{it})$.

A.6 Derivation for M

$$
\begin{aligned}
&\mathbf{M}\\
=&\mathbf{b}(\boldsymbol{\theta})\boldsymbol{\Omega}^{\mathbf{s}}\mathbf{b}(\boldsymbol{\theta})'\\
=&\begin{bmatrix}
\mathbf{b}_{11}(\boldsymbol{\theta}) & & & \\
& \mathbf{b}_{12}(\boldsymbol{\theta}) & & \\
& & \ddots & \\
& & & \mathbf{b}_{GT}(\boldsymbol{\theta})
\end{bmatrix}
\begin{bmatrix}
\frac{\boldsymbol{\Omega}^{\mathbf{s}}_{11}}{\rho_1\kappa_1} & & & \\
& \frac{\boldsymbol{\Omega}^{\mathbf{s}}_{12}}{\rho_1\kappa_2} & & \\
& & \ddots & \\
& & & \frac{\boldsymbol{\Omega}^{\mathbf{s}}_{GT}}{\rho_G\kappa_T}
\end{bmatrix}\cdot\\
&\begin{bmatrix}
\mathbf{b}_{11}(\boldsymbol{\theta}) & & & \\
& \mathbf{b}_{12}(\boldsymbol{\theta}) & & \\
& & \ddots & \\
& & & \mathbf{b}_{GT}(\boldsymbol{\theta})
\end{bmatrix}'\\
=&\begin{bmatrix}
\frac{\mathbf{b}_{11}(\boldsymbol{\theta})\boldsymbol{\Omega}^{\mathbf{s}}_{11}\mathbf{b}_{11}(\boldsymbol{\theta})'}{\rho_1\kappa_1} & & & \\
& \frac{\mathbf{b}_{12}(\boldsymbol{\theta})\boldsymbol{\Omega}^{\mathbf{s}}_{12}\mathbf{b}_{12}(\boldsymbol{\theta})'}{\rho_1\kappa_2} & & \\
& & \ddots & \\
& & & \frac{\mathbf{b}_{GT}(\boldsymbol{\theta})\boldsymbol{\Omega}^{\mathbf{s}}_{GT}\mathbf{b}_{GT}(\boldsymbol{\theta})'}{\rho_G\kappa_T}
\end{bmatrix}
\end{aligned}
$$

A.7 Equivalence of the two ways of computing $\hat{\mathbf{M}}$

Proof. Expanding $\hat{\mathbf{M}} = \mathbf{b}(\check{\boldsymbol{\theta}})\hat{\boldsymbol{\Omega}}^{\mathbf{S}}\mathbf{b}(\check{\boldsymbol{\theta}})'$ shows that

$$
\begin{aligned}
&\mathbf{b}(\check{\boldsymbol{\theta}})\hat{\boldsymbol{\Omega}}^{\mathbf{S}}\mathbf{b}(\check{\boldsymbol{\theta}})' \\
&= \begin{bmatrix} \mathbf{b}_{11}(\check{\boldsymbol{\theta}}) & & & \\ & \mathbf{b}_{12}(\check{\boldsymbol{\theta}}) & & \\ & & \ddots & \\ & & & \mathbf{b}_{GT}(\check{\boldsymbol{\theta}}) \end{bmatrix} \begin{bmatrix} \frac{\hat{\boldsymbol{\Omega}}^{\mathbf{S}}_{11}}{n_{11}/n} & & & \\ & \frac{\hat{\boldsymbol{\Omega}}^{\mathbf{S}}_{12}}{n_{12}/n} & & \\ & & \ddots & \\ & & & \frac{\hat{\boldsymbol{\Omega}}^{\mathbf{S}}_{GT}}{n_{GT}/n} \end{bmatrix}. \\
&\quad \begin{bmatrix} \mathbf{b}_{11}(\check{\boldsymbol{\theta}}) & & & \\ & \mathbf{b}_{12}(\check{\boldsymbol{\theta}}) & & \\ & & \ddots & \\ & & & \mathbf{b}_{GT}(\check{\boldsymbol{\theta}}) \end{bmatrix}' \\
&= \begin{bmatrix} \frac{\mathbf{b}_{11}(\check{\boldsymbol{\theta}})\hat{\boldsymbol{\Omega}}^{\mathbf{S}}_{11}\mathbf{b}_{11}(\check{\boldsymbol{\theta}})'}{n_{11}/n} & & & \\ & \frac{\mathbf{b}_{12}(\check{\boldsymbol{\theta}})\hat{\boldsymbol{\Omega}}^{\mathbf{S}}_{12}\mathbf{b}_{12}(\check{\boldsymbol{\theta}})'}{n_{12}/n} & & \\ & & \ddots & \\ & & & \frac{\mathbf{b}_{GT}(\check{\boldsymbol{\theta}})\hat{\boldsymbol{\Omega}}^{\mathbf{S}}_{GT}\mathbf{b}_{GT}(\check{\boldsymbol{\theta}})'}{n_{GT}/n} \end{bmatrix}
\end{aligned}
$$

For each (g,t),

$$
\begin{aligned}
&\mathbf{b}_{gt}(\check{\boldsymbol{\theta}})\hat{\boldsymbol{\Omega}}^{\mathbf{S}}_{gt}\mathbf{b}_{gt}(\check{\boldsymbol{\theta}})' \\
&= \left[\mathbf{I}_P \otimes (-1, \check{\boldsymbol{\beta}}'_{gt})\right] [\hat{\Gamma}_{pq,gt}]_P \left[\mathbf{I}_P \otimes (-1, \check{\boldsymbol{\beta}}'_{gt})\right]' \\
&= \begin{bmatrix} (-1,\check{\boldsymbol{\beta}}'_{gt}) & & & \\ & (-1,\check{\boldsymbol{\beta}}'_{gt}) & & \\ & & \ddots & \\ & & & (-1,\check{\boldsymbol{\beta}}'_{gt}) \end{bmatrix}_P \begin{bmatrix} \hat{\Gamma}_{11,gt} & \hat{\Gamma}_{12,gt} & \cdots & \hat{\Gamma}_{1P,gt} \\ \hat{\Gamma}_{21,gt} & \hat{\Gamma}_{22,gt} & \cdots & \hat{\Gamma}_{PP,gt} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Gamma}_{P1,gt} & \hat{\Gamma}_{P2,gt} & \cdots & \hat{\Gamma}_{PP,gt} \end{bmatrix} \cdot \\
&\quad \begin{bmatrix} (-1,\check{\boldsymbol{\beta}}'_{gt}) & & & \\ & (-1,\check{\boldsymbol{\beta}}'_{gt}) & & \\ & & \ddots & \\ & & & (-1,\check{\boldsymbol{\beta}}'_{gt}) \end{bmatrix}'_P \\
&= \begin{bmatrix} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{11,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} & \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{12,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} & \cdots & \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{1P,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} \\ \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{21,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} & \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{22,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} & \cdots & \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{PP,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} \\ \vdots & \vdots & \ddots & \vdots \\ \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{P1,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' & \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{P2,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} & \cdots & \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{PP,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} \end{bmatrix}
\end{aligned}
$$

For each $(p,q;g,t)$, recall $\mathbf{v}_{it} = (y_{it}, \mathbf{x}_{it})$ and notice that

$$
\begin{aligned}
\mathbf{v}_{it} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} &= -(y_{it} - \mathbf{x}_{it}\check{\boldsymbol{\beta}}_{gt}) \\
&= -(y_{it} - \mathbf{x}_{it}\check{\boldsymbol{\beta}} - (\check{\eta}_t + \check{\alpha}_g)) = -\check{u}_{it}
\end{aligned}
$$

and that

$$
\hat{\boldsymbol{\mu}}^{z_p\mathbf{v}}_{gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} = n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} z_{pit} \mathbf{v}_{it} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} = n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} z_{pit} \check{u}_{it}
$$

we have

$$
\begin{aligned}
&\begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' \hat{\Gamma}_{pp,gt} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} \\
=&\begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix}' n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} (z_{pit}\mathbf{v}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{z_p\mathbf{v}})'(z_{qit}\mathbf{v}_{it} - \hat{\boldsymbol{\mu}}_{gt}^{z_q\mathbf{v}}) \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} \\
=&n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} \left[z_{pit}\mathbf{v}_{it} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} - \hat{\boldsymbol{\mu}}_{gt}^{z_p\mathbf{v}} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} \right]' \left[z_{qit}\mathbf{v}_{it} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} - \hat{\boldsymbol{\mu}}_{gt}^{z_q\mathbf{v}} \begin{pmatrix} -1 \\ \check{\boldsymbol{\beta}}_{gt} \end{pmatrix} \right] \\
=&n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} \left[z_{pit}\check{u}_{it} - n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} z_{pit}\check{u}_{it} \right]' \left[z_{qit}\check{u}_{it} - n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} z_{pit}\check{u}_{it} \right] \\
=&n_{gt}^{-1} \sum_{i=1}^{n_t} r_{it,g} \left[z_{pit}\check{u}_{it} - \check{\mu}_{gt}^{z_pu} \right]' \left[z_{qit}\check{u}_{it} - \check{\mu}_{gt}^{z_pu} \right] = \hat{\tau}_{pq}.
\end{aligned}
$$

Hence the two ways are numerically equivalent. □

A.8 Further algebra on $\boldsymbol{\Xi}_{gt}$

In general,

$$
\begin{aligned}
\boldsymbol{\Xi}_{gt} &= Var[\mathbf{z}'_{it}y_{it} - \mathbf{z}'_{it}\mathbf{x}_{it}\boldsymbol{\beta} - \mathbf{z}'_{it}(\eta_t + \alpha_g)|g] \\
&= Var[\mathbf{z}'_{it}(\varepsilon_i^f + u_{it})|g] \\
&= E\left[\left(\mathbf{z}'_{it}(\varepsilon_i^f + u_{it}) - E[\mathbf{z}'_{it}(\varepsilon_i^f + u_{it})|g] \right) \left(\mathbf{z}'_{it}(\varepsilon_i^f + u_{it}) - E[\mathbf{z}'_{it}(\varepsilon_i^f + u_{it})|g] \right)' \right] \\
&= E\left[(\varepsilon_i^f + u_{it})^2 \mathbf{z}'_{it}\mathbf{z}_{it}|g \right] - E\left[\mathbf{z}'_{it}(\varepsilon_i^f + u_{it})|g \right] E\left[(\varepsilon_i^f + u_{it})\mathbf{z}_{it}|g \right] \\
&= E\left[(\varepsilon_i^f + u_{it})^2 \mathbf{z}'_{it}\mathbf{z}_{it}|g \right] \\
&= E\left[(\varepsilon_i^f + u_{it})^2|g \right] E\left[\mathbf{z}'_{it}\mathbf{z}_{it}|g \right] \ if\ independent \\
&= \sigma^2_{\varepsilon^f + u}(g) \cdot \boldsymbol{\Omega}^{\mathbf{z}} \ if\ independent\ of\ g_i\ \&\ mean\ zero
\end{aligned}
$$

For $\mathbf{z}_{it} = (1, x_{2it})$ in Case 3,

$$\begin{aligned}
\mathbf{\Xi}_{gt} &= E\left[(\varepsilon_i^f + u_{it})^2 \mathbf{z}_{it}' \mathbf{z}_{it} | g\right] \\
&= E\left[(\varepsilon_i^f + u_{it})^2 | g\right] E\left[\mathbf{z}_{it}' \mathbf{z}_{it} | g\right] \;\; if \;\; independent \\
&= \sigma^2_{\varepsilon f + u} E\left[(1, x_{2it})'(1, x_{2it}) | g\right] \\
&= \sigma^2_{\varepsilon f + u} E\left[\begin{pmatrix} 1 & x_{2it} \\ x_{2it} & x_{2it}^2 \end{pmatrix} | g\right] \\
&= \sigma^2_{\varepsilon f + u} \begin{pmatrix} 1 & gt/150 \\ gt/150 & 2 + (gt/150)^2 \end{pmatrix}
\end{aligned}$$

APPENDIX B

ADDITIONAL TABLES

Table B.1 Small panel with $G = 6$, $T = 4$. Case 1.a: $x_{2it} \sim N(gt/6, 1)$, $n_{gt} = 200$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0021	0.1590	0.1555	0.1397	0.1413	1.0020	0.1585
		(0.1534)	(0.0122)	(0.0326)	(0.0424)	(0.0628)	(0.1531)	(0.0121)
	x_3	0.9973	0.0887	0.0866	0.0826	0.0878	0.9976	0.0884
		(0.0901)	(0.0040)	(0.0176)	(0.0195)	(0.0315)	(0.0904)	(0.0040)
	x_4	1.0127	0.2848	0.2780	0.2624	0.2943	1.0129	0.2838
		(0.2818)	(0.0137)	(0.0570)	(0.0659)	(0.1070)	(0.2834)	(0.0137)
IV: z	x_2	1.0530	0.1504	0.1486	0.1225	0.1240	1.0532	0.1486
		(0.1447)	(0.0104)	(0.0206)	(0.0304)	(0.0514)	(0.1452)	(0.0103)
	x_3	0.9880	0.0870	0.0858	0.0690	0.0736	0.9875	0.0860
		(0.0884)	(0.0037)	(0.0105)	(0.0148)	(0.0263)	(0.0892)	(0.0036)
	x_4	1.0230	0.2788	0.2749	0.2187	0.2455	1.0221	0.2755
		(0.2764)	(0.0129)	(0.0349)	(0.0496)	(0.0870)	(0.2774)	(0.0127)
IV: x_2	x_2	1.2350	0.1360	0.0931	0.1025	0.1344	1.4768	0.0326
		(0.1365)	(0.0085)	(0.0216)	(0.0257)	(0.0535)	(0.0329)	(0.0009)
	x_3	1.0028	0.1100	0.0528	0.0709	0.0708	0.9063	0.0802
		(0.1091)	(0.0048)	(0.0118)	(0.0235)	(0.0304)	(0.0833)	(0.0027)
	x_4	0.9988	0.3415	0.1670	0.2146	0.2608	1.1332	0.2694
		(0.3368)	(0.0184)	(0.0384)	(0.0699)	(0.1138)	(0.2681)	(0.0116)
IV: x_3	x_2	1.0329	0.1634	0.1218	0.1115	0.1253	1.0504	0.1424
		(0.1561)	(0.0117)	(0.0190)	(0.0258)	(0.0532)	(0.1404)	(0.0098)
	x_3	0.9879	0.0463	0.0353	0.0367	0.0406	0.9885	0.0415
		(0.0458)	(0.0016)	(0.0049)	(0.0064)	(0.0138)	(0.0419)	(0.0012)
	x_4	1.0083	0.2932	0.2332	0.2191	0.2567	1.0254	0.2617
		(0.2885)	(0.0134)	(0.0333)	(0.0421)	(0.0918)	(0.2607)	(0.0112)
IV: x_4	x_2	1.0061	0.1588	0.1007	0.1011	0.1173	1.0505	0.1479
		(0.1546)	(0.0119)	(0.0175)	(0.0269)	(0.0509)	(0.1463)	(0.0105)
	x_3	0.9975	0.0863	0.0542	0.0564	0.0693	0.9892	0.0812
		(0.0862)	(0.0036)	(0.0088)	(0.0120)	(0.0252)	(0.0811)	(0.0032)
	x_4	1.0097	0.1585	0.1271	0.1303	0.1408	1.0143	0.1011
		(0.1550)	(0.0043)	(0.0206)	(0.0268)	(0.0491)	(0.1008)	(0.0023)
IV: all	x_2	1.2449	0.1267	0.0574	0.0867	0.1108	1.4718	0.0319
		(0.1268)	(0.0079)	(0.0111)	(0.0218)	(0.0446)	(0.0335)	(0.0008)
	x_3	0.9943	0.0881	0.0301	0.0514	0.0582	0.9687	0.0392
		(0.0871)	(0.0034)	(0.0055)	(0.0157)	(0.0226)	(0.0406)	(0.0009)
	x_4	0.9839	0.3163	0.1043	0.1798	0.2246	1.0243	0.0972
		(0.3119)	(0.0155)	(0.0198)	(0.0578)	(0.0964)	(0.0983)	(0.0016)

Table B.2 Small panel with $G = 6$, $T = 4$. Case 1.b: $x_{2it} \sim N(gt/6, 1)$, $n_{gt} = 1000$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9940	0.0722	0.0700	0.0628	0.0639	0.9939	0.0721
		(0.0715)	(0.0018)	(0.0144)	(0.0180)	(0.0286)	(0.0714)	(0.0018)
	x_3	1.0037	0.0398	0.0385	0.0367	0.0390	1.0037	0.0397
		(0.0391)	(0.0007)	(0.0079)	(0.0085)	(0.0146)	(0.0391)	(0.0007)
	x_4	1.0067	0.1282	0.1244	0.1174	0.1333	1.0069	0.1281
		(0.1274)	(0.0024)	(0.0256)	(0.0308)	(0.0486)	(0.1276)	(0.0024)
IV: z	x_2	0.9940	0.0718	0.0714	0.0517	0.0531	0.9940	0.0716
		(0.0709)	(0.0018)	(0.0088)	(0.0142)	(0.0233)	(0.0710)	(0.0018)
	x_3	1.0038	0.0397	0.0395	0.0301	0.0322	1.0037	0.0396
		(0.0389)	(0.0007)	(0.0049)	(0.0069)	(0.0120)	(0.0390)	(0.0007)
	x_4	1.0064	0.1278	0.1272	0.0963	0.1100	1.0067	0.1275
		(0.1274)	(0.0024)	(0.0156)	(0.0247)	(0.0399)	(0.1278)	(0.0024)
IV: x_2	x_2	0.9979	0.0360	0.0336	0.0312	0.0306	0.9987	0.0205
		(0.0372)	(0.0011)	(0.0064)	(0.0057)	(0.0120)	(0.0208)	(0.0002)
	x_3	1.0036	0.0464	0.0270	0.0286	0.0373	1.0026	0.0371
		(0.0459)	(0.0009)	(0.0052)	(0.0071)	(0.0140)	(0.0357)	(0.0005)
	x_4	1.0219	0.1536	0.0751	0.0839	0.1172	1.0086	0.1259
		(0.1538)	(0.0033)	(0.0145)	(0.0273)	(0.0518)	(0.1241)	(0.0023)
IV: x_3	x_2	0.9998	0.0768	0.0565	0.0512	0.0573	0.9946	0.0680
		(0.0778)	(0.0020)	(0.0078)	(0.0110)	(0.0248)	(0.0658)	(0.0016)
	x_3	1.0025	0.0207	0.0158	0.0165	0.0181	1.0031	0.0188
		(0.0206)	(0.0003)	(0.0022)	(0.0028)	(0.0065)	(0.0186)	(0.0002)
	x_4	1.0097	0.1337	0.1059	0.0992	0.1159	1.0070	0.1202
		(0.1327)	(0.0025)	(0.0147)	(0.0195)	(0.0408)	(0.1218)	(0.0020)
IV: x_4	x_2	0.9959	0.0728	0.0457	0.0458	0.0534	0.9936	0.0708
		(0.0711)	(0.0018)	(0.0075)	(0.0114)	(0.0238)	(0.0695)	(0.0017)
	x_3	1.0040	0.0387	0.0242	0.0251	0.0313	1.0045	0.0370
		(0.0381)	(0.0006)	(0.0040)	(0.0053)	(0.0117)	(0.0370)	(0.0006)
	x_4	1.0028	0.0705	0.0567	0.0580	0.0635	1.0011	0.0454
		(0.0686)	(0.0007)	(0.0093)	(0.0121)	(0.0218)	(0.0442)	(0.0003)
IV: all	x_2	1.0002	0.0377	0.0256	0.0271	0.0300	0.9986	0.0203
		(0.0390)	(0.0010)	(0.0030)	(0.0045)	(0.0112)	(0.0208)	(0.0002)
	x_3	1.0028	0.0250	0.0132	0.0150	0.0201	1.0031	0.0182
		(0.0248)	(0.0003)	(0.0015)	(0.0024)	(0.0073)	(0.0179)	(0.0002)
	x_4	1.0154	0.1166	0.0515	0.0599	0.0872	1.0014	0.0448
		(0.1157)	(0.0018)	(0.0060)	(0.0167)	(0.0357)	(0.0440)	(0.0003)

Table B.3 Small panel with $G = 6$, $T = 4$. Case 1.1: $x_{2it} \sim N(gt/6, 1)$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9898	0.1606	0.1568	0.1402	0.1430	0.9900	0.1601
		(0.1652)	(0.0089)	(0.0342)	(0.0423)	(0.0615)	(0.1652)	(0.0089)
	x_3	1.0083	0.0889	0.0868	0.0823	0.0871	1.0081	0.0886
		(0.0875)	(0.0034)	(0.0185)	(0.0208)	(0.0326)	(0.0874)	(0.0034)
	x_4	1.0103	0.2863	0.2794	0.2619	0.2986	1.0113	0.2853
		(0.2886)	(0.0124)	(0.0593)	(0.0716)	(0.1097)	(0.2882)	(0.0123)
IV: z	x_2	0.9896	0.1565	0.1546	0.1186	0.1203	0.9905	0.1547
		(0.1599)	(0.0084)	(0.0198)	(0.0308)	(0.0477)	(0.1599)	(0.0083)
	x_3	1.0091	0.0879	0.0868	0.0682	0.0716	1.0084	0.0869
		(0.0864)	(0.0033)	(0.0105)	(0.0160)	(0.0268)	(0.0863)	(0.0033)
	x_4	1.0095	0.2817	0.2783	0.2179	0.2447	1.0111	0.2785
		(0.2832)	(0.0119)	(0.0341)	(0.0541)	(0.0908)	(0.2818)	(0.0117)
IV: x_2	x_2	0.9966	0.0809	0.0752	0.0701	0.0687	0.9978	0.0453
		(0.0829)	(0.0052)	(0.0153)	(0.0131)	(0.0255)	(0.0476)	(0.0010)
	x_3	1.0086	0.1035	0.0606	0.0642	0.0817	1.0065	0.0819
		(0.1007)	(0.0043)	(0.0122)	(0.0165)	(0.0322)	(0.0808)	(0.0027)
	x_4	1.0183	0.3417	0.1684	0.1871	0.2611	1.0130	0.2750
		(0.3441)	(0.0167)	(0.0341)	(0.0628)	(0.1184)	(0.2811)	(0.0113)
IV: x_3	x_2	1.0019	0.1681	0.1241	0.1127	0.1275	0.9931	0.1473
		(0.1685)	(0.0089)	(0.0180)	(0.0253)	(0.0542)	(0.1513)	(0.0073)
	x_3	1.0038	0.0464	0.0353	0.0367	0.0399	1.0035	0.0416
		(0.0468)	(0.0013)	(0.0048)	(0.0060)	(0.0140)	(0.0433)	(0.0010)
	x_4	1.0200	0.2956	0.2340	0.2198	0.2580	1.0173	0.2633
		(0.2924)	(0.0123)	(0.0327)	(0.0441)	(0.0882)	(0.2629)	(0.0101)
IV: x_4	x_2	0.9909	0.1610	0.1020	0.1020	0.1189	0.9899	0.1529
		(0.1641)	(0.0087)	(0.0177)	(0.0263)	(0.0507)	(0.1565)	(0.0079)
	x_3	1.0105	0.0865	0.0544	0.0563	0.0693	1.0093	0.0814
		(0.0840)	(0.0030)	(0.0091)	(0.0126)	(0.0266)	(0.0800)	(0.0027)
	x_4	1.0028	0.1582	0.1276	0.1301	0.1415	1.0005	0.1008
		(0.1582)	(0.0036)	(0.0210)	(0.0278)	(0.0498)	(0.0996)	(0.0016)
IV: all	x_2	0.9993	0.0839	0.0566	0.0602	0.0668	0.9979	0.0443
		(0.0849)	(0.0045)	(0.0070)	(0.0104)	(0.0250)	(0.0479)	(0.0010)
	x_3	1.0059	0.0559	0.0296	0.0335	0.0443	1.0036	0.0398
		(0.0547)	(0.0017)	(0.0035)	(0.0056)	(0.0167)	(0.0421)	(0.0008)
	x_4	1.0151	0.2579	0.1144	0.1336	0.1932	1.0013	0.0984
		(0.2575)	(0.0091)	(0.0140)	(0.0385)	(0.0819)	(0.1007)	(0.0015)

Table B.4 Small panel with $G = 6$, $T = 4$. Case 1.1: $x_{2it} \sim N(gt/6, 1)$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0027	0.0722	0.0697	0.0618	0.0622	1.0027	0.0721
		(0.0722)	(0.0017)	(0.0145)	(0.0181)	(0.0281)	(0.0722)	(0.0017)
	x_3	0.9971	0.0398	0.0384	0.0366	0.0390	0.9972	0.0398
		(0.0401)	(0.0007)	(0.0079)	(0.0086)	(0.0138)	(0.0401)	(0.0007)
	x_4	1.0010	0.1279	0.1235	0.1167	0.1325	1.0012	0.1278
		(0.1304)	(0.0024)	(0.0256)	(0.0306)	(0.0490)	(0.1304)	(0.0024)
IV: z	x_2	1.0028	0.0718	0.0709	0.0510	0.0516	1.0029	0.0716
		(0.0715)	(0.0017)	(0.0087)	(0.0143)	(0.0229)	(0.0715)	(0.0017)
	x_3	0.9971	0.0397	0.0392	0.0300	0.0323	0.9973	0.0396
		(0.0400)	(0.0007)	(0.0048)	(0.0069)	(0.0114)	(0.0401)	(0.0007)
	x_4	1.0009	0.1275	0.1259	0.0957	0.1092	1.0009	0.1272
		(0.1304)	(0.0024)	(0.0152)	(0.0245)	(0.0403)	(0.1302)	(0.0024)
IV: x_2	x_2	1.0030	0.0360	0.0332	0.0308	0.0301	0.9992	0.0205
		(0.0353)	(0.0011)	(0.0066)	(0.0058)	(0.0115)	(0.0204)	(0.0002)
	x_3	0.9962	0.0465	0.0267	0.0284	0.0359	0.9980	0.0372
		(0.0461)	(0.0009)	(0.0053)	(0.0072)	(0.0137)	(0.0374)	(0.0005)
	x_4	1.0016	0.1535	0.0742	0.0832	0.1166	0.9999	0.1256
		(0.1567)	(0.0034)	(0.0148)	(0.0290)	(0.0554)	(0.1278)	(0.0023)
IV: x_3	x_2	1.0005	0.0769	0.0565	0.0511	0.0574	1.0006	0.0680
		(0.0754)	(0.0018)	(0.0075)	(0.0112)	(0.0255)	(0.0674)	(0.0015)
	x_3	0.9989	0.0207	0.0158	0.0165	0.0180	0.9998	0.0188
		(0.0211)	(0.0003)	(0.0021)	(0.0026)	(0.0060)	(0.0190)	(0.0002)
	x_4	1.0006	0.1335	0.1057	0.0993	0.1159	0.9981	0.1198
		(0.1349)	(0.0025)	(0.0139)	(0.0182)	(0.0404)	(0.1211)	(0.0021)
IV: x_4	x_2	1.0030	0.0729	0.0456	0.0452	0.0518	1.0031	0.0708
		(0.0731)	(0.0017)	(0.0074)	(0.0114)	(0.0234)	(0.0707)	(0.0017)
	x_3	0.9965	0.0388	0.0242	0.0251	0.0310	0.9970	0.0370
		(0.0391)	(0.0006)	(0.0039)	(0.0053)	(0.0117)	(0.0369)	(0.0006)
	x_4	1.0027	0.0705	0.0564	0.0579	0.0628	1.0032	0.0454
		(0.0713)	(0.0007)	(0.0091)	(0.0120)	(0.0222)	(0.0461)	(0.0003)
IV: all	x_2	1.0022	0.0378	0.0255	0.0268	0.0299	0.9991	0.0203
		(0.0369)	(0.0010)	(0.0029)	(0.0046)	(0.0113)	(0.0203)	(0.0002)
	x_3	0.9980	0.0250	0.0132	0.0149	0.0197	0.9998	0.0182
		(0.0252)	(0.0004)	(0.0015)	(0.0024)	(0.0069)	(0.0183)	(0.0002)
	x_4	1.0025	0.1164	0.0512	0.0595	0.0863	1.0026	0.0449
		(0.1185)	(0.0019)	(0.0059)	(0.0175)	(0.0377)	(0.0460)	(0.0003)

Table B.5 Small panel with $G = 6$, $T = 4$. Case 2.a: $x_{2it} \sim N(gt/6, 1) + f_i$, $n_{gt} = 200$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0021	0.1590	0.1555	0.1397	0.1413	1.0020	0.1585
		(0.1534)	(0.0122)	(0.0326)	(0.0424)	(0.0628)	(0.1531)	(0.0121)
	x_3	0.9973	0.0887	0.0866	0.0826	0.0878	0.9976	0.0884
		(0.0901)	(0.0040)	(0.0176)	(0.0195)	(0.0315)	(0.0904)	(0.0040)
	x_4	1.0127	0.2848	0.2780	0.2624	0.2943	1.0129	0.2838
		(0.2818)	(0.0137)	(0.0570)	(0.0659)	(0.1070)	(0.2834)	(0.0137)
IV: z	x_2	1.0530	0.1504	0.1486	0.1225	0.1240	1.0532	0.1486
		(0.1447)	(0.0104)	(0.0206)	(0.0304)	(0.0514)	(0.1452)	(0.0103)
	x_3	0.9880	0.0870	0.0858	0.0690	0.0736	0.9875	0.0860
		(0.0884)	(0.0037)	(0.0105)	(0.0148)	(0.0263)	(0.0892)	(0.0036)
	x_4	1.0230	0.2788	0.2749	0.2187	0.2455	1.0221	0.2755
		(0.2764)	(0.0129)	(0.0349)	(0.0496)	(0.0870)	(0.2774)	(0.0127)
IV: x_2	x_2	1.2350	0.1360	0.0931	0.1025	0.1344	1.4768	0.0326
		(0.1365)	(0.0085)	(0.0216)	(0.0257)	(0.0535)	(0.0329)	(0.0009)
	x_3	1.0028	0.1100	0.0528	0.0709	0.0708	0.9063	0.0802
		(0.1091)	(0.0048)	(0.0118)	(0.0235)	(0.0304)	(0.0833)	(0.0027)
	x_4	0.9988	0.3415	0.1670	0.2146	0.2608	1.1332	0.2694
		(0.3368)	(0.0184)	(0.0384)	(0.0699)	(0.1138)	(0.2681)	(0.0116)
IV: x_3	x_2	1.0329	0.1634	0.1218	0.1115	0.1253	1.0504	0.1424
		(0.1561)	(0.0117)	(0.0190)	(0.0258)	(0.0532)	(0.1404)	(0.0098)
	x_3	0.9879	0.0463	0.0353	0.0367	0.0406	0.9885	0.0415
		(0.0458)	(0.0016)	(0.0049)	(0.0064)	(0.0138)	(0.0419)	(0.0012)
	x_4	1.0083	0.2932	0.2332	0.2191	0.2567	1.0254	0.2617
		(0.2885)	(0.0134)	(0.0333)	(0.0421)	(0.0918)	(0.2607)	(0.0112)
IV: x_4	x_2	1.0061	0.1588	0.1007	0.1011	0.1173	1.0505	0.1479
		(0.1546)	(0.0119)	(0.0175)	(0.0269)	(0.0509)	(0.1463)	(0.0105)
	x_3	0.9975	0.0863	0.0542	0.0564	0.0693	0.9892	0.0812
		(0.0862)	(0.0036)	(0.0088)	(0.0120)	(0.0252)	(0.0811)	(0.0032)
	x_4	1.0097	0.1585	0.1271	0.1303	0.1408	1.0143	0.1011
		(0.1550)	(0.0043)	(0.0206)	(0.0268)	(0.0491)	(0.1008)	(0.0023)
IV: all	x_2	1.2449	0.1267	0.0574	0.0867	0.1108	1.4718	0.0319
		(0.1268)	(0.0079)	(0.0111)	(0.0218)	(0.0446)	(0.0335)	(0.0008)
	x_3	0.9943	0.0881	0.0301	0.0514	0.0582	0.9687	0.0392
		(0.0871)	(0.0034)	(0.0055)	(0.0157)	(0.0226)	(0.0406)	(0.0009)
	x_4	0.9839	0.3163	0.1043	0.1798	0.2246	1.0243	0.0972
		(0.3119)	(0.0155)	(0.0198)	(0.0578)	(0.0964)	(0.0983)	(0.0016)

Table B.6 Small panel with $G = 6$, $T = 4$. Case 2.b: $x_{2it} \sim N(gt/6, 1) + f_i$, $n_{gt} = 1000$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9993	0.0721	0.0698	0.0628	0.0642	0.9992	0.0720
		(0.0717)	(0.0025)	(0.0144)	(0.0181)	(0.0287)	(0.0716)	(0.0025)
	x_3	1.0027	0.0398	0.0385	0.0367	0.0390	1.0027	0.0397
		(0.0391)	(0.0008)	(0.0079)	(0.0085)	(0.0146)	(0.0392)	(0.0008)
	x_4	1.0082	0.1283	0.1244	0.1174	0.1333	1.0084	0.1282
		(0.1274)	(0.0026)	(0.0256)	(0.0309)	(0.0487)	(0.1277)	(0.0026)
IV: z	x_2	1.0107	0.0712	0.0709	0.0526	0.0545	1.0107	0.0711
		(0.0706)	(0.0025)	(0.0088)	(0.0140)	(0.0235)	(0.0706)	(0.0024)
	x_3	1.0004	0.0396	0.0394	0.0301	0.0323	1.0004	0.0395
		(0.0389)	(0.0008)	(0.0048)	(0.0068)	(0.0120)	(0.0390)	(0.0008)
	x_4	1.0113	0.1277	0.1271	0.0966	0.1102	1.0115	0.1274
		(0.1272)	(0.0026)	(0.0156)	(0.0247)	(0.0401)	(0.1276)	(0.0025)
IV: x_2	x_2	1.2316	0.0609	0.0555	0.0664	0.1070	1.4756	0.0147
		(0.0608)	(0.0017)	(0.0095)	(0.0135)	(0.0301)	(0.0147)	(0.0002)
	x_3	1.0107	0.0491	0.0312	0.0354	0.0398	0.9080	0.0365
		(0.0480)	(0.0009)	(0.0052)	(0.0098)	(0.0134)	(0.0352)	(0.0005)
	x_4	1.0146	0.1532	0.0992	0.1064	0.1261	1.1474	0.1238
		(0.1537)	(0.0034)	(0.0166)	(0.0313)	(0.0536)	(0.1225)	(0.0023)
IV: x_3	x_2	1.0101	0.0763	0.0562	0.0512	0.0571	1.0094	0.0675
		(0.0776)	(0.0026)	(0.0078)	(0.0110)	(0.0247)	(0.0658)	(0.0022)
	x_3	1.0016	0.0207	0.0158	0.0165	0.0181	1.0023	0.0188
		(0.0206)	(0.0003)	(0.0022)	(0.0028)	(0.0065)	(0.0185)	(0.0002)
	x_4	1.0115	0.1337	0.1058	0.0991	0.1160	1.0087	0.1201
		(0.1325)	(0.0026)	(0.0147)	(0.0195)	(0.0408)	(0.1217)	(0.0022)
IV: x_4	x_2	1.0027	0.0726	0.0456	0.0458	0.0537	1.0094	0.0703
		(0.0712)	(0.0025)	(0.0075)	(0.0114)	(0.0238)	(0.0690)	(0.0024)
	x_3	1.0030	0.0387	0.0242	0.0251	0.0313	1.0018	0.0369
		(0.0382)	(0.0007)	(0.0040)	(0.0053)	(0.0117)	(0.0370)	(0.0007)
	x_4	1.0039	0.0705	0.0567	0.0580	0.0635	1.0017	0.0454
		(0.0687)	(0.0008)	(0.0092)	(0.0121)	(0.0218)	(0.0442)	(0.0004)
IV: all	x_2	1.2392	0.0574	0.0337	0.0573	0.0867	1.4726	0.0146
		(0.0575)	(0.0016)	(0.0050)	(0.0117)	(0.0257)	(0.0146)	(0.0002)
	x_3	1.0034	0.0394	0.0175	0.0263	0.0337	0.9803	0.0179
		(0.0387)	(0.0006)	(0.0025)	(0.0065)	(0.0100)	(0.0176)	(0.0002)
	x_4	0.9986	0.1423	0.0609	0.0899	0.1093	1.0102	0.0442
		(0.1424)	(0.0029)	(0.0088)	(0.0260)	(0.0448)	(0.0429)	(0.0003)

Table B.7 Small panel with $G = 6$, $T = 4$. Case 2.1: $x_{2it} \sim N(gt/6, 1) + f_i$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0147	0.1591	0.1553	0.1399	0.1450	1.0148	0.1586
		(0.1624)	(0.0126)	(0.0351)	(0.0428)	(0.0643)	(0.1624)	(0.0125)
	x_3	1.0034	0.0889	0.0868	0.0824	0.0875	1.0033	0.0886
		(0.0871)	(0.0041)	(0.0187)	(0.0209)	(0.0327)	(0.0871)	(0.0041)
	x_4	1.0175	0.2864	0.2795	0.2623	0.2987	1.0184	0.2854
		(0.2885)	(0.0133)	(0.0597)	(0.0720)	(0.1101)	(0.2881)	(0.0132)
IV: z	x_2	1.0622	0.1505	0.1490	0.1216	0.1253	1.0638	0.1487
		(0.1535)	(0.0110)	(0.0202)	(0.0303)	(0.0516)	(0.1538)	(0.0109)
	x_3	0.9950	0.0872	0.0863	0.0687	0.0725	0.9942	0.0862
		(0.0855)	(0.0037)	(0.0106)	(0.0159)	(0.0269)	(0.0855)	(0.0037)
	x_4	1.0297	0.2804	0.2775	0.2190	0.2455	1.0314	0.2772
		(0.2812)	(0.0124)	(0.0341)	(0.0543)	(0.0905)	(0.2796)	(0.0122)
IV: x_2	x_2	1.2409	0.1363	0.0920	0.1010	0.1331	1.4705	0.0324
		(0.1366)	(0.0084)	(0.0235)	(0.0265)	(0.0548)	(0.0328)	(0.0009)
	x_3	1.0163	0.1100	0.0525	0.0702	0.0707	0.9150	0.0804
		(0.1082)	(0.0046)	(0.0126)	(0.0238)	(0.0311)	(0.0793)	(0.0026)
	x_4	1.0187	0.3425	0.1662	0.2122	0.2587	1.1427	0.2708
		(0.3458)	(0.0175)	(0.0403)	(0.0745)	(0.1200)	(0.2719)	(0.0113)
IV: x_3	x_2	1.0486	0.1630	0.1211	0.1114	0.1277	1.0592	0.1422
		(0.1633)	(0.0119)	(0.0189)	(0.0259)	(0.0546)	(0.1457)	(0.0099)
	x_3	0.9996	0.0463	0.0352	0.0366	0.0399	1.0002	0.0415
		(0.0465)	(0.0016)	(0.0049)	(0.0061)	(0.0140)	(0.0430)	(0.0012)
	x_4	1.0287	0.2948	0.2336	0.2195	0.2580	1.0250	0.2626
		(0.2912)	(0.0129)	(0.0328)	(0.0442)	(0.0887)	(0.2612)	(0.0107)
IV: x_4	x_2	1.0223	0.1584	0.1008	0.1012	0.1200	1.0598	0.1478
		(0.1602)	(0.0123)	(0.0184)	(0.0268)	(0.0525)	(0.1503)	(0.0108)
	x_3	1.0058	0.0864	0.0545	0.0564	0.0696	0.9974	0.0812
		(0.0836)	(0.0036)	(0.0092)	(0.0127)	(0.0268)	(0.0793)	(0.0032)
	x_4	1.0077	0.1581	0.1276	0.1302	0.1417	1.0033	0.1008
		(0.1577)	(0.0043)	(0.0212)	(0.0279)	(0.0498)	(0.0991)	(0.0022)
IV: all	x_2	1.2502	0.1271	0.0568	0.0854	0.1108	1.4657	0.0318
		(0.1276)	(0.0078)	(0.0120)	(0.0223)	(0.0458)	(0.0335)	(0.0008)
	x_3	1.0080	0.0881	0.0299	0.0510	0.0579	0.9815	0.0392
		(0.0864)	(0.0032)	(0.0059)	(0.0159)	(0.0234)	(0.0415)	(0.0009)
	x_4	1.0039	0.3170	0.1038	0.1782	0.2232	1.0100	0.0970
		(0.3193)	(0.0145)	(0.0206)	(0.0619)	(0.1021)	(0.0996)	(0.0016)

Table B.8 Small panel with $G = 6$, $T = 4$. Case 2.1: $x_{2it} \sim N(gt/6, 1) + f_i$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0080	0.0719	0.0694	0.0615	0.0619	1.0080	0.0719
		(0.0721)	(0.0024)	(0.0145)	(0.0179)	(0.0280)	(0.0721)	(0.0024)
	x_3	0.9961	0.0398	0.0384	0.0366	0.0390	0.9961	0.0398
		(0.0401)	(0.0008)	(0.0079)	(0.0086)	(0.0138)	(0.0401)	(0.0008)
	x_4	1.0026	0.1279	0.1235	0.1166	0.1322	1.0028	0.1278
		(0.1304)	(0.0027)	(0.0255)	(0.0305)	(0.0489)	(0.1303)	(0.0027)
IV: z	x_2	1.0191	0.0711	0.0702	0.0514	0.0520	1.0193	0.0709
		(0.0714)	(0.0024)	(0.0086)	(0.0138)	(0.0227)	(0.0715)	(0.0023)
	x_3	0.9939	0.0396	0.0391	0.0300	0.0323	0.9940	0.0395
		(0.0400)	(0.0008)	(0.0047)	(0.0068)	(0.0115)	(0.0400)	(0.0008)
	x_4	1.0057	0.1273	0.1258	0.0957	0.1090	1.0057	0.1270
		(0.1302)	(0.0026)	(0.0151)	(0.0244)	(0.0403)	(0.1300)	(0.0026)
IV: x_2	x_2	1.2376	0.0610	0.0545	0.0645	0.1023	1.4775	0.0146
		(0.0602)	(0.0018)	(0.0092)	(0.0132)	(0.0293)	(0.0147)	(0.0002)
	x_3	1.0016	0.0492	0.0308	0.0350	0.0389	0.9031	0.0365
		(0.0485)	(0.0010)	(0.0051)	(0.0098)	(0.0128)	(0.0364)	(0.0006)
	x_4	0.9972	0.1535	0.0979	0.1051	0.1255	1.1404	0.1235
		(0.1544)	(0.0035)	(0.0163)	(0.0310)	(0.0525)	(0.1251)	(0.0023)
IV: x_3	x_2	1.0108	0.0763	0.0562	0.0509	0.0571	1.0152	0.0674
		(0.0749)	(0.0025)	(0.0076)	(0.0111)	(0.0253)	(0.0672)	(0.0021)
	x_3	0.9980	0.0207	0.0158	0.0165	0.0180	0.9991	0.0188
		(0.0211)	(0.0003)	(0.0021)	(0.0026)	(0.0060)	(0.0189)	(0.0003)
	x_4	1.0025	0.1334	0.1057	0.0993	0.1158	0.9998	0.1198
		(0.1349)	(0.0027)	(0.0139)	(0.0182)	(0.0403)	(0.1210)	(0.0022)
IV: x_4	x_2	1.0096	0.0725	0.0454	0.0450	0.0515	1.0191	0.0701
		(0.0730)	(0.0024)	(0.0074)	(0.0114)	(0.0233)	(0.0703)	(0.0023)
	x_3	0.9955	0.0388	0.0241	0.0251	0.0310	0.9943	0.0370
		(0.0391)	(0.0008)	(0.0039)	(0.0053)	(0.0116)	(0.0368)	(0.0007)
	x_4	1.0038	0.0704	0.0564	0.0578	0.0627	1.0038	0.0454
		(0.0713)	(0.0009)	(0.0091)	(0.0120)	(0.0221)	(0.0460)	(0.0005)
IV: all	x_2	1.2440	0.0575	0.0332	0.0555	0.0824	1.4744	0.0145
		(0.0568)	(0.0017)	(0.0048)	(0.0114)	(0.0251)	(0.0147)	(0.0002)
	x_3	0.9957	0.0394	0.0172	0.0260	0.0329	0.9777	0.0179
		(0.0391)	(0.0007)	(0.0025)	(0.0065)	(0.0093)	(0.0180)	(0.0002)
	x_4	0.9832	0.1426	0.0601	0.0886	0.1086	1.0106	0.0442
		(0.1436)	(0.0030)	(0.0086)	(0.0260)	(0.0443)	(0.0448)	(0.0003)

Table B.9 Small panel with $G = 6$, $T = 4$. Case 3.a: $x_{2it} \sim N(gt/6, 1) + z_i$, $n_{gt} = 200$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9764	0.1586	0.1548	0.1389	0.1403	0.9765	0.1581
		(0.1533)	(0.0116)	(0.0325)	(0.0425)	(0.0621)	(0.1528)	(0.0116)
	x_3	1.0024	0.0887	0.0865	0.0825	0.0874	1.0027	0.0884
		(0.0902)	(0.0038)	(0.0175)	(0.0195)	(0.0315)	(0.0903)	(0.0038)
	x_4	1.0051	0.2850	0.2779	0.2626	0.2944	1.0055	0.2840
		(0.2823)	(0.0131)	(0.0569)	(0.0663)	(0.1072)	(0.2839)	(0.0131)
IV: z	x_2	1.0018	0.0463	0.0456	0.0508	0.0499	1.0020	0.0452
		(0.0460)	(0.0012)	(0.0055)	(0.0077)	(0.0154)	(0.0462)	(0.0011)
	x_3	0.9980	0.0826	0.0813	0.0667	0.0713	0.9976	0.0816
		(0.0841)	(0.0027)	(0.0098)	(0.0150)	(0.0257)	(0.0850)	(0.0026)
	x_4	1.0087	0.2766	0.2723	0.2202	0.2462	1.0080	0.2734
		(0.2738)	(0.0118)	(0.0341)	(0.0495)	(0.0869)	(0.2751)	(0.0116)
IV: x_2	x_2	1.0042	0.0440	0.0487	0.0489	0.0422	1.0041	0.0329
		(0.0431)	(0.0020)	(0.0088)	(0.0092)	(0.0154)	(0.0331)	(0.0007)
	x_3	0.9967	0.1092	0.0679	0.0707	0.0881	0.9978	0.0815
		(0.1076)	(0.0046)	(0.0126)	(0.0175)	(0.0335)	(0.0846)	(0.0026)
	x_4	0.9932	0.3584	0.1885	0.2042	0.2764	1.0090	0.2734
		(0.3591)	(0.0192)	(0.0365)	(0.0643)	(0.1203)	(0.2779)	(0.0117)
IV: x_3	x_2	0.9840	0.1633	0.1215	0.1104	0.1236	0.9850	0.1423
		(0.1582)	(0.0111)	(0.0189)	(0.0257)	(0.0529)	(0.1420)	(0.0091)
	x_3	0.9923	0.0464	0.0353	0.0367	0.0403	0.9919	0.0416
		(0.0460)	(0.0014)	(0.0049)	(0.0064)	(0.0137)	(0.0421)	(0.0011)
	x_4	0.9989	0.2941	0.2333	0.2192	0.2552	1.0189	0.2625
		(0.2890)	(0.0129)	(0.0332)	(0.0421)	(0.0917)	(0.2617)	(0.0107)
IV: x_4	x_2	0.9740	0.1582	0.1002	0.1004	0.1162	0.9828	0.1474
		(0.1543)	(0.0113)	(0.0173)	(0.0270)	(0.0502)	(0.1441)	(0.0096)
	x_3	1.0025	0.0863	0.0541	0.0563	0.0689	1.0010	0.0812
		(0.0862)	(0.0034)	(0.0087)	(0.0119)	(0.0252)	(0.0810)	(0.0030)
	x_4	1.0045	0.1586	0.1270	0.1303	0.1406	1.0111	0.1011
		(0.1553)	(0.0036)	(0.0205)	(0.0269)	(0.0490)	(0.1014)	(0.0017)
IV: all	x_2	1.0052	0.0404	0.0296	0.0335	0.0379	1.0043	0.0322
		(0.0399)	(0.0013)	(0.0037)	(0.0051)	(0.0126)	(0.0333)	(0.0007)
	x_3	0.9925	0.0565	0.0312	0.0348	0.0458	0.9914	0.0397
		(0.0559)	(0.0017)	(0.0039)	(0.0058)	(0.0160)	(0.0414)	(0.0008)
	x_4	0.9959	0.2653	0.1211	0.1423	0.2010	1.0133	0.0985
		(0.2635)	(0.0101)	(0.0157)	(0.0383)	(0.0807)	(0.1012)	(0.0015)

Table B.10 Small panel with $G = 6$, $T = 4$. Case 3.b: $x_{2it} \sim N(gt/6, 1) + z_i$, $n_{gt} = 1000$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9937	0.0720	0.0698	0.0627	0.0639	0.9936	0.0720
		(0.0714)	(0.0023)	(0.0144)	(0.0181)	(0.0283)	(0.0713)	(0.0023)
	x_3	1.0038	0.0398	0.0385	0.0366	0.0391	1.0038	0.0397
		(0.0391)	(0.0007)	(0.0079)	(0.0085)	(0.0146)	(0.0391)	(0.0007)
	x_4	1.0066	0.1283	0.1244	0.1174	0.1332	1.0068	0.1282
		(0.1273)	(0.0025)	(0.0256)	(0.0308)	(0.0487)	(0.1276)	(0.0025)
IV: z	x_2	0.9982	0.0206	0.0205	0.0229	0.0223	0.9982	0.0205
		(0.0191)	(0.0002)	(0.0025)	(0.0034)	(0.0073)	(0.0191)	(0.0002)
	x_3	1.0029	0.0372	0.0371	0.0295	0.0317	1.0029	0.0371
		(0.0356)	(0.0005)	(0.0045)	(0.0067)	(0.0121)	(0.0357)	(0.0005)
	x_4	1.0077	0.1261	0.1258	0.0982	0.1118	1.0080	0.1259
		(0.1244)	(0.0023)	(0.0154)	(0.0243)	(0.0403)	(0.1249)	(0.0023)
IV: x_2	x_2	0.9964	0.0192	0.0218	0.0217	0.0189	0.9987	0.0148
		(0.0191)	(0.0004)	(0.0039)	(0.0040)	(0.0069)	(0.0143)	(0.0001)
	x_3	1.0049	0.0489	0.0307	0.0320	0.0406	1.0027	0.0370
		(0.0489)	(0.0009)	(0.0055)	(0.0077)	(0.0153)	(0.0355)	(0.0005)
	x_4	1.0179	0.1624	0.0859	0.0921	0.1246	1.0084	0.1258
		(0.1651)	(0.0036)	(0.0154)	(0.0290)	(0.0553)	(0.1239)	(0.0023)
IV: x_3	x_2	0.9995	0.0763	0.0562	0.0510	0.0570	0.9945	0.0675
		(0.0773)	(0.0025)	(0.0078)	(0.0109)	(0.0245)	(0.0654)	(0.0021)
	x_3	1.0026	0.0207	0.0158	0.0165	0.0180	1.0031	0.0188
		(0.0206)	(0.0003)	(0.0022)	(0.0028)	(0.0065)	(0.0186)	(0.0002)
	x_4	1.0097	0.1338	0.1059	0.0992	0.1159	1.0070	0.1202
		(0.1327)	(0.0025)	(0.0147)	(0.0195)	(0.0408)	(0.1218)	(0.0020)
IV: x_4	x_2	0.9956	0.0726	0.0456	0.0457	0.0533	0.9933	0.0703
		(0.0710)	(0.0023)	(0.0075)	(0.0114)	(0.0236)	(0.0692)	(0.0022)
	x_3	1.0041	0.0387	0.0242	0.0251	0.0314	1.0046	0.0369
		(0.0382)	(0.0007)	(0.0040)	(0.0054)	(0.0117)	(0.0371)	(0.0006)
	x_4	1.0028	0.0705	0.0567	0.0580	0.0634	1.0011	0.0454
		(0.0686)	(0.0007)	(0.0093)	(0.0121)	(0.0219)	(0.0442)	(0.0003)
IV: all	x_2	0.9979	0.0178	0.0133	0.0150	0.0168	0.9986	0.0147
		(0.0171)	(0.0003)	(0.0016)	(0.0022)	(0.0058)	(0.0144)	(0.0001)
	x_3	1.0037	0.0252	0.0141	0.0158	0.0208	1.0031	0.0182
		(0.0252)	(0.0004)	(0.0017)	(0.0026)	(0.0076)	(0.0179)	(0.0002)
	x_4	1.0120	0.1203	0.0553	0.0646	0.0910	1.0014	0.0448
		(0.1207)	(0.0019)	(0.0067)	(0.0176)	(0.0377)	(0.0440)	(0.0003)

Table B.11 Small panel with $G = 6$, $T = 4$. Case 3.1: $x_{2it} \sim N(gt/6, 1) + z_i$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9909	0.1590	0.1552	0.1393	0.1423	0.9911	0.1585
		(0.1623)	(0.0117)	(0.0349)	(0.0425)	(0.0603)	(0.1623)	(0.0117)
	x_3	1.0079	0.0890	0.0868	0.0825	0.0873	1.0077	0.0887
		(0.0870)	(0.0037)	(0.0185)	(0.0210)	(0.0329)	(0.0870)	(0.0037)
	x_4	1.0113	0.2868	0.2798	0.2625	0.2991	1.0123	0.2858
		(0.2880)	(0.0127)	(0.0594)	(0.0714)	(0.1095)	(0.2877)	(0.0127)
IV: z	x_2	0.9980	0.0463	0.0457	0.0508	0.0496	0.9979	0.0452
		(0.0430)	(0.0012)	(0.0054)	(0.0073)	(0.0159)	(0.0437)	(0.0012)
	x_3	1.0070	0.0827	0.0820	0.0670	0.0704	1.0065	0.0818
		(0.0804)	(0.0027)	(0.0097)	(0.0159)	(0.0269)	(0.0803)	(0.0027)
	x_4	1.0131	0.2780	0.2757	0.2210	0.2490	1.0138	0.2749
		(0.2794)	(0.0115)	(0.0336)	(0.0535)	(0.0888)	(0.2780)	(0.0113)
IV: x_2	x_2	0.9965	0.0436	0.0488	0.0490	0.0431	0.9981	0.0327
		(0.0421)	(0.0020)	(0.0092)	(0.0094)	(0.0154)	(0.0324)	(0.0007)
	x_3	1.0109	0.1090	0.0688	0.0712	0.0886	1.0064	0.0816
		(0.1051)	(0.0048)	(0.0129)	(0.0177)	(0.0343)	(0.0801)	(0.0026)
	x_4	1.0141	0.3598	0.1914	0.2046	0.2767	1.0140	0.2749
		(0.3642)	(0.0184)	(0.0358)	(0.0669)	(0.1251)	(0.2797)	(0.0113)
IV: x_3	x_2	1.0028	0.1633	0.1213	0.1105	0.1241	0.9952	0.1426
		(0.1629)	(0.0111)	(0.0187)	(0.0252)	(0.0517)	(0.1465)	(0.0092)
	x_3	1.0036	0.0465	0.0353	0.0367	0.0400	1.0034	0.0417
		(0.0467)	(0.0014)	(0.0049)	(0.0061)	(0.0140)	(0.0433)	(0.0010)
	x_4	1.0203	0.2961	0.2343	0.2202	0.2588	1.0180	0.2638
		(0.2919)	(0.0125)	(0.0328)	(0.0442)	(0.0881)	(0.2629)	(0.0103)
IV: x_4	x_2	0.9917	0.1583	0.1006	0.1008	0.1172	0.9920	0.1477
		(0.1605)	(0.0112)	(0.0182)	(0.0264)	(0.0498)	(0.1498)	(0.0098)
	x_3	1.0102	0.0865	0.0545	0.0564	0.0695	1.0088	0.0813
		(0.0836)	(0.0033)	(0.0091)	(0.0128)	(0.0268)	(0.0795)	(0.0029)
	x_4	1.0034	0.1584	0.1277	0.1303	0.1417	1.0010	0.1010
		(0.1576)	(0.0037)	(0.0210)	(0.0278)	(0.0498)	(0.0992)	(0.0017)
IV: all	x_2	0.9980	0.0402	0.0296	0.0335	0.0379	0.9981	0.0321
		(0.0376)	(0.0013)	(0.0037)	(0.0050)	(0.0125)	(0.0330)	(0.0007)
	x_3	1.0071	0.0564	0.0314	0.0350	0.0458	1.0036	0.0397
		(0.0550)	(0.0018)	(0.0039)	(0.0059)	(0.0171)	(0.0421)	(0.0008)
	x_4	1.0122	0.2656	0.1223	0.1432	0.2011	1.0013	0.0984
		(0.2666)	(0.0100)	(0.0153)	(0.0408)	(0.0858)	(0.1007)	(0.0015)

Table B.12 Small panel with $G = 6$, $T = 4$. Case 3.1: $x_{2it} \sim N(gt/6, 1) + z_i$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0027	0.0720	0.0695	0.0618	0.0623	1.0027	0.0720
		(0.0721)	(0.0023)	(0.0146)	(0.0182)	(0.0281)	(0.0722)	(0.0023)
	x_3	0.9971	0.0398	0.0384	0.0366	0.0390	0.9972	0.0398
		(0.0401)	(0.0008)	(0.0080)	(0.0086)	(0.0138)	(0.0401)	(0.0008)
	x_4	1.0010	0.1280	0.1235	0.1166	0.1324	1.0012	0.1279
		(0.1305)	(0.0025)	(0.0256)	(0.0305)	(0.0489)	(0.1304)	(0.0025)
IV: z	x_2	1.0012	0.0206	0.0203	0.0227	0.0222	1.0012	0.0205
		(0.0196)	(0.0002)	(0.0024)	(0.0034)	(0.0073)	(0.0197)	(0.0002)
	x_3	0.9975	0.0373	0.0369	0.0295	0.0320	0.9976	0.0372
		(0.0372)	(0.0005)	(0.0045)	(0.0069)	(0.0115)	(0.0372)	(0.0005)
	x_4	1.0004	0.1258	0.1245	0.0977	0.1110	1.0005	0.1256
		(0.1286)	(0.0023)	(0.0150)	(0.0240)	(0.0398)	(0.1283)	(0.0023)
IV: x_2	x_2	1.0005	0.0192	0.0216	0.0216	0.0189	1.0000	0.0148
		(0.0186)	(0.0004)	(0.0041)	(0.0040)	(0.0069)	(0.0145)	(0.0001)
	x_3	0.9963	0.0490	0.0305	0.0317	0.0389	0.9978	0.0371
		(0.0476)	(0.0010)	(0.0057)	(0.0081)	(0.0149)	(0.0372)	(0.0005)
	x_4	1.0004	0.1621	0.0849	0.0912	0.1240	1.0000	0.1255
		(0.1665)	(0.0038)	(0.0161)	(0.0305)	(0.0578)	(0.1274)	(0.0023)
IV: x_3	x_2	1.0004	0.0765	0.0563	0.0511	0.0573	1.0003	0.0675
		(0.0751)	(0.0024)	(0.0076)	(0.0113)	(0.0254)	(0.0673)	(0.0019)
	x_3	0.9989	0.0207	0.0158	0.0165	0.0180	0.9998	0.0188
		(0.0211)	(0.0003)	(0.0021)	(0.0026)	(0.0060)	(0.0190)	(0.0002)
	x_4	1.0005	0.1335	0.1057	0.0993	0.1158	0.9980	0.1199
		(0.1350)	(0.0025)	(0.0139)	(0.0182)	(0.0403)	(0.1212)	(0.0021)
IV: x_4	x_2	1.0028	0.0727	0.0455	0.0451	0.0518	1.0026	0.0703
		(0.0730)	(0.0023)	(0.0075)	(0.0115)	(0.0233)	(0.0699)	(0.0022)
	x_3	0.9965	0.0388	0.0242	0.0251	0.0310	0.9971	0.0370
		(0.0391)	(0.0007)	(0.0039)	(0.0053)	(0.0116)	(0.0369)	(0.0006)
	x_4	1.0027	0.0705	0.0564	0.0579	0.0628	1.0032	0.0454
		(0.0713)	(0.0007)	(0.0091)	(0.0120)	(0.0221)	(0.0461)	(0.0003)
IV: all	x_2	1.0004	0.0178	0.0132	0.0149	0.0170	1.0000	0.0147
		(0.0172)	(0.0003)	(0.0016)	(0.0022)	(0.0057)	(0.0145)	(0.0001)
	x_3	0.9983	0.0252	0.0140	0.0156	0.0203	0.9998	0.0182
		(0.0251)	(0.0004)	(0.0017)	(0.0025)	(0.0072)	(0.0183)	(0.0002)
	x_4	1.0008	0.1200	0.0549	0.0640	0.0902	1.0026	0.0449
		(0.1228)	(0.0020)	(0.0066)	(0.0181)	(0.0386)	(0.0460)	(0.0003)

Table B.13 Small panel with $G = 6$, $T = 4$. Case 4.a: $x_{2it} \sim N(gt/6, 1) + z_i + f_i$, $n_{gt} = 200$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0020	0.1570	0.1533	0.1380	0.1399	1.0020	0.1565
		(0.1521)	(0.0139)	(0.0329)	(0.0425)	(0.0620)	(0.1516)	(0.0138)
	x_3	0.9972	0.0888	0.0866	0.0825	0.0875	0.9975	0.0884
		(0.0901)	(0.0043)	(0.0176)	(0.0196)	(0.0314)	(0.0903)	(0.0043)
	x_4	1.0131	0.2854	0.2783	0.2628	0.2947	1.0133	0.2844
		(0.2825)	(0.0140)	(0.0571)	(0.0662)	(0.1072)	(0.2839)	(0.0139)
IV: z	x_2	1.0084	0.0460	0.0454	0.0505	0.0495	1.0086	0.0450
		(0.0456)	(0.0014)	(0.0055)	(0.0078)	(0.0154)	(0.0459)	(0.0013)
	x_3	0.9967	0.0825	0.0813	0.0667	0.0712	0.9963	0.0816
		(0.0841)	(0.0027)	(0.0098)	(0.0150)	(0.0257)	(0.0850)	(0.0027)
	x_4	1.0104	0.2764	0.2722	0.2201	0.2462	1.0097	0.2732
		(0.2735)	(0.0119)	(0.0342)	(0.0495)	(0.0869)	(0.2747)	(0.0117)
IV: x_2	x_2	1.2218	0.0954	0.0741	0.0775	0.0979	1.3252	0.0267
		(0.0963)	(0.0072)	(0.0173)	(0.0191)	(0.0426)	(0.0276)	(0.0006)
	x_3	1.0029	0.1121	0.0529	0.0726	0.0728	0.9357	0.0803
		(0.1114)	(0.0048)	(0.0122)	(0.0237)	(0.0311)	(0.0841)	(0.0026)
	x_4	0.9978	0.3470	0.1676	0.2207	0.2681	1.0932	0.2699
		(0.3442)	(0.0185)	(0.0396)	(0.0705)	(0.1154)	(0.2715)	(0.0115)
IV: x_3	x_2	1.0304	0.1585	0.1187	0.1090	0.1223	1.0463	0.1377
		(0.1521)	(0.0128)	(0.0190)	(0.0251)	(0.0512)	(0.1369)	(0.0107)
	x_3	0.9880	0.0463	0.0352	0.0367	0.0405	0.9887	0.0415
		(0.0459)	(0.0016)	(0.0049)	(0.0064)	(0.0137)	(0.0420)	(0.0012)
	x_4	1.0083	0.2938	0.2335	0.2194	0.2572	1.0252	0.2622
		(0.2888)	(0.0135)	(0.0333)	(0.0422)	(0.0922)	(0.2608)	(0.0112)
IV: x_4	x_2	1.0055	0.1557	0.0990	0.0995	0.1154	1.0488	0.1426
		(0.1523)	(0.0133)	(0.0177)	(0.0271)	(0.0500)	(0.1402)	(0.0110)
	x_3	0.9975	0.0863	0.0542	0.0564	0.0691	0.9894	0.0811
		(0.0861)	(0.0038)	(0.0088)	(0.0120)	(0.0252)	(0.0805)	(0.0033)
	x_4	1.0098	0.1587	0.1272	0.1304	0.1410	1.0140	0.1013
		(0.1552)	(0.0043)	(0.0206)	(0.0268)	(0.0490)	(0.1008)	(0.0023)
IV: all	x_2	1.1804	0.0713	0.0412	0.0544	0.0698	1.3193	0.0263
		(0.0720)	(0.0056)	(0.0074)	(0.0116)	(0.0279)	(0.0285)	(0.0006)
	x_3	0.9944	0.0897	0.0309	0.0520	0.0576	0.9766	0.0393
		(0.0890)	(0.0033)	(0.0054)	(0.0160)	(0.0234)	(0.0416)	(0.0009)
	x_4	0.9829	0.3207	0.1070	0.1838	0.2289	1.0171	0.0975
		(0.3180)	(0.0151)	(0.0196)	(0.0584)	(0.0978)	(0.0999)	(0.0015)

Table B.14 Small panel with $G = 6$, $T = 4$. Case 4.b: $x_{2it} \sim N(gt/6, 1) + z_i + f_i$, $n_{gt} = 1000$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9989	0.0719	0.0697	0.0627	0.0642	0.9989	0.0719
		(0.0716)	(0.0029)	(0.0145)	(0.0181)	(0.0284)	(0.0715)	(0.0029)
	x_3	1.0027	0.0398	0.0385	0.0366	0.0391	1.0028	0.0397
		(0.0391)	(0.0009)	(0.0079)	(0.0085)	(0.0146)	(0.0392)	(0.0009)
	x_4	1.0081	0.1283	0.1244	0.1174	0.1332	1.0083	0.1282
		(0.1274)	(0.0027)	(0.0256)	(0.0309)	(0.0488)	(0.1277)	(0.0027)
IV: z	x_2	0.9996	0.0206	0.0205	0.0229	0.0223	0.9996	0.0205
		(0.0191)	(0.0003)	(0.0025)	(0.0034)	(0.0073)	(0.0191)	(0.0003)
	x_3	1.0026	0.0372	0.0371	0.0295	0.0317	1.0026	0.0371
		(0.0356)	(0.0005)	(0.0045)	(0.0067)	(0.0121)	(0.0357)	(0.0005)
	x_4	1.0082	0.1261	0.1258	0.0982	0.1118	1.0085	0.1259
		(0.1244)	(0.0023)	(0.0154)	(0.0243)	(0.0403)	(0.1249)	(0.0023)
IV: x_2	x_2	1.2166	0.0421	0.0405	0.0458	0.0711	1.3220	0.0120
		(0.0423)	(0.0014)	(0.0082)	(0.0106)	(0.0254)	(0.0120)	(0.0001)
	x_3	1.0113	0.0500	0.0286	0.0357	0.0402	0.9385	0.0365
		(0.0491)	(0.0009)	(0.0057)	(0.0104)	(0.0145)	(0.0353)	(0.0005)
	x_4	1.0149	0.1559	0.0910	0.1077	0.1274	1.1027	0.1241
		(0.1568)	(0.0035)	(0.0182)	(0.0325)	(0.0546)	(0.1224)	(0.0023)
IV: x_3	x_2	1.0097	0.0758	0.0559	0.0509	0.0569	1.0091	0.0671
		(0.0771)	(0.0030)	(0.0079)	(0.0110)	(0.0243)	(0.0653)	(0.0025)
	x_3	1.0017	0.0207	0.0158	0.0165	0.0181	1.0024	0.0188
		(0.0206)	(0.0003)	(0.0022)	(0.0028)	(0.0065)	(0.0186)	(0.0003)
	x_4	1.0115	0.1337	0.1058	0.0991	0.1159	1.0087	0.1201
		(0.1325)	(0.0026)	(0.0147)	(0.0195)	(0.0408)	(0.1216)	(0.0022)
IV: x_4	x_2	1.0024	0.0724	0.0455	0.0457	0.0536	1.0090	0.0698
		(0.0710)	(0.0029)	(0.0076)	(0.0114)	(0.0236)	(0.0689)	(0.0027)
	x_3	1.0031	0.0387	0.0242	0.0251	0.0314	1.0019	0.0369
		(0.0382)	(0.0008)	(0.0040)	(0.0053)	(0.0117)	(0.0371)	(0.0007)
	x_4	1.0038	0.0705	0.0567	0.0580	0.0635	1.0017	0.0454
		(0.0687)	(0.0008)	(0.0093)	(0.0121)	(0.0219)	(0.0442)	(0.0004)
IV: all	x_2	1.1734	0.0314	0.0235	0.0345	0.0533	1.3182	0.0120
		(0.0317)	(0.0011)	(0.0031)	(0.0060)	(0.0158)	(0.0121)	(0.0001)
	x_3	1.0038	0.0401	0.0174	0.0252	0.0304	0.9883	0.0180
		(0.0395)	(0.0006)	(0.0023)	(0.0070)	(0.0111)	(0.0182)	(0.0002)
	x_4	0.9986	0.1446	0.0608	0.0877	0.1063	1.0073	0.0444
		(0.1451)	(0.0028)	(0.0081)	(0.0268)	(0.0455)	(0.0444)	(0.0003)

Table B.15 Small panel with $G = 6$, $T = 4$. Case 4.1: $x_{2it} \sim N(gt/6, 1) + z_i + f_i$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0153	0.1575	0.1537	0.1388	0.1440	1.0155	0.1569
		(0.1599)	(0.0143)	(0.0356)	(0.0427)	(0.0631)	(0.1598)	(0.0142)
	x_3	1.0031	0.0890	0.0868	0.0825	0.0876	1.0029	0.0887
		(0.0867)	(0.0043)	(0.0187)	(0.0211)	(0.0331)	(0.0866)	(0.0043)
	x_4	1.0184	0.2869	0.2799	0.2628	0.2990	1.0194	0.2859
		(0.2879)	(0.0135)	(0.0597)	(0.0718)	(0.1099)	(0.2877)	(0.0135)
IV: z	x_2	1.0048	0.0462	0.0455	0.0506	0.0494	1.0048	0.0451
		(0.0431)	(0.0014)	(0.0054)	(0.0074)	(0.0159)	(0.0437)	(0.0014)
	x_3	1.0057	0.0827	0.0820	0.0670	0.0704	1.0052	0.0818
		(0.0804)	(0.0027)	(0.0097)	(0.0159)	(0.0270)	(0.0803)	(0.0027)
	x_4	1.0149	0.2779	0.2756	0.2210	0.2488	1.0157	0.2748
		(0.2792)	(0.0115)	(0.0336)	(0.0535)	(0.0889)	(0.2778)	(0.0113)
IV: x_2	x_2	1.2218	0.0951	0.0728	0.0752	0.0947	1.3194	0.0266
		(0.0947)	(0.0071)	(0.0188)	(0.0191)	(0.0420)	(0.0266)	(0.0006)
	x_3	1.0175	0.1119	0.0522	0.0715	0.0722	0.9439	0.0806
		(0.1094)	(0.0048)	(0.0129)	(0.0243)	(0.0320)	(0.0790)	(0.0026)
	x_4	1.0190	0.3481	0.1655	0.2170	0.2634	1.1004	0.2714
		(0.3531)	(0.0175)	(0.0410)	(0.0755)	(0.1217)	(0.2737)	(0.0112)
IV: x_3	x_2	1.0478	0.1585	0.1185	0.1092	0.1239	1.0572	0.1379
		(0.1584)	(0.0127)	(0.0191)	(0.0256)	(0.0524)	(0.1415)	(0.0106)
	x_3	0.9997	0.0463	0.0352	0.0366	0.0400	1.0003	0.0415
		(0.0465)	(0.0016)	(0.0049)	(0.0061)	(0.0139)	(0.0431)	(0.0012)
	x_4	1.0287	0.2953	0.2339	0.2199	0.2586	1.0252	0.2631
		(0.2908)	(0.0130)	(0.0328)	(0.0441)	(0.0884)	(0.2615)	(0.0108)
IV: x_4	x_2	1.0224	0.1557	0.0994	0.1000	0.1179	1.0573	0.1430
		(0.1568)	(0.0136)	(0.0187)	(0.0266)	(0.0514)	(0.1437)	(0.0116)
	x_3	1.0056	0.0864	0.0545	0.0565	0.0698	0.9976	0.0811
		(0.0832)	(0.0038)	(0.0092)	(0.0128)	(0.0270)	(0.0789)	(0.0033)
	x_4	1.0083	0.1583	0.1278	0.1303	0.1418	1.0035	0.1010
		(0.1572)	(0.0044)	(0.0212)	(0.0279)	(0.0499)	(0.0988)	(0.0022)
IV: all	x_2	1.1794	0.0710	0.0407	0.0531	0.0677	1.3142	0.0262
		(0.0707)	(0.0057)	(0.0078)	(0.0114)	(0.0273)	(0.0275)	(0.0006)
	x_3	1.0090	0.0896	0.0306	0.0514	0.0573	0.9894	0.0393
		(0.0875)	(0.0032)	(0.0057)	(0.0163)	(0.0243)	(0.0424)	(0.0009)
	x_4	1.0036	0.3215	0.1062	0.1814	0.2260	1.0073	0.0974
		(0.3253)	(0.0143)	(0.0198)	(0.0624)	(0.1035)	(0.1018)	(0.0015)

Table B.16 Small panel with $G = 6$, $T = 4$. Case 4.1: $x_{2it} \sim N(gt/6, 1) + z_i + f_i$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0079	0.0718	0.0693	0.0615	0.0620	1.0079	0.0717
		(0.0721)	(0.0028)	(0.0146)	(0.0181)	(0.0280)	(0.0721)	(0.0028)
	x_3	0.9961	0.0398	0.0384	0.0365	0.0390	0.9961	0.0398
		(0.0401)	(0.0009)	(0.0079)	(0.0086)	(0.0138)	(0.0400)	(0.0009)
	x_4	1.0026	0.1279	0.1235	0.1165	0.1321	1.0027	0.1279
		(0.1305)	(0.0027)	(0.0255)	(0.0305)	(0.0488)	(0.1304)	(0.0027)
IV: z	x_2	1.0025	0.0205	0.0203	0.0226	0.0221	1.0026	0.0204
		(0.0196)	(0.0003)	(0.0024)	(0.0034)	(0.0073)	(0.0196)	(0.0003)
	x_3	0.9972	0.0373	0.0369	0.0295	0.0320	0.9973	0.0372
		(0.0372)	(0.0006)	(0.0045)	(0.0069)	(0.0115)	(0.0372)	(0.0006)
	x_4	1.0008	0.1258	0.1244	0.0977	0.1109	1.0009	0.1255
		(0.1285)	(0.0023)	(0.0150)	(0.0240)	(0.0398)	(0.1283)	(0.0023)
IV: x_2	x_2	1.2219	0.0422	0.0397	0.0445	0.0677	1.3238	0.0120
		(0.0420)	(0.0015)	(0.0079)	(0.0105)	(0.0250)	(0.0121)	(0.0001)
	x_3	1.0015	0.0501	0.0281	0.0352	0.0389	0.9336	0.0366
		(0.0490)	(0.0010)	(0.0056)	(0.0104)	(0.0138)	(0.0367)	(0.0005)
	x_4	0.9986	0.1561	0.0895	0.1057	0.1275	1.0948	0.1238
		(0.1587)	(0.0036)	(0.0179)	(0.0324)	(0.0540)	(0.1257)	(0.0023)
IV: x_3	x_2	1.0106	0.0759	0.0560	0.0509	0.0570	1.0148	0.0669
		(0.0747)	(0.0029)	(0.0076)	(0.0111)	(0.0251)	(0.0670)	(0.0024)
	x_3	0.9980	0.0207	0.0158	0.0165	0.0180	0.9991	0.0188
		(0.0211)	(0.0003)	(0.0021)	(0.0026)	(0.0060)	(0.0189)	(0.0003)
	x_4	1.0024	0.1335	0.1057	0.0992	0.1158	0.9997	0.1198
		(0.1349)	(0.0027)	(0.0139)	(0.0181)	(0.0402)	(0.1211)	(0.0023)
IV: x_4	x_2	1.0094	0.0723	0.0453	0.0449	0.0515	1.0185	0.0697
		(0.0729)	(0.0028)	(0.0075)	(0.0115)	(0.0233)	(0.0696)	(0.0026)
	x_3	0.9955	0.0388	0.0241	0.0251	0.0310	0.9944	0.0370
		(0.0391)	(0.0008)	(0.0039)	(0.0053)	(0.0116)	(0.0368)	(0.0007)
	x_4	1.0038	0.0705	0.0564	0.0578	0.0627	1.0038	0.0454
		(0.0713)	(0.0009)	(0.0091)	(0.0119)	(0.0220)	(0.0461)	(0.0005)
IV: all	x_2	1.1771	0.0316	0.0232	0.0336	0.0513	1.3198	0.0120
		(0.0312)	(0.0012)	(0.0030)	(0.0058)	(0.0157)	(0.0123)	(0.0001)
	x_3	0.9954	0.0402	0.0172	0.0248	0.0294	0.9849	0.0180
		(0.0395)	(0.0007)	(0.0022)	(0.0071)	(0.0103)	(0.0185)	(0.0002)
	x_4	0.9840	0.1447	0.0601	0.0861	0.1059	1.0069	0.0444
		(0.1473)	(0.0029)	(0.0078)	(0.0270)	(0.0456)	(0.0458)	(0.0003)

Table B.17 Small panel with $G = 6$, $T = 4$. Case 5.a: $x_{2it} \sim N(gt/2, 1) + z_i + f_i$, $n_{gt} = 200$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9951	0.0539	0.0527	0.0470	0.0472	0.9951	0.0537
		(0.0524)	(0.0021)	(0.0106)	(0.0142)	(0.0211)	(0.0522)	(0.0021)
	x_3	1.0007	0.0887	0.0866	0.0825	0.0875	1.0010	0.0884
		(0.0906)	(0.0033)	(0.0175)	(0.0194)	(0.0316)	(0.0908)	(0.0032)
	x_4	1.0074	0.2843	0.2777	0.2625	0.2945	1.0078	0.2834
		(0.2817)	(0.0129)	(0.0569)	(0.0661)	(0.1070)	(0.2834)	(0.0128)
IV: z	x_2	1.0029	0.0359	0.0354	0.0358	0.0352	1.0030	0.0352
		(0.0360)	(0.0010)	(0.0043)	(0.0053)	(0.0117)	(0.0361)	(0.0010)
	x_3	0.9966	0.0847	0.0833	0.0687	0.0727	0.9962	0.0836
		(0.0863)	(0.0028)	(0.0101)	(0.0143)	(0.0256)	(0.0871)	(0.0028)
	x_4	1.0107	0.2777	0.2733	0.2200	0.2460	1.0099	0.2744
		(0.2754)	(0.0120)	(0.0345)	(0.0494)	(0.0865)	(0.2765)	(0.0118)
IV: x_2	x_2	1.0240	0.0787	0.0380	0.0456	0.0469	1.2598	0.0247
		(0.0782)	(0.0044)	(0.0098)	(0.0120)	(0.0193)	(0.0252)	(0.0007)
	x_3	0.9991	0.1118	0.0487	0.0692	0.0668	0.8485	0.0824
		(0.1104)	(0.0048)	(0.0126)	(0.0237)	(0.0319)	(0.0867)	(0.0028)
	x_4	0.9839	0.3461	0.1403	0.2007	0.2496	1.2129	0.2737
		(0.3425)	(0.0188)	(0.0369)	(0.0720)	(0.1127)	(0.2806)	(0.0119)
IV: x_3	x_2	1.0004	0.0572	0.0422	0.0381	0.0426	1.0026	0.0502
		(0.0552)	(0.0022)	(0.0059)	(0.0087)	(0.0187)	(0.0496)	(0.0018)
	x_3	0.9908	0.0463	0.0353	0.0368	0.0405	0.9907	0.0415
		(0.0461)	(0.0013)	(0.0049)	(0.0063)	(0.0136)	(0.0422)	(0.0010)
	x_4	1.0021	0.2932	0.2331	0.2190	0.2554	1.0212	0.2617
		(0.2886)	(0.0129)	(0.0332)	(0.0420)	(0.0911)	(0.2612)	(0.0107)
IV: x_4	x_2	0.9950	0.0544	0.0343	0.0343	0.0395	1.0018	0.0522
		(0.0534)	(0.0022)	(0.0056)	(0.0091)	(0.0173)	(0.0515)	(0.0020)
	x_3	1.0009	0.0863	0.0541	0.0563	0.0689	0.9971	0.0815
		(0.0867)	(0.0029)	(0.0087)	(0.0119)	(0.0253)	(0.0818)	(0.0027)
	x_4	1.0061	0.1583	0.1269	0.1303	0.1407	1.0121	0.1008
		(0.1549)	(0.0036)	(0.0205)	(0.0269)	(0.0491)	(0.1011)	(0.0017)
IV: all	x_2	1.0226	0.0745	0.0227	0.0388	0.0409	1.2477	0.0240
		(0.0740)	(0.0040)	(0.0052)	(0.0101)	(0.0166)	(0.0255)	(0.0006)
	x_3	0.9952	0.1014	0.0284	0.0565	0.0575	0.9575	0.0399
		(0.1000)	(0.0041)	(0.0065)	(0.0189)	(0.0265)	(0.0421)	(0.0009)
	x_4	0.9774	0.3369	0.0857	0.1775	0.2239	1.0247	0.0986
		(0.3333)	(0.0176)	(0.0199)	(0.0634)	(0.1009)	(0.1011)	(0.0016)

Table B.18 Small panel with $G = 6$, $T = 4$. Case 5.b: $x_{2it} \sim N(gt/2, 1) + z_i + f_i$, $n_{gt} = 1000$, sampling rate= 1%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9985	0.0241	0.0233	0.0209	0.0213	0.9985	0.0241
		(0.0239)	(0.0004)	(0.0048)	(0.0060)	(0.0095)	(0.0239)	(0.0004)
	x_3	1.0034	0.0397	0.0385	0.0366	0.0390	1.0034	0.0397
		(0.0391)	(0.0006)	(0.0079)	(0.0085)	(0.0146)	(0.0391)	(0.0006)
	x_4	1.0072	0.1282	0.1244	0.1174	0.1333	1.0074	0.1281
		(0.1273)	(0.0024)	(0.0256)	(0.0308)	(0.0486)	(0.1276)	(0.0024)
IV: z	x_2	0.9991	0.0160	0.0160	0.0161	0.0157	0.9992	0.0160
		(0.0150)	(0.0002)	(0.0019)	(0.0022)	(0.0052)	(0.0151)	(0.0002)
	x_3	1.0031	0.0382	0.0380	0.0304	0.0324	1.0030	0.0381
		(0.0368)	(0.0006)	(0.0047)	(0.0065)	(0.0120)	(0.0369)	(0.0006)
	x_4	1.0075	0.1268	0.1264	0.0981	0.1116	1.0078	0.1265
		(0.1251)	(0.0023)	(0.0156)	(0.0243)	(0.0404)	(0.1256)	(0.0023)
IV: x_2	x_2	1.0251	0.0351	0.0193	0.0217	0.0229	1.2599	0.0111
		(0.0355)	(0.0008)	(0.0041)	(0.0052)	(0.0093)	(0.0109)	(0.0001)
	x_3	1.0089	0.0498	0.0247	0.0312	0.0307	0.8485	0.0374
		(0.0489)	(0.0009)	(0.0052)	(0.0104)	(0.0149)	(0.0368)	(0.0006)
	x_4	1.0059	0.1552	0.0715	0.0902	0.1118	1.2361	0.1259
		(0.1551)	(0.0035)	(0.0150)	(0.0330)	(0.0517)	(0.1265)	(0.0023)
IV: x_3	x_2	1.0010	0.0257	0.0189	0.0171	0.0191	0.9998	0.0227
		(0.0260)	(0.0005)	(0.0026)	(0.0037)	(0.0083)	(0.0221)	(0.0004)
	x_3	1.0023	0.0207	0.0158	0.0165	0.0180	1.0028	0.0188
		(0.0206)	(0.0002)	(0.0022)	(0.0028)	(0.0065)	(0.0186)	(0.0002)
	x_4	1.0104	0.1337	0.1059	0.0991	0.1159	1.0076	0.1201
		(0.1325)	(0.0025)	(0.0147)	(0.0195)	(0.0408)	(0.1217)	(0.0020)
IV: x_4	x_2	0.9993	0.0243	0.0153	0.0153	0.0178	0.9995	0.0237
		(0.0238)	(0.0004)	(0.0025)	(0.0038)	(0.0079)	(0.0233)	(0.0004)
	x_3	1.0037	0.0387	0.0242	0.0251	0.0313	1.0036	0.0369
		(0.0381)	(0.0006)	(0.0040)	(0.0053)	(0.0117)	(0.0370)	(0.0005)
	x_4	1.0032	0.0705	0.0567	0.0579	0.0635	1.0013	0.0454
		(0.0686)	(0.0007)	(0.0093)	(0.0121)	(0.0218)	(0.0442)	(0.0003)
IV: all	x_2	1.0231	0.0333	0.0117	0.0188	0.0197	1.2498	0.0109
		(0.0337)	(0.0008)	(0.0021)	(0.0043)	(0.0081)	(0.0109)	(0.0001)
	x_3	1.0052	0.0452	0.0147	0.0257	0.0267	0.9681	0.0182
		(0.0444)	(0.0008)	(0.0026)	(0.0082)	(0.0123)	(0.0183)	(0.0002)
	x_4	0.9991	0.1512	0.0445	0.0801	0.1005	1.0142	0.0448
		(0.1511)	(0.0033)	(0.0080)	(0.0290)	(0.0463)	(0.0447)	(0.0003)

Table B.19 Small panel with $G = 6$, $T = 4$. Case 5.1: $x_{2it} \sim N(gt/2, 1) + z_i + f_i$, $n_{gt} = 200$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	0.9993	0.0539	0.0527	0.0471	0.0479	0.9994	0.0537
		(0.0550)	(0.0021)	(0.0113)	(0.0141)	(0.0206)	(0.0550)	(0.0021)
	x_3	1.0066	0.0888	0.0868	0.0823	0.0871	1.0064	0.0885
		(0.0874)	(0.0032)	(0.0185)	(0.0208)	(0.0326)	(0.0874)	(0.0032)
	x_4	1.0131	0.2859	0.2794	0.2623	0.2988	1.0141	0.2849
		(0.2886)	(0.0124)	(0.0595)	(0.0718)	(0.1102)	(0.2882)	(0.0124)
IV: z	x_2	1.0017	0.0360	0.0355	0.0359	0.0353	1.0017	0.0353
		(0.0342)	(0.0010)	(0.0042)	(0.0051)	(0.0115)	(0.0347)	(0.0010)
	x_3	1.0058	0.0848	0.0841	0.0688	0.0715	1.0052	0.0838
		(0.0826)	(0.0028)	(0.0100)	(0.0154)	(0.0270)	(0.0828)	(0.0028)
	x_4	1.0151	0.2793	0.2767	0.2209	0.2483	1.0159	0.2761
		(0.2798)	(0.0117)	(0.0339)	(0.0534)	(0.0897)	(0.2784)	(0.0115)
IV: x_2	x_2	1.0280	0.0785	0.0376	0.0449	0.0466	1.2559	0.0245
		(0.0797)	(0.0042)	(0.0107)	(0.0125)	(0.0206)	(0.0244)	(0.0006)
	x_3	1.0159	0.1117	0.0483	0.0684	0.0662	0.8574	0.0825
		(0.1086)	(0.0048)	(0.0134)	(0.0242)	(0.0324)	(0.0821)	(0.0027)
	x_4	1.0101	0.3470	0.1393	0.1977	0.2463	1.2250	0.2751
		(0.3527)	(0.0178)	(0.0385)	(0.0766)	(0.1195)	(0.2813)	(0.0116)
IV: x_3	x_2	1.0062	0.0571	0.0420	0.0381	0.0432	1.0058	0.0502
		(0.0569)	(0.0022)	(0.0059)	(0.0086)	(0.0184)	(0.0514)	(0.0018)
	x_3	1.0022	0.0464	0.0353	0.0367	0.0399	1.0023	0.0416
		(0.0468)	(0.0013)	(0.0049)	(0.0061)	(0.0139)	(0.0432)	(0.0010)
	x_4	1.0234	0.2950	0.2338	0.2197	0.2577	1.0203	0.2628
		(0.2921)	(0.0123)	(0.0328)	(0.0443)	(0.0886)	(0.2627)	(0.0102)
IV: x_4	x_2	1.0004	0.0543	0.0344	0.0344	0.0400	1.0051	0.0522
		(0.0550)	(0.0021)	(0.0058)	(0.0087)	(0.0171)	(0.0536)	(0.0020)
	x_3	1.0088	0.0864	0.0544	0.0563	0.0694	1.0049	0.0815
		(0.0839)	(0.0028)	(0.0091)	(0.0127)	(0.0267)	(0.0800)	(0.0026)
	x_4	1.0047	0.1580	0.1276	0.1303	0.1418	1.0017	0.1007
		(0.1579)	(0.0037)	(0.0211)	(0.0279)	(0.0502)	(0.0993)	(0.0017)
IV: all	x_2	1.0262	0.0743	0.0225	0.0382	0.0409	1.2446	0.0239
		(0.0755)	(0.0038)	(0.0056)	(0.0106)	(0.0176)	(0.0250)	(0.0006)
	x_3	1.0115	0.1012	0.0283	0.0559	0.0572	0.9700	0.0399
		(0.0984)	(0.0040)	(0.0069)	(0.0194)	(0.0271)	(0.0428)	(0.0009)
	x_4	1.0035	0.3378	0.0852	0.1750	0.2212	1.0149	0.0985
		(0.3431)	(0.0165)	(0.0207)	(0.0674)	(0.1072)	(0.1026)	(0.0017)

Table B.20 Small panel with $G = 6$, $T = 4$. Case 5.2: $x_{2it} \sim N(gt/2, 1) + z_i + f_i$, $n_{gt} = 1000$, sampling rate= 0.2%. se_n, se_r and se_c are the non-roust, robust and cluster-robust standard errors, receptively.

		MD Identity					MD Optimal	
		$\check{\beta}$	$\widehat{se(\check{\beta})}$	$\widehat{se_n(\check{\beta})}$	$\widehat{se_r(\check{\beta})}$	$\widehat{se_c(\check{\beta})}$	$\hat{\beta}$	$\widehat{se(\hat{\beta})}$
IV: none	x_2	1.0015	0.0241	0.0232	0.0206	0.0207	1.0015	0.0241
		(0.0241)	(0.0004)	(0.0048)	(0.0060)	(0.0094)	(0.0241)	(0.0004)
	x_3	0.9967	0.0398	0.0384	0.0366	0.0390	0.9968	0.0398
		(0.0401)	(0.0006)	(0.0079)	(0.0086)	(0.0138)	(0.0401)	(0.0006)
	x_4	1.0016	0.1279	0.1235	0.1166	0.1324	1.0018	0.1278
		(0.1304)	(0.0024)	(0.0255)	(0.0305)	(0.0489)	(0.1304)	(0.0024)
IV: z	x_2	1.0018	0.0160	0.0158	0.0160	0.0157	1.0019	0.0159
		(0.0151)	(0.0002)	(0.0019)	(0.0022)	(0.0054)	(0.0151)	(0.0002)
	x_3	0.9966	0.0383	0.0378	0.0304	0.0327	0.9967	0.0382
		(0.0383)	(0.0006)	(0.0046)	(0.0066)	(0.0114)	(0.0384)	(0.0006)
	x_4	1.0017	0.1265	0.1251	0.0975	0.1107	1.0018	0.1262
		(0.1293)	(0.0023)	(0.0151)	(0.0239)	(0.0399)	(0.1290)	(0.0023)
IV: x_2	x_2	1.0291	0.0351	0.0190	0.0214	0.0219	1.2617	0.0111
		(0.0343)	(0.0008)	(0.0039)	(0.0050)	(0.0086)	(0.0109)	(0.0001)
	x_3	0.9982	0.0500	0.0245	0.0309	0.0296	0.8421	0.0375
		(0.0489)	(0.0010)	(0.0050)	(0.0106)	(0.0143)	(0.0381)	(0.0006)
	x_4	0.9876	0.1554	0.0708	0.0891	0.1111	1.2310	0.1255
		(0.1574)	(0.0036)	(0.0146)	(0.0336)	(0.0533)	(0.1299)	(0.0023)
IV: x_3	x_2	1.0014	0.0257	0.0189	0.0171	0.0192	1.0019	0.0227
		(0.0253)	(0.0004)	(0.0025)	(0.0037)	(0.0085)	(0.0226)	(0.0003)
	x_3	0.9985	0.0207	0.0158	0.0165	0.0180	0.9995	0.0188
		(0.0212)	(0.0003)	(0.0021)	(0.0026)	(0.0060)	(0.0190)	(0.0002)
	x_4	1.0012	0.1334	0.1057	0.0993	0.1158	0.9987	0.1198
		(0.1350)	(0.0025)	(0.0139)	(0.0181)	(0.0403)	(0.1211)	(0.0021)
IV: x_4	x_2	1.0017	0.0243	0.0152	0.0151	0.0173	1.0028	0.0237
		(0.0244)	(0.0004)	(0.0025)	(0.0038)	(0.0078)	(0.0237)	(0.0004)
	x_3	0.9962	0.0388	0.0241	0.0251	0.0309	0.9961	0.0370
		(0.0392)	(0.0006)	(0.0039)	(0.0053)	(0.0117)	(0.0369)	(0.0005)
	x_4	1.0031	0.0705	0.0564	0.0579	0.0628	1.0034	0.0454
		(0.0713)	(0.0007)	(0.0091)	(0.0120)	(0.0221)	(0.0460)	(0.0004)
IV: all	x_2	1.0270	0.0333	0.0116	0.0185	0.0188	1.2514	0.0109
		(0.0325)	(0.0008)	(0.0020)	(0.0042)	(0.0074)	(0.0110)	(0.0001)
	x_3	0.9952	0.0453	0.0146	0.0254	0.0258	0.9645	0.0182
		(0.0445)	(0.0008)	(0.0026)	(0.0084)	(0.0118)	(0.0187)	(0.0002)
	x_4	0.9816	0.1514	0.0440	0.0790	0.0997	1.0142	0.0448
		(0.1535)	(0.0033)	(0.0078)	(0.0297)	(0.0478)	(0.0461)	(0.0003)

BIBLIOGRAPHY

BIBLIOGRAPHY

Arellano, Manuel. 1987. "PRACTITIONERS' CORNER: Computing Robust Standard Errors for Within-groups Estimators." *Oxford bulletin of Economics and Statistics*, 49(4): 431–434.

Breusch, Trevor, Hailong Qian, Peter Schmidt, and Donald Wyhowski. 1999. "Redundancy of moment conditions." *Journal of econometrics*, 91(1): 89–111.

Collado, M Dolores. 1997. "Estimating dynamic models from time series of independent cross-sections." *Journal of Econometrics*, 82(1): 37–62.

Deaton, Angus. 1985. "Panel data from time series of cross-sections." *Journal of econometrics*, 30(1): 109–126.

Girma, Sourafel. 2000. "A quasi-differencing approach to dynamic modelling from a time series of independent cross-sections." *Journal of Econometrics*, 98(2): 365–383.

Hansen, Christian B. 2007*a*. "Asymptotic properties of a robust variance matrix estimator for panel data when T is large." *Journal of Econometrics*, 141(2): 597–620.

Hansen, Christian B. 2007*b*. "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects." *Journal of Econometrics*, 140(2): 670–694.

Heckman, James J, and V Joseph Hotz. 1989. "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training." *Journal of the American statistical Association*, 84(408): 862–874.

Imbens, Guido, and Jeffrey M Wooldridge. 2007. *What's new in econometrics?* NBER.

Kezdi, Gabor. 2003. "Robust standard error estimation in fixed-effects panel models." *Available at SSRN 596988*.

McKenzie, David J. 2004. "Asymptotic theory for heterogeneous dynamic pseudo-panels." *Journal of Econometrics*, 120(2): 235–262.

Moffitt, Robert. 1993. "Identification and estimation of dynamic models with a time series of repeated cross-sections." *Journal of Econometrics*, 59(1): 99–123.

Newey, Whitney K, and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." *Handbook of econometrics*, 4: 2111–2245.

Verbeek, Marno. 2008. "Pseudo-panels and repeated cross-sections." In *The Econometrics of Panel Data*. 369–383. Springer.

Verbeek, Marno, and Francis Vella. 2005. "Estimating dynamic models from repeated cross-sections." *Journal of econometrics*, 127(1): 83–102.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* . 2nd ed., Boston MA:MIT Press.

CHAPTER 3

A FLEXIBLE PLUG-IN G-FORMULA FOR CONTROLLED DIRECT EFFECTS IN MEDIATION ANALYSIS

3.1 Introduction

In the literature of epidemiology and biostatistics, the term g-methods (e.g., Westreich et al., 2012) are often used to collectively refer to g-formula (Robins, 1986), g-estimation of structural nested models (Robins, 1998), and inverse probability weighting of marginal structural models (Horvitz and Thompson, 1952; Robins, 1989; Hernan and Robins, 2015), all of which are useful approaches in estimating the effects of time-varying treatments in the presence of time-varying confounders. The g-formula in its original form is non-parametric and is the foundation for the other two. While non-parametric g-formula is flexible in its model specification, it is also quite demanding on data. Therefore, we often introduce semi-parametric or parametric modeling. Despite the fact that parametric models are almost always misspecified, the parametric g-formula often yields satisfactory estimates as long as the specified models are reasonably flexible. However, most applications of the parametric g-formula (e.g., Westreich et al., 2012; Taubman et al., 2009; Young et al., 2011; Danaei et al., 2013; Lajous et al., 2013; Garcia-Aymerich et al., 2014) still use Monte Carlo integration to calculate because closed-form expressions of the treatment effects of interest are either non-existent or tedious to derive.

The application of these g-method to mediation analysis is straightforward as mediation analysis is conceptually equivalent to a sequential treatment of two periods. Since Robins and Greenland (1992) conceptualize the natural and controlled effects in mediation analysis using the potential outcome (counterfactual) framework, several mediation analysis methods are developed from g-methods. These methods include, among others, the parametric g-formula in Daniel, De Stavola and Cousens (2011) and Valeri and Vanderweele (2013),

and the parametric version of the sequential g-estimation of structural nested mean models (SNMMs) studied by Vansteelandt (2009). As in the case of g-methods, these methods also rely on Monte Carlo integration to calculate the treatment effects of interest, which can be computationally demanding. Meanwhile, since there is no closed-form expressions, we almost always need to bootstrap the standard errors, which raises the computation intensity rapidly. When the estimation itself is time-consuming, the problem gets amplified even further. This includes, but not limited to, maximum likelihood estimation when it converges slowly and most semi-parametric or non-parametric techniques that require cross validation for tuning parameter selection.

In view of this limitation, in this chapter we propose a so called flexible plug-in g-formula for controlled direct effects (CDE) in mediation analysis. The key assumption needed is that the conditional expectation of the outcome is linear in time-varying confounders. This partial linearity allows us to replace the confounders with their fitted values, which results in a plug-in estimator for CDE. At the same time, it also relaxes the fully linear assumptions that are commonly used in empirical studies, which gives us more flexibility in choosing the functional form of the outcome conditional mean. As a result, we have a better chance to be closer to the true underlying model.

Besides the partial linearity assumption, another necessary condition for the consistency of the flexible plug-in g-formula is the sequential ignorability assumption (Robins, 1986). To check the robustness of the estimator to a particular violation of the sequential ignorability assumption, we present a sensitivity analysis that is similar in spirit to that proposed by Imai, Keele and Tingley (2010). The proposed estimator is evaluated in a small simulation and its use is illustrated in a longitudinal cohort study.

The rest of this chapter is organized as follows. We first set up the counterfactual mediation analysis framework in the second section. In the third section, we review the general g-formula as well as the sequential g-estimation. In the fourth section, we present in detail the flexible plug-in g-formula. In particular, it is compared to the sequential g-

estimation in some commonly used linear specifications. The next two sections outline the sensitivity analysis and an empirical application, respectively. The last section concludes.

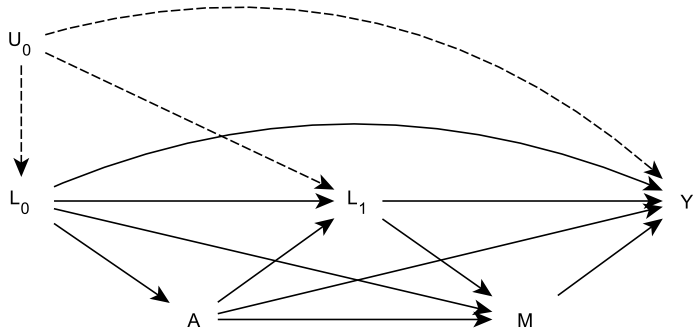
3.2 Framework

A causal mediation analysis is typically guided by a directed acyclic graphs (DAG) (Pearl, 2009). We use the DAG $\mathcal{G}$ in Figure 1 for illustration, but our method can be applied in similar models that satisfy the assumptions below. Throughout this paper, a DAG is viewed as a graphical representation of an underlying non-parametric structural equation model with independent errors (NPSEM-IE) (Pearl, 2009).[1] A DAG and the associated NPSEM-IE are related as usual: "there is an equation for each variable in the model, specifying that variable as a function of its parents in the graph" (Richardson and Robins, 2013). The counterfactual outcomes, defined by intervening on certain variables in the NPSEM-IE model, are used in constructing the CDEs of interest.

Specifically, assume we have a longitudinal study in which each respondent was interviewed 3 times at $k = 0, 1, 2$, with a one-time treatment A at $k = 0$. Each interview generates data L_k. The purpose is to learn to what extent the treatment effect of A on Y is mediated by a mediator M. (A, M) are the intervention nodes, and $\{L_k : k = 0, 1, 2\}$ are observed non-intervention nodes or confounders. L_0 contains all baseline information, and L_1 contains all post-treatment non-intervention nodes that occur before M and confounds the mediator-outcome relationship. As noted in Pearl (2014), we cannot identify natural mediation effects non-parametrically because of the existence of L_1. Hence, we focus on the CDEs in this paper. The model allows for a type of harmless unobserved variables, collec-

[1] Although the NPSEM-IE model makes many more counterfactual independence assumptions than, and is a strict submodel of, the finest fully randomized causally interpretable structured tree graph (FFRCISTG) model of (Robins, 1986; Robins and Richardson, 2010), the generality of the latter does not play an essential role in our paper. The adoption of NPSEM-IE, however, makes the description of the model straightforward and easy to follow, since it is built on the traditional structural equation models and imposes simple distribution assumptions on the errors (exogenous variables).

Figure 3.1 A directed acyclic graph for a longitudinal study with three time points. (A, M) are the intervention nodes, (L_0, L_1, Y) the non-intervention nodes, and U_0 the unobservables.



$\mathcal{G}$

tively denoted by U_0 that are parents of the observed non-intervention nodes. Based on the back-door criterion the observed non-intervention nodes block the confounding effects of U_0 Pearl (2009).

Each node (including the unobservable U_0) has an exogenous error (not shown in the graph) attached solely to itself. For example, ε_Y is the exogenous error for Y, as ε_{U_0} is for U_0. All errors are unobserved and are assumed to be jointly independent in NPSEM-IE. The independence assumption will be relaxed in our sensitivity analysis in which a more general NPSEM allowing correlations between ε's will be used.

Let Y^{am} be the counterfactual outcome where A and M are intervened to be fixed at a and m. In the context of NPSEM-IEs, such an intervention would correspond to the operation of deleting the equations for A and M from the system and substituting $A = a$ and $M = m$ in the system. This operation is called the *do* operation in Pearl (2009). In this paper, we consider a binary A so that $a \in \{0, 1\}$. An extension to multi-valued A is trivial. For a fixed m, we are interested in the CDE(m) defined as $E(Y^{1m} - Y^{0m})$, or $E(Y(1, m) - Y(0, m))$.

3.3 Existing Methods

In the context of DAG $\mathcal{G}$, we discuss two existing methods for estimating the CDE. The first method is the g-formula for the marginal distribution of Y^{am} introduced in Robins (1986) and revisited in Richardson and Robins (2013), which is the basis for the parametric g-formula and in particular for the flexible plug-in estimator. The second method is the sequential g-estimation Vansteelandt (2009), which is, as we will show, numerically equivalent to the flexible plug-in estimator in certain linear cases.

3.3.1 The g-Formula

In Richardson and Robins (2013), the term g-formula refers to the unextended g-formula of Robins (1986), the extended g-formula of Robins, Hernán and SiEBERT (2004), or the g-formula for a sequence of treatments and a single response. Among the three, the last one is of interest in the context of DAG $\mathcal{G}$ and the estimation of CDEs.

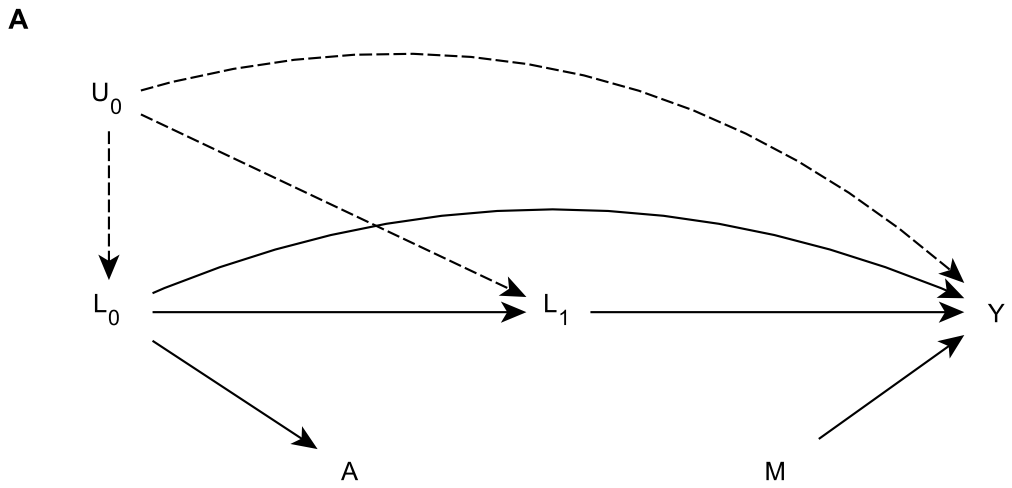
Let $P(Y^{am} = y)$ be the distribution of the potential response Y under the sequence of interventions $(A = a, M = m)$, from which we can construct the CDE. Assume the consistency rule holds Robins (1994), that is, a potential response under a hypothetical condition that happened to take place is precisely the observed response. In addition, the following form of sequential ignorability condition from Robins (2000) is imposed:

$$\begin{array}{c} Y^{am} \perp\!\!\!\perp A \mid L_0, \\ Y^{am} \perp\!\!\!\perp M \mid L_0, A, L_1, \end{array} \quad for\ all\ a, m \tag{3.1}$$

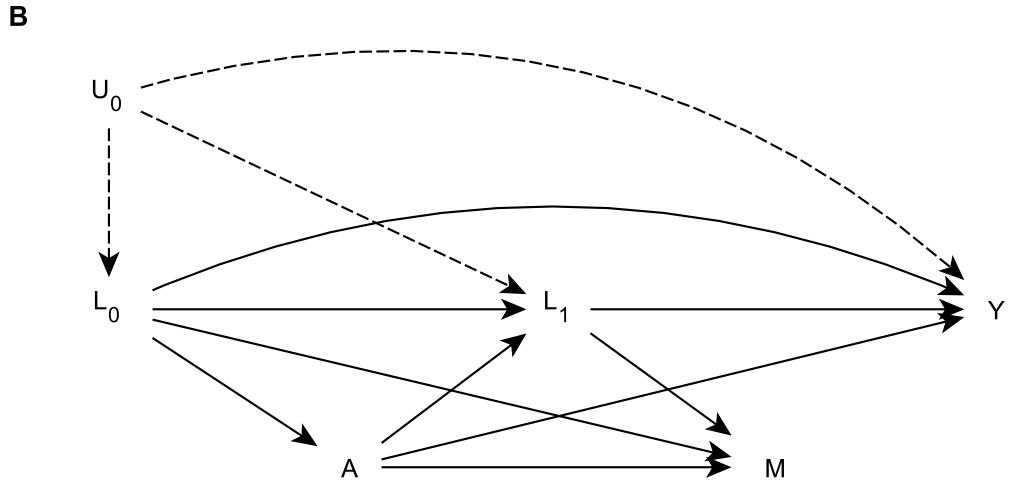
where $\perp\!\!\!\perp$ represents distributional independence. Note that condition (3.1) summarizes the set of independence conditions for all possible values of a and m, which are needed to identify CDEs for all possible values of M.

A logically equivalent algorithm of finding the right conditioning set of variables as the sequential ignorability conditions in (3.1) is the sequential back-door criterion developed in

Figure 3.2 Subgraphs for $\mathcal{G}$ where an upper bar means arrows pointing to a node are removed and an under bar means arrows emitting from a node are removed.



$\mathcal{G}_{\underline{A}\overline{M}}$



$\mathcal{G}(\underline{M})$

Pearl and Robins (1995), which states

$$\begin{aligned}(Y \perp\!\!\!\perp A \mid L_0)_{\mathcal{G}_{\underline{A}\overline{M}}}, \\ (Y \perp\!\!\!\perp M \mid L_0, A, L_1)_{\mathcal{G}_{\underline{M}}},\end{aligned} \tag{3.2}$$

where $\mathcal{G}(\underline{M})$ denotes the subgraph obtained by removing from $\mathcal{G}$ all arrows emerging from M (Figure 3.2, B), and $\mathcal{G}_{\underline{A}\overline{M}}$ denotes the removal of both incoming arrows to M and outgoing arrows from A (Figure 3.2, A).

Under either (3.1) or (3.2), the g-formula for the expected counterfactual outcome is

$$E(Y^{am}) = \int\int\int y f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m) f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dy dl_1 dl_0 \tag{3.3}$$

$$= \int\int E(y|l_0,a,l_1,m) f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dl_1 dl_0 \tag{3.4}$$

where, e.g, $P(y|l_0,a,l_1,m)$ is shorthand for $P(Y=y|L_0=l_0,A=a,L_1=l_1,M=m)$. See Appendix A for detail. Then the CDE is

$$\begin{aligned}&E(Y^{1m} - Y^{0m}) \\ =&\int\int E(y|l_0,1,l_1,m) f_{L_1|L_0}(l_1|l_0,1) f_{L_0}(l_0) dl_1 dl_0 \\ &- \int\int E(y|l_0,0,l_1,m) f_{L_1|L_0}(l_1|l_0,) f_{L_0}(l_0) dl_1 dl_0.\end{aligned} \tag{3.5}$$

Equation (3.3) is non-parametric in the sense that no parametric assumptions are made yet for $f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m)$, $f_{L_1|L_0}(l_1|l_0,a)$ and $f_{L_0}(l_0)$. Equation (3.4) adds an additional assumption that the conditional mean $E(y|l_0,a,l_1,m)$ exits and is finite.

Two straightforward estimation strategies to estimate $E(Y^{am})$ follow from equations (3.3) and (3.4). The first strategy exploits equation (3.3), with either non-parametric or parametric specification of the distributions. This strategy generally involves Monte Carlo simulation and numerical integration for calculating CDEs (Daniel, De Stavola and Cousens, 2011; Imai, Keele and Tingley, 2010; Hicks and Tingley, 2011).

The second strategy exploits equation (3.4), which avoids the estimation of the density function $f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m)$. Instead, we only estimates the conditional mean

$E(Y|l_0, a, l_1, m)$. We can keep the non-parametric feature or choose suitable parametric models to reduce computation burdens. Many applications have shown that a properly chosen parametric method can provide good approximations (Westreich et al., 2012; Taubman et al., 2009; Young et al., 2011; Danaei et al., 2013; Lajous et al., 2013; Garcia-Aymerich et al., 2014). The estimator proposed in this chapter falls in the second estimation strategy and imposes a particular parametric assumption on $E(Y|l_0, a, l_1, m)$ that simplifies equation (3.4) even further (details below).

3.3.2 The Sequential g-formula estimator

The sequential g-estimation for CDEs is a two-step estimator based on an SNMM (Vansteelandt, 2009). The idea is to first partial out the effect of the mediator on the outcome and then regress the adjusted outcome on the treatment, the confounders, and possibly their interactions to identify the direct effect. The sequential g-formula estimator assumes an additive separable functional form in the conditional mean equation for Y

$$E(Y|l_0, a, l_1, m) = \quad q_A(l_0, a, l_1; \boldsymbol{\gamma}) + q_M(l_0, a, l_1, m; \boldsymbol{\gamma}) \tag{3.6}$$

where $q_A(\cdot)$ and $q_M(\cdot)$ are arbitrary known functions with finite dimensional parameter $\boldsymbol{\gamma}$, satisfying $q_M(l_0, a, l_1, m = 0; \boldsymbol{\gamma}) = 0$. For example, we can assume $q_A = \gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{L_1} l_1$ and $q_M = \gamma_M m$. In addition, assume an SNMM for $E(Y^{am} - Y^{0m}|l_0) = \varphi_A a$ where φ_A is the CDE. Then the sequential g-estimation procedure is:

1. regress Y on $(1, L_0, A, L_1, M)$ and obtain the ordinary least square (OLS) estimator $\hat{\gamma}_M$ for γ_M and generate $\hat{Y}_{-M} \equiv Y - \hat{\gamma}_M M$, and

2. regress $\hat{Y}_{-M}$ on $(1, L_0, A)$. Denote by $\hat{\varphi}_A$ the OLS estimator for the coefficient of A. It can be shown $\hat{\varphi}_A$ is a consistent estimator for φ_A.[2]

[2]Under the sequential ignorability conditions and additive separability (3.6), Vansteelandt (2009) gave the key identification result $E[Y - q_M(L_0, A, L_1, M; \boldsymbol{\gamma})|L_0 = l_0, A = a] =$

See Appendices C and D for the validity of the estimation procedure.

Table 3.1 lists five typical examples of SNMMs compatible with DAG $\mathcal{G}$ that can be estimated using the sequential g-formula estimator. The simple example above is Model 1 in Table 3.1. The first three models were discussed in Vansteelandt (2009), and we added the latter two models with more flexible specifications. One difficulty in applying the sequential g-formula estimator is to find the proper $q_A(\cdot)$ and $q_M(\cdot)$ functions for a given SNMM, as can been seen in Table 3.1. The derivation of standard errors of the estimator is also no mean feat.

The sequential g-estimation is not always as simple as it looks in Model 1. For example, in a model with up to two-way interactions (Model 5 in Table 3.1), we need to estimate φ_{AM}, φ_{AL_0} and φ_{AML_0} in $E(Y^{1m} - Y^{0m}) = \varphi_A + \varphi_{AM}m + \varphi_{AL_0}E(L_0) + \varphi_{AML_0}mE(L_0)$. It turns out that in this case only $E(L_1|L_0 = l_0, A = 0) = \pi_0 + \pi_{L_0}l_0$ is not enough to identify φ_{AM} or φ_{AML_0}. What is needed is the stronger assumption

$$E(L_1|L_0 = l_0, A = a) = \pi_0 + \pi_{L_0}l_0 + \pi_A a + \pi_{AL_0}a \times l_0$$

which implies (by setting $\gamma_{AML_0} = \gamma_{AML_1} = 0$ in Appendix B)

$$\varphi_{AM} = \gamma_{AM} + \gamma_{ML_1}\pi_A,$$

$$\varphi_{AML_0} = \gamma_{ML_1}\pi_{AL_0},$$

Clearly, the above equations show that φ_{AM} and φ_{AML_0} cannot be obtained directly from the first-step regression, since now an additional regression for $E(L_1|L_0 = l_0, A = a)$ is needed to estimate $\boldsymbol{\pi}$.

$E\left[Y^{a0}|L_0 = l_0\right]$. We show that, in addition to the assumptions in Vansteelandt (2009), a necessary and sufficient condition for the validity of the second-step regression is $f(l_0, a) \equiv E(L_1|L_0 = l_0, A = 0) = \pi_0 + \pi_{L_0}l_0$. See Appendix B for detail.

Table 3.1 Compare the plug-in estimator with the sequential g-estimator under different specifications for the outcome conditional mean and different structural nested mean models.

	Structural Nested Mean Model (SNMM) $E(Y^{am} - Y^{0m}\|l_0)$	Sequential g-formula estimator $E(Y\|l_0, a, l_1, m) =$ $q_A(l_0, a, l_1; \gamma) + q_M(l_0, a, l_1, m; \gamma)$	Flexible Plug-in g-formula estimator $E(Y\|l_0, a, l_1, m) =$ $h_0(a, m)l_0 + h_1(a, m)l_1 + h(a, m)$
1	$\varphi_A a$	$q_A = \gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{L_1} l_1$ $q_M = \gamma_M m$ $f(l_0, 0) = \pi_0 + \pi_{L_0} l_0$	$h_0 = \gamma_{L_0},\ h_1 = \gamma_{L_1}$ $h = \gamma_0 + \gamma_A a + \gamma_M m$ $f(l_0, a) = \pi_0 + \pi_{L_0} l_0 + \pi_A a$
2	$\varphi_A a + \varphi_{AM} a \cdot m$	$q_A = \gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{L_1} l_1$ $q_M = \gamma_M m + \gamma_{AM} a \cdot m$ $f(l_0, 0) = \pi_0 + \pi_{L_0} l_0$	$h_0 = \gamma_{L_0},\ h_1 = \gamma_{L_1}$ $h = \gamma_0 + \gamma_A a + \gamma_M m + \gamma_{AM} a \cdot m$ $f(l_0, a) = \pi_0 + \pi_{L_0} l_0 + \pi_A a$
3	$\varphi_A a + \varphi_{AL_0} a \cdot l_0$	$q_A = \gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{L_1} l_1$ $+\gamma_{AL_0} a \cdot l_0$ $q_M = \gamma_M m$ $f(l_0, 0) = \pi_0 + \pi_{L_0} l_0$	$h_0 = \gamma_{L_0} + \gamma_{AL_0} a,\ h_1 = \gamma_{L_1}$ $h = \gamma_0 + \gamma_A a + \gamma_M m$ $f(l_0, a) = \pi_0 + \pi_{L_0} l_0 + \pi_A a$ $+\pi_{AL_0} a \cdot l_0$
4	$\varphi_A a + \varphi_{AM} a \cdot m$ $+\varphi_{AL_0} a \cdot l_0$	$q_A = \gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{L_1} l_1$ $+\gamma_{AL_0} a \cdot l_0$ $q_M = \gamma_M m + \gamma_{AM} a \cdot m$ $f(l_0, 0) = \pi_0 + \pi_{L_0} l_0$	$h_0 = \gamma_{L_0} + \gamma_{AL_0} a,\ h_1 = \gamma_{L_1}$ $h = \gamma_0 + \gamma_A a + \gamma_M m + \gamma_{AM} a \cdot m$ $f(l_0, a) = \pi_0 + \pi_{L_0} l_0 + \pi_A a$ $+\pi_{AL_0} a \cdot l_0$
5	$\varphi_A a + \varphi_{AM} a \cdot m +$ $+\varphi_{AL_0} a \cdot l_0 + \varphi_{AL_0} a \cdot l_0$ $+\varphi_{AML_0} a \cdot m \cdot l_0$	$q_A = \gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{L_1} l_1$ $+\gamma_{AL_0} a \cdot l_0 + \gamma_{AL_1} a \cdot l_1$ $q_M = \gamma_M m + \gamma_{AM} a \cdot m + \gamma_{ML_0} m \cdot l_0$ $+\gamma_{ML_1} m \cdot l_1$ $f(l_0, a) = \pi_0 + \pi_{L_0} l_0 + \pi_A a$ $+\pi_{AL_0} a \cdot l_0$	$h_0 = \gamma_{L_0} + \gamma_{AL_0} a + \gamma_{ML_0} m$ $h_1 = \gamma_{L_1} + \gamma_{AL_1} a + \gamma_{ML_1} m$ $h = \gamma_0 + \gamma_A a + \gamma_M m + \gamma_{AM} a \cdot m$ $f(l_0, a) = \pi_0 + \pi_{L_0} l_0 + \pi_A a$ $+\pi_{AL_0} a \cdot l_0$

Notation: $q_A \equiv q_A(l_0, a, l_1; \gamma)$, $q_M \equiv q_M(l_0, a, l_1, m; \gamma)$, $f(l_0, a) \equiv E(L_1 | L_0 = l_0, A = a)$.

3.4 The Flexible Plug-in g-formula estimator

3.4.1 The Partial Linearity Assumption and the Plug-in g-formula estimator

Although the idea of using linear outcome conditional mean is not particularly new (Robins, 2000; Van der Wal et al., 2009), to the best of our knowledge, the flexible plug-in g-formula estimator proposed here is the first to make full use of this idea. This parametric g-formula has a closed-form expression for CDE and thus does not require numerical integration.

Specifically, let the conditional expectation of Y given (L_0, A, L_1, M) be linear in L_0 and L_1, namely

$$E(Y|l_0, a, l_1, m) = h_0(a, m; \boldsymbol{\gamma})l_0 + h_1(a, m; \boldsymbol{\gamma})l_1 + h(a, m; \boldsymbol{\gamma}) \tag{3.7}$$

where the h_k's are arbitrary known functions of (a, m) known up to certain parameters $\boldsymbol{\gamma}$, for $k = \{0, 1, \emptyset\}$. This is of course a strong parametric assumption since it ignores any interaction among confounders. We should think of Equation (3.7) as the first order Taylor approximation to any function of (L_0, L_1). The plug-in g-formula estimator can be extended to include higher order terms of confounders.

Equations (3.4) and (3.7) lead to the proposed flexible plug-in g-formula estimator for $E(Y^{am})$:

$$E(Y^{am}) = h_0(a, m; \boldsymbol{\gamma})E(L_0) + h_1(a, m; \boldsymbol{\gamma})E\left[E(L_1|L_0, A = a)\right] + h(a, m; \boldsymbol{\gamma}). \tag{3.8}$$

See Appendix A for proof.

The last column of Table 3.1 shows that, by varying the specifications of h_k's, equation (3.7) can provide the same specification on $E(Y|l_0, a, l_1, m)$ as equation (3.6) in all the five models there. The separability in q_A and q_M for the sequential g-formula estimator and the estimating equation that is linear in confounders for the plug-in g-formula estimator do not nest within each other. Neither estimator is strictly more flexible than the other. Note that the unknown parameters $\boldsymbol{\gamma}$ in (3.7) are not the structural parameters that would appear in the structural equation $Y = f_Y(L_0, A, L_1, M, \epsilon_Y)$ in the NPSEM-IE. Equation (3.7) is

an estimating equation only. The parameters in this equation are not of direct interest in general, but they eventually identify $E(Y^{am})$.

Note also that as long as equation (3.7) holds, the presence of U_0 does not affect the identification of CDE. However, we do need a model for $E(L_1|L_0, A)$. This model needs not be linear, and can potentially be semi- or non-parametric if the dimension of the conditioning set is low. For example, if L_1 is binary, a logistic model can be used. But in the application below, we use the following linear model for simplicity:

$$E(L_1|L_0, A) = \pi_0 + \pi_A A + \pi_{L_0} L_0 + \pi_{AL_0} A \times L_0, \tag{3.9}$$

and thus

$$E\left[E(L_1|L_0, A = a)\right] = \pi_0 + \pi_A a + \pi_{L_0} E(L_0) + \pi_{AL_0} a \times E(L_0). \tag{3.10}$$

3.4.2 Estimation Procedure for the Flexible Plug-in g-formula estimator of CDE

Given the discussion above, the CDE can be estimated as follows:

1. estimate $\boldsymbol{\gamma}$ in equation (3.7) using a proper method, e.g., a quasi-maximum likelihood estimator, which is consistent given correctly specified conditional mean (Wooldridge, 2010); and

2. estimate $E(L_0)$ using a proper method, e.g., the sample mean, and get $\widehat{E(L_0)}$.

3. Estimate $E(L_1|L_0, A)$ using a proper method and get $\widehat{E\left[E(L_1 \mid L_0, A = a)\right]}$. For example, regress L_1 on $(1, A, L_0, AL_0)$.

4. Plug $\widehat{E(L_0)}$ and $\widehat{E\left[E(L_1 \mid L_0, A = a)\right]}$ into (3.8) to obtain $\widehat{E(Y^{am})}$. Then $\widehat{\text{CDE}(m)} = \widehat{E(Y^{am})} - \widehat{E(Y^{0m})}$.

$\widehat{\text{CDE}(m)}$ is the flexible plug-in estimator for the CDE evaluated at m. Bootstrap can be used to obtain the standard errors for the estimated CDEs.

3.4.3 plim of Parametric g-Formula is the Flexible Plug-in g-formula estimator

By the law of large numbers, the flexible plug-in g-formula estimator of CDE is the plim of the corresponding parametric g-formula in section 3.1. This is simply because the former is an analytical solution for the integral which the latter is trying to evaluate via Monte Carlo simulation. We verify this claim using Model 5 in Table 3.1 through a simulation study with the aid of the Stata command gformula developed in Daniel, De Stavola and Cousens (2011). In the simulation study, we let the number of Monte Carlo simulations used by the gformula command increase towards infinity. The results show that the estimates obtained using the gformula command indeed come closer and closer to those obtained using the flexible plug-in g-formula as the simulations increase.

One reason for using Model 5 is that the specification is complex enough to make noticeable difference in computation time between the two methods. Obviously, if we are only interested in a point estimate of the controlled direct effect, the flexible plug-in g-formula does not gain us much. However, since we also need to obtain the standard errors for the estimators, and bootstrap is often inevitable in g-methods, the flexible plug-in g-formula can save considerable amount of computation time.

3.4.4 Flexible Plug-in g-formula estimator Is Numerically Equivalent to Sequential g-formula estimator

We show that, in each of the five models in Table 3.1, the flexible plug-in g-formula estimator and the sequential g-formula estimator are numerically identical. See Appendix C for proof. It is worth emphasizing the following two conditions that are met by each of the five models in Table 3.1.

First, the SNMM used by the sequential g-formula estimator must be compatible with the specification on $E(Y|l_0, a, l_1, m)$. Note that a given SNMM and a compatible specification on $E(Y|l_0, a, l_1, m)$ implies a specification on $E(L_1|L_0, A = 0)$ (or on $E(L_1|L_0, A)$ in Model 5) as shown in Table 3.1. Second, the specification on $E(L_1|A, L_0)$ used by the flexible

plug-in estimator must be properly chosen. By "properly chosen", we mean the specification on $E(L_1|A, L_0)$ must be of the same level of flexibility as the second-step regression of the sequential g-formula estimator. See the remark in Appendix C for more detail.

Later when we discuss the issue of "one single parameter", the two estimators are not identical except for the no interaction case. The sequential g-formula estimator forces $E(Y^{am} - Y^{0m})$ to be φa when it is actually not, a typical case in which the incompatibility issue arises.

3.4.5 Simulation

We use a simple simulation study to evaluate the equivalency. Assume the data generating process (DGP) is as follows:

$$\begin{aligned} U_0 &= \varepsilon_{U_0} \\ L_0 &= U_0 + \varepsilon_{L_0} \\ A &= 1[\text{expit}(L_0) \geq \varepsilon_A] \\ L_1 &= U_0 + L_0 + A + \varepsilon_{L_1} \\ M &= 50 \times \text{expit}(L_0 + A + L_1 + \varepsilon_M) \\ Y &= L_0 \times \log(1 + A + M^2) + L_1 \times (A + M) + U_0 + \varepsilon_Y \end{aligned}$$

where all ε's except ε_A are standard normal, ε_A is uniform on $(0, 1)$, and $\text{expit}(x) = \exp(x)/(\exp(x) + 1)$ is the inverse of the logit transformation. All ε's are independent of each other. For simplicity, all coefficients are set to unity. The resulting true CDE is $\text{CDE}(m) = m + 1$ for any fixed value m. Under this DGP, the following estimating equation that is linear in L_0 and L_1 holds:

$$E(Y|L_0, A, L_1, M) = \underbrace{\Big[\log(1 + A + M^2)\Big]}_{h_0(a,m)} L_0 + \underbrace{(\frac{1}{3} + A + M)}_{h_1(a,m)} L_1 + \underbrace{\Big(-\frac{1}{3}A\Big)}_{h(a,m)}. \qquad (3.11)$$

Table 3.2 Simulation results: flexible plug-in g-formula v.s. sequential g-estimator

A: Simulation Results for Model 4

		$E(L_1\|A,L_0) = \pi_0+\pi_A A+\pi_{L_0}L_0$			$E(L_1\|A,L_0) = \pi_0+\pi_A A+\pi_{L_0}L_0+\pi_{AL_0}A\times L_0$		
m	ture CDE	FPG	SG	Difference	FPG	SG	Difference
1	2	57.0846	57.0781	$-6.481e^{-3}$	57.0781	57.0781	0 (0)
		(8.5109)	(8.5106)	(.1090)	(8.5106)	(8.5106)	
25	26	40.1442	40.1377	$-6.481e^{-3}$	40.1377	40.1377	0 (0)
		(5.1294)	(5.1285)	(.1090)	(5.1285)	(5.1285)	
50	51	22.4979	22.4914	$-6.481e^{-3}$	22.4914	22.4914	0 (0)
		(7.1983)	(7.1973)	(.1090)	(7.1973)	(7.1973)	

B: Simulation Results for Model 5

		$E(L_1\|A,L_0) = \pi_0+\pi_A A+\pi_{L_0}L_0$			$E(L_1\|A,L_0) = \pi_0+\pi_A A+\pi_{L_0}L_0+\pi_{AL_0}A\times L_0$		
m	ture CDE	FPG	SG	Difference	FPG	SG	Difference
1	2	.0426 (.4303)	.04219 (.4303)	$-4.951e^{-4}$ (.0236)	.0419 (.4303)	.0419 (.4303)	0 (0)
25	26	24.9052	24.9047	$-4.951e^{-4}$	24.8994	24.8994	0 (0)
		(2.3424)	(2.3425)	(.0236)	(2.3411)	(2.3411)	
50	51	50.8037	50.8032	$-4.951e^{-4}$	50.7926	50.7926	0 (0)
		(4.6871)	(4.6871)	(.0236)	(4.6847)	(4.6847)	

For illustration purposes, we only show the estimation results for the two estimators in Models 4 and Model 5 in Table 3.1. The following two specifications for $E(L_1|A, L_0)$ which differ in their flexibility are considered for both Model 4 and Model 5:

$$E(L_1|A, L_0) = \pi_0 + \pi_A A + \pi_{L_0} L_0, \tag{3.12}$$

$$E(L_1|L_0, A) = \pi_0 + \pi_A A + \pi_{L_0} L_0 + \pi_{AL_0} A \times L_0. \tag{3.13}$$

The specifications for $E(L_1|A, L_0)$ affect the estimates of the flexible plug-in estimator in both Model 4 and Model 5. As for the sequential g-formula estimator, the specifications for $E(L_1|A, L_0)$ have no effect in Model 4 since no estimates involve the estimation of $\boldsymbol{\pi}$, which is in turn a result of the specification that L_1 does not interact with M; but they do have an effect in Model 5 since both φ_{AM} and φ_{AML_0} need the estimation of $\boldsymbol{\pi}$. When the same specification on $E(Y|L_0, A, L_1, M)$ and the same proper specification on $E(L_1|L_0, A)$ are used, the flexible plug-in g-formula estimator and the sequential g-formula estimator must give exactly the same estimates on each occasion.

The simulation consists of 1000 runs with 500 observations in each sample. To save space, we report in Table 3.2 the results for CDE(1), CDE(25) and CDE(50) only, m ranges from 1 to 50. For each estimator, the average and standard deviation (in parenthesis) over the 1000 simulations are reported. We also calculate the difference of the two estimates in each simulation run and report the average and standard deviation of the difference over the 1000 simulations.

There are two important observations from the simulation results in Table 3.2. First, the flexibility of the specification on $E(Y|L_0, A, L_1, M)$ is important. Although Models 4 and 5 both misspecify the true estimating equation, however, compared to Model 5, Model 4 is more restrictive, leading to larger biases in both estimators. Note that in terms of $q_A(\cdot)$ and $q_M(\cdot)$, Model 4 is typical in causal mediation analysis, partly because researchers usually think the two-way interactions $A \times L_1$ and $M \times L_1$ are unnecessary. On the other hand, if

the flexible plug-in g-formula estimator is used, the complete set of two-way interactions in Model 5 becomes typical.

Second, when $E(L_1|A, L_0) = \pi_0 + \pi_A A + \pi_{L_0} L_0$, the two estimates are close but not the same. Because the CDE is linear in m for both estimators, the difference does not depend on m. When the proper specification $E(L_1|A, L_0) = \pi_0 + \pi_A A + \pi_{L_0} L_0 + \pi_{AL_0} A \times L_0$ is used, the two estimates become identical. This supports the numerical equivalence claim made in the last section.

3.4.6 Comparison of Flexible plug-in g-formula estimator with Sequential g-formula estimator

In cases where the two estimators are identical, the flexible plug-in g-formula estimator inherits everything the sequential g-formula estimator has. However, the difference in the estimation procedures grants the former several advantage over the latter in applications. We discuss several points of importance in this respect.

First, when applying the sequential g-formula estimator, one needs to make sure the specifications for $q_A(\cdot)$ and $q_M(\cdot)$ are compatible with the chosen SNMM. For example, $E(Y^{am} - Y^{0m}) = \varphi a$ is not compatible with the $q_M(\cdot)$ in Model 2. As $q_A(\cdot)$ and $q_M(\cdot)$ become more complex, it becomes more difficult to find the corresponding SNMM. The flexible plug-in g-formula estimator avoids this issue, because it starts from assuming the specifications on $E(Y|l_0, a, l_1, m)$ and $E(L_1|a, l_0)$, and the model for CDEs follows naturally. The resulting SNMM can even be nonlinear depending on the specifications on $E(Y|l_0, a, l_1, m)$ and $E(L_1|a, l_0)$.

Second, unless the SNMM is forced to be $E(Y^{am} - Y^{0m}) = \varphi a$, the sequential g-formula estimator does not always depend on "one single parameter" for CDE Vansteelandt (2009). For example, if there is a strong belief that there is treatment-mediator interaction in $q_M(\cdot)$ as in Model 2, then $E(Y^{am} - Y^{0m}) = \varphi_A a + \varphi_{AM} a \times m$ is a function of two parameters. The aforementioned compatibility issue will arise if, in order to force CDE to depend on "one

single parameter", the SNMM is assumed to be $E(Y^{am} - Y^{0m}) = \varphi a$. In an incompatible case it is difficult to interpret what effect the single parameter φ captures. Without further investigation, all we can say is that it is some average of the CDE evaluated at different values of m, and it is unknown whether it is practically relevant. As a result, the test of existence of CDE based on this average becomes less useful than a test evaluated at different values of m.

Third, there are interesting specifications on $E(Y|l_0, a, l_1, m)$ that the sequential g-estimation does not allow. The feasibility of the sequential g-formula estimator hinges on the additive separability between $q_A(\cdot)$ and $q_M(\cdot)$, and clearly, not all specifications for $E(Y|l_0, a, l_1, m)$ satisfy this restriction. Practically interesting examples include cases where $h = log(\gamma_0 + \gamma_A a + \gamma_M m^2)$ or $h = exp(\gamma_0 + \gamma_A a + \gamma_M m^2)$, i.e. we use the link function idea of generalized linear models to enrich specifications on the h_k functions. To be fair, however, we also note that there are specifications that the flexible plug-in g-formula estimator cannot handle. For example, if the conditional expectation of Y is nonlinear in L_0 and L_1, equation (3.7) will not hold, but equation (3.6) may still be satisfied provided that the nonlinearity does not interfere with the additive separability requirement. Extensions of equation (3.7) to be nonlinear in L_0 and L_1 are possible but complicated, in which case the original parametric g-formula might be a better choice. In sum, the sequential g-formula estimator has the potential of allowing nonlinearity in confounders but generally not in the treatment and mediator, and for the flexible plug-in g-formula estimator the converse is true. In this sense these two estimators complement each other.

Finally, the sequential g-estimation procedure changes in a nontrivial way as the specification for $q_A(\cdot)$ and $q_M(\cdot)$ changes, unless one does not care about compatibility and always uses $E(Y^{am} - Y^{0m}) = \varphi a$. The two steps of the procedure must be derived and tailored individually, and become more complex as we move from Model 1 to Model 5 (see Appendix B). In particular, in Model 5 (or whenever there are ML_1 and/or AML_1 interactions), φ_{AM} in SNMM can not be estimated simply by the first-step regression anymore, and the

derivation of the formula is not trivial. Moreover, there is an additional parameter φ_{AML_0} to be estimated. These scenarios make the sequential g-formula estimator inconvenient to use when one would like to build more flexibility into the model. On the other hand, the estimation procedure for the plug-in g-formula works uniformly across different settings, and there is no derivation by hand because the work is done by the computer.

3.5 Sensitivity Analysis

The untestable sequential ignorability conditions in (3.1) is crucial for any g-formula driven estimators. The second part of the assumption is particularly vulnerable in mediation analyses since sequential randomization is not always the case in practice. In this section, we provide a sensitivity analysis for one type of violation of the sequential ignorability conditions. For illustration purposes, this section only shows the sensitivity analysis for Model 1. Similar procedures can be derived for other models in Table 3.1.

Recall the exogenous parents ε's omitted from DAG $\mathcal{G}$ are assumed to be jointly independent in NPSEM-IE. Suppose now ε_M and ε_Y are correlated, then the original g-formula and consequently the flexible plug-in g-formula estimator do not work any more. To identify the CDE in the analysis, we use the following conditions to perform a sensitivity analysis.

1. The unobservable U_0 does not enter the structural equation of Y, i.e. the arrow from U_0 to Y in Figure 1 is deleted.

2. The structural equation for Y is linear in its coefficients. If we use $f_Y(L_0, A, L_1, M)$ to denote the structural equation for Y, then in Model 1, the linearity assumption means $f_Y(\cdot) = \gamma_0 + \gamma_{L_0} L_0 + \gamma_A A + \gamma_{L_1} L_1 + \gamma_M M + \varepsilon_Y$, which is stronger than a linear estimation equation.

3. The structural equation for M is additive separable in ε_M, i.e. $M = f_M(L_0, A, L_1) + \varepsilon_M$ for some function f_M.

Given these assumptions, the counterfactual Y^{am} is

$$Y^{am} = \gamma_0 + \gamma_{L_0} L_0 + \gamma_A a + \gamma_{L_1} L_1 + \gamma_M m + \varepsilon_Y,$$

and the CDE is

$$E\left(Y^{1m} - Y^{0m}\right) = \gamma_A.$$

To consistently estimate γ_A, let $\mathbf{Z} = (1, L_0, A, L_1, M)$ and $\boldsymbol{\gamma}_Z = (\gamma_0, \gamma_{L_0}, \gamma_A, \gamma_{L_1}, \gamma_M)$. Then we can write

$$Y = \mathbf{Z}\boldsymbol{\gamma}_Z + \varepsilon_Y.$$

Define $\boldsymbol{\gamma}_Z^{OLS} = \left[E(\mathbf{Z}'\mathbf{Z})\right]^{-1} E(\mathbf{Z}'Y)$. Because ε_M and ε_Y are correlated, in general $\boldsymbol{\gamma}_Z^{OLS}$ is not equal to $\boldsymbol{\gamma}_Z$. But under conditions (1-3), $\boldsymbol{\gamma}_Z$ can be derived through a bias correction term:

$$\boldsymbol{\gamma}_Z = \boldsymbol{\gamma}_Z^{OLS} - \left[E(\mathbf{Z}'\mathbf{Z})\right]^{-1} \begin{bmatrix} \mathbf{0} \\ \\ \sigma_{\varepsilon_M \varepsilon_Y} \end{bmatrix}$$

where $\sigma_{\varepsilon_M \varepsilon_Y}$ is the covariance between ε_M and ε_Y, and $\mathbf{0}$ is a 4×1 vector. (See Appendix D for derivation.)

To see how sensitive the flexible plug-in g-formula estimator is to this particular violation of sequential ignorability conditions, we let $\sigma_{\varepsilon_M \varepsilon_Y}$ vary within some range and estimate the CDE accordingly. As a rule of thumb, the covariance between M and Y could provide a reference on the choice of the range for $\sigma_{\varepsilon_M \varepsilon_Y}$, since ε_M (ε_Y) only represents part of the variation in M (Y) if one believes that the chosen model is a sound one.

Compared to the sensitivity analysis in Imai, Keele and Tingley (2010), the sensitivity analysis in this section relaxes the structural assumption on the mediator. One major reason that this relaxation can be made is that CDE is the parameter of interest in this paper instead of the natural direct effect, which is not nonparametrically identified in a model with post-treatment confounders.

3.6 An Application

In a longitudinal study by Breslau, Johnson and Lucia (2001) and Breslau, Paneth and Lucia (2004), a random sample of low birthweight (LBW, < 2500 grams) and normal birthweight (NBW, ≥ 2500 grams) infants is selected from two socioeconomically disparate populations in southeast Michigan and followed over 17 years. The goal is to study the long term impact of LBW on academic achievements. The first assessment occurs when the children are 6 years old, the second assessment occurs when the children are 11, and the last assessment when the children are 17. A test from the Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R) by Woodcock, Johnson and Mather (1990) is used to measure their academic achievement in reading at ages 11 and 17. The WJ-R tests are age standardized with a mean of 100 and a standard deviation of 15. An earlier paper by Breslau, Johnson and Lucia (2001) found that the reading score for LBW children at age 11 is 3.6 points lower than those of NBW children. However, the difference became trivial and insignificant after adjusting for their IQ, visual-motor-integration (VMI) function from Beery (1989) and phonologic awareness (PA) from Rosner and Simon (1971) at age 6. Their conclusion is thus that the deficit in reading score in LBW children at age 11 relative to NBW children is accounted for (mediated) mostly by the deficit in their cognitive skills at age 6. In the follow-up study Breslau, Paneth and Lucia (2004), a similar conclusion is obtained for reading score at age 17.

In this application, we estimate the CDE of LBW on reading scores at age 17 when a behavior problem index is used as the mediator. The behavior problem index is constructed by summing up 8 binary indicators for different behavior problems at age 17, including ever smoked a cigarette, ever smoked cigarettes daily, ever used alcohol, ever used marijuana, ever used cocaine, ever used crack, ever used any hallucinogen, and ever used inhalants. The index ranges from 0 to 8. Based on Breslau, Johnson and Lucia (2001); Breslau, Paneth and Lucia (2004); Luo et al. (2014), we put the subject's gender and residence at birth and mother's IQ, education and marital status as the baseline confounders in L_0. For post-

treatment confounders in L_1, the subject's IQ, VMI and PA at age 6 were used. After deleting observations that contain missing values, we obtain a sample of 704 complete cases out of the total 713 who were assessed at age 17. The five models in Table 3.1 were applied, and the results are presented in Figure 3.3.

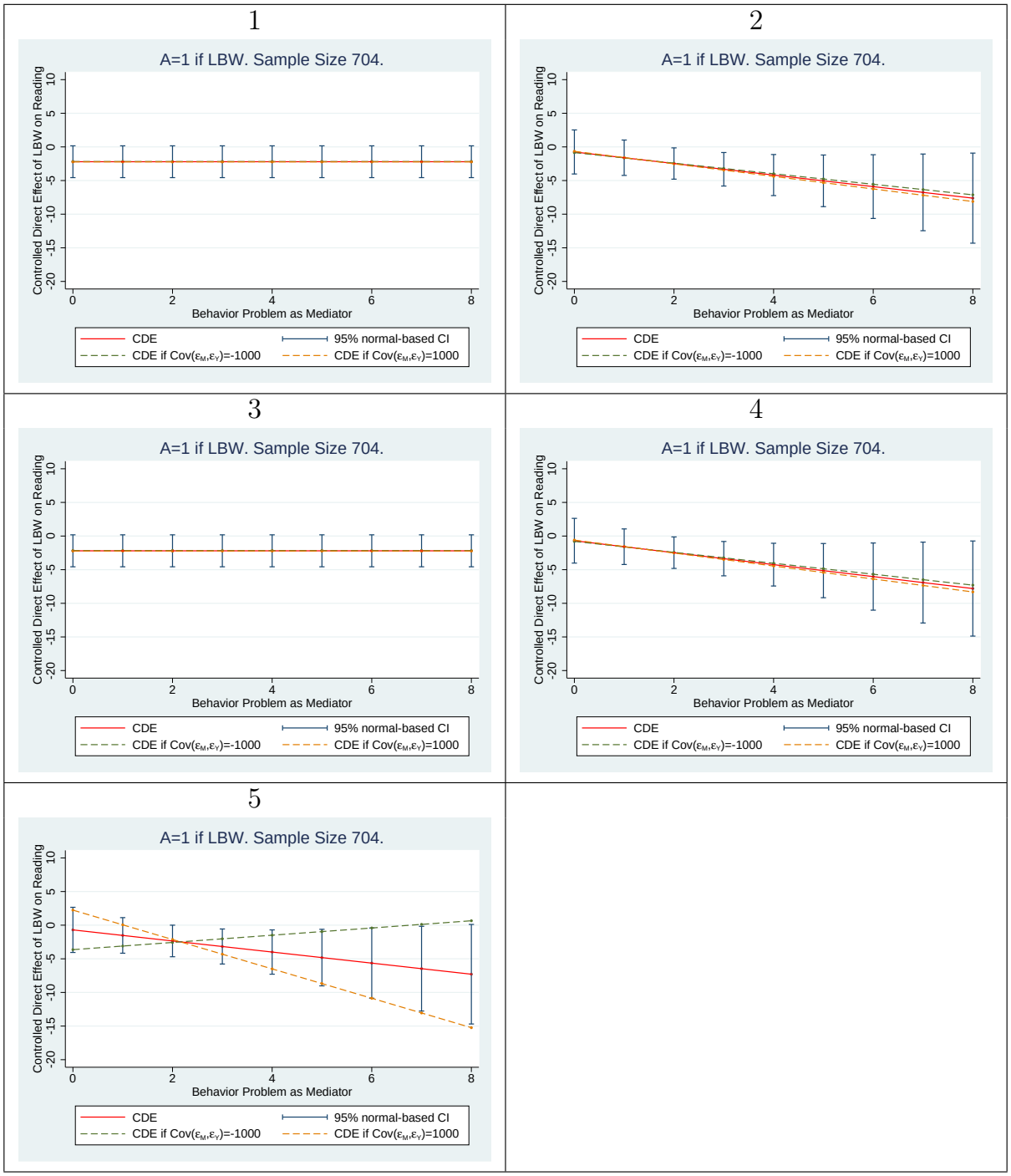
The results for Model 1 show a negative constant CDE estimate, and the effect is not statistically significant according to the normal-based bootstrap confidence intervals. When the interaction $A \times M$ is added as in Model 2, the CDE estimates show a downward trend as the behavior problem became more severe. Specifically, the CDE estimate decreases from $-.75$ to -7.62 as the mediating behavior problem changes from 0 to 8. The effect is significant when the number of behavior problems is greater than 1. Model 3 includes the $M \times L_0$ interaction, but the result is almost identical to Model 1 because the resulting SNMM is not a function of m. Model 4 has both $A \times M$ and $M \times L_0$ interactions, and its result is similar to Model 2 but with slightly wider confidence intervals. Finally, Model 5 included $M \times L_1$ interactions on top of Model 4 but led to similar results.

The downward trends in Model 2, 4 and 5 mainly comes from the negative effect of the $A \times M$ interaction, although this interaction is not statistically significant. Ignoring other channels, this negative interaction effect essentially indicated that even if the immediate effect of the behavior problem on reading is shut down by controlling the behavior problem, a more severe behavior problem would still exacerbate the negative effect of LBW on reading by altering the mechanism through which LBW exerts its effect.

3.7 Conclusion

In this chapter, wee formalize the idea of using partially linear conditional mean models of the outcome and propose a flexible plug-in g-formula estimator in for controlled direct effects causal mediation analysis. Partial linearity of outcome conditional expectation is of interest because under this linear assumption, we can replace the confounders in the conditional

Figure 3.3 Controlled direct effect of LBW on reading with bad behavior as mediator, Model 1 to 5



mean model for the outcome by properly fitted values of the confounders, which results in a plug-in estimator for the controlled direct effects. The flexible plug-in g-formula is closed-form and thus can save some computation time by avoiding Monte Carlo integration that the traditional parametric g-formula usually relies on to evaluate integrals.

We also show that under certain conditions the flexible plug-in g-formula estimator is numerically equivalent to the sequential g-formula estimator in the literature. Although the sequential g-formula estimator is supposed to be a parametric version of the g-estimation of structural nested mean models, this equivalence result indicates that it can also be viewed as a particular parametric g-formula. Indeed, since the g-estimation of structural nested models is a semiparametric version of the original g-formula, when stronger parametric assumptions are imposed, we should expect it to come close to the parametric g-formula. Therefore the equivalence result provides a new insight of the connections between parametric g-formula and g-estimation of structural nested mean models.

The interest in the flexible plug-in g-formula estimator is manifold. First, in the linear case, the flexible plug-in g-formula estimator provides an closed-form expression without introducing additional assumptions than those commonly made in empirical studies. In view of the fact that linear regression is often the first choice for modeling continuous outcomes in parametric g-formula, the flexible plug-in g-formula actually imposes no stronger parametric assumption than some of the parametric g-formulae that already exist in the literature.

Second, the flexible plug-in g-formula estimator connects to the sequential g-estimation but may be more straightforward to use. The two estimators also complement each other in giving practitioners choices of reasonable specifications in their context. For example, if there is reason to believe the functional form for Y should be nonlinear in A and M, we can use the flexible plug-in estimator; and on the other hand, if the nonlinearity lies in the confounding factors, we can use the sequential g-formula estimator. In any case, if the results from both methods are similar, the estimates are more likely to be robust.

APPENDICES

APPENDIX A

PROOFS AND ALGEBRA ON G-FORMULA

A.1 A direct proof of the g-formula for $f_{Y^{am}}(y)$

The proof for the g-formula has been given in a series of paper by Robins and his colleagues. Here we repeat it for the continuous outcome case for easy reference. We adopt the convention to use upper case letters for random variables and lower case letters for realizations.

Under sequential ignorability and consistency,

$$f_{Y^{am}}(y) = \sum_{l_0,l_1} f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m) f_{L_1|L_0,A}(l_1|l_0,a) f_{L_0}(l_0)$$

where, e.g., $f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m)$ is the shorthanded notation for the conditional density function of Y given (L_0, A, L_1, M).

Proof. It can be shown that

$$\begin{aligned}
f_{Y^{am}}(y) &= \int f_{Y^{am}|L_0}(y|l_0) f_{L_0}(l_0) dl_0 \\
&= \int f_{Y^{am}|L_0,A}(y|l_0,a) f_{L_0}(l_0) dl_0 \\
&= \int\int f_{Y^{am}|L_0,A,L_1}(y|l_0,a,l_1) f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dl_1 dl_0 \\
&= \int\int f_{Y^{am}|L_0,A,L_1,M}(y|l_0,a,l_1,m) f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dl_1 dl_0 \\
&= \int\int f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m) f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dl_1 dl_0
\end{aligned}$$

where the first and the third equality uses law of total probability, the second uses $Y^{am} \perp\!\!\!\perp A|L_0$, the fourth uses $(Y^{am} \perp\!\!\!\perp M|L_0, A = a, L_1)$, and the last uses the consistency axiom. □

Remark 6. Based on the g-formula for $f_{Y^{am}}(y)$, it follows that the g-formula for $E(Y^{am})$ is

$$\begin{aligned}
E(Y^{am}) &= \int y \left[\int \int f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m) f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dl_1 dl_0 \right] dy \\
&= \int \int \int y f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m) f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dy dl_1 dl_0 \\
&= \int \int \left[\int y f_{Y|L_0,A,L_1,M}(y|l_0,a,l_1,m) dy \right] f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dl_1 dl_0 \\
&= \int \int E(y|l_0,a,l_1,m) f_{L_1|L_0}(l_1|l_0,a) f_{L_0}(l_0) dl_1 dl_0.
\end{aligned}$$

A.2 The flexible plug-in g-formula for $E(Y^{am})$

If sequential ignorability and consistency hold, and the outcome conditional mean is given by

$$E(Y|l_0,a,l_1,m) = h_0(a,m)l_0 + h_1(a,m)l_1 + h(a,m),$$

then

$$E(Y^{am}) = h_0(a,m)E(L_0) + h_1(a,m)E\left[E(L_1|L_0,A=a)\right] + h(a,m).$$

Proof. Under the assumptions, we get

$$\begin{aligned}
E(Y^{am}) &= \iiint y f(y|l_0,a,l_1,m) f(l_1|l_0,a) f(l_0) dy dl_1 dl_0 \\
&= \int \int \left[\int y f(y|l_0,a,l_1,m) dy \right] f(l_1|l_0,a) f(l_0) dl_1 dl_0 \\
&= \int \int E(Y|l_0,a,l_1,m) f(l_1|l_0,a) f(l_0) dl_1 dl_0 \\
&= \int \left[\int E(Y|l_0,a,l_1,m) f(l_1|l_0,a) dl_1 \right] f(l_0) dl_0 \\
&= \int \left\{ \int \left[h_0(a,m)l_0 + h_1(a,m)l_1 + h(a,m) \right] f(l_1|l_0,a) dl_1 \right\} f(l_0) dl_0 \\
&= \int \left[h_0(a,m)l_0 + h_1(a,m)E(L_1|l_0,a) + h(a,m) \right] f(l_0) dl_0 \\
&= h_0(a,m)E(L_0) + h_1(a,m) \int E(L_1|l_0,a) f(l_0) dl_0 + h(a,m) \\
&= h_0(a,m)E(L_0) + h_1(a,m)E\left[E(L_1|L_0,A=a)\right] + h(a,m)
\end{aligned}$$

□

Remark 7. Note that $E\left[E(L_1|L_0, A=a)\right] \neq E(L_1|A=a)$. For example, $L_1 = exp(L_0) + A$. Then $E(L_1|L_0, A=a) = exp(L_0) + a$, and the LHS is $E\left[exp(L_0)+a\right] = E\left[exp(L_0)\right] + a$. But the RHS is $E\left[exp(L_0)+A|A=a\right] = E\left[exp(L_0)|A=a\right] + a$. If $A \perp\!\!\!\perp L_0$, the equality holds.

In general, let the structural model for L_1 be $L_1 = f(L_0, A, \varepsilon_{L_1})$. Then the LHS is $E\left[f(L_0, a, \varepsilon_{L_1})|L_0\right]$. The RHS is $E\left[f(L_0, a, \varepsilon_{L_1})|A=a\right]$. If we assume ε_{L_1} is independent of (L_0, A), the LHS becomes $E\left[f(L_0, a, \varepsilon_{L_1})\right]$. If in addition $A \perp\!\!\!\perp L_0$, the RHS becomes $E\left[f(L_0, a, \varepsilon_{L_1})\right]$ and the equality holds.

APPENDIX B

PROOFS AND ALGEBRA ON SEQUENTIAL G-ESTIMATOR

B.1 Validity of the second step of sequential g-estimator

The proof of the validity of the sequential g-estimator in the Appendix of Vansteelandt (2009) needs to be extended because L_0 is included in our analysis. Specifically, we want to show that, in presence of L_0, and under additive separability, $q_M(l_0, a, l_1, 0; \gamma) = 0$, and sequential ignorability,

$$E\left[Y - q_M(L_0, A, L_1, M; \gamma)|L_0, A\right] = E\left[Y^{A0}|L_0\right],$$

i.e. for any (l_0, a), show

$$E\left[Y - q_M(L_0, A, L_1, M; \gamma)|L_0 = l_0, A = a\right] = E\left[Y^{a0}|L_0 = l_0\right]. \tag{B.1}$$

Proof. Under the assumptions, we have

$$E\left[Y - q_M(l_0, a, l_1, m; \gamma)|L_0 = l_0, A = a, L_1 = l_1, M = m\right] = q_A(l_0, a, l_1; \gamma).$$

The last equality holds for any m. Therefore

$$E\left[Y - q_M(l_0, a, l_1, M; \gamma)|L_0 = l_0, A = a, L_1 = l_1, M\right] = q_A(l_0, a, l_1; \gamma).$$

Take expectation of both sides conditional on $L_0 = l_0, A = a, L_1 = l_1$ and use iterated expectation, we get

$$E\left[Y - q_M(l_0, a, l_1, M; \gamma)|L_0 = l_0, A = a, L_1 = l_1\right] = q_A(l_0, a, l_1; \gamma).$$

Now, sequential ignorability implies

$$\begin{aligned}
&E\left[Y^{a0}|L_0 = l_0, A = a, L_1 = l_1\right] \\
=&E\left[Y^{a0}|L_0 = l_0, A = a, L_1 = l_1, M = 0\right] \\
=&E\left[Y|L_0 = l_0, A = a, L_1 = l_1, M = 0\right] \\
=&E\left[q_A(L_0, A, L_1; \gamma) + q_M(L_0, A, L_1, M; \gamma)|L_0 = l_0, A = a, L_1 = l_1, M = 0\right] \\
=&E\left[q_A(l_0, a, l_1; \gamma) + q_M(l_0, a, l_1, 0; \gamma)|L_0 = l_0, A = a, L_1 = l_1, M = 0\right] \\
=&q_A(l_0, a, l_1; \gamma)
\end{aligned}$$

The equality holds for any l_1. Therefore

$$E\left[Y - q_M(l_0, a, L_1, M; \gamma)|L_0 = l_0, A = a, L_1\right] = E\left[Y^{a0}|L_0 = l_0, A = a, L_1\right].$$

Take expectation of both sides conditional on $(L_0 = l_0, A = a)$, we get

$$E\left[Y - q_M(l_0, a, L_1, M; \gamma)|L_0 = l_0, A = a\right] = E\left[Y^{a0}|L_0 = l_0, A = a\right].$$

Lastly, notice that $Y^{a0} \perp\!\!\!\perp A|L_0$, we have

$$E\left[Y - q_M(l_0, a, L_1, M; \gamma)|L_0 = l_0, A = a\right] = E\left[Y^{a0}|L_0 = l_0\right].$$

□

We provide an alternative proof below.

Proof. (alternative proof) First, we show that

$$E\left[Y - q_M(L_0, A, L_1, M; \gamma)|L_0, A, L_1, M\right] = E\left(Y^{A0}|L_0, A, L_1\right).$$

Specifically,

$$
\begin{aligned}
& E\left[Y - q_M(L_0, A, L_1, M; \gamma) | L_0, A, L_1, M\right] \\
=& q_A(L_0, A, L_1; \gamma) \\
=& E\left[q_A(L_0, A, L_1; \gamma) | L_0, A, L_1, M = 0\right] \\
=& E\left[q_A(L_0, A, L_1; \gamma) + q_M(L_0, A, L_1, 0; \gamma) | L_0, A, L_1, M = 0\right] \\
=& E\left[q_A(L_0, A, L_1; \gamma) + q_M(L_0, A, L_1, M; \gamma) | L_0, A, L_1, M = 0\right] \\
=& E\left[Y | L_0, A, L_1, M = 0\right] \\
=& E\left[Y^{A0} | L_0, A, L_1, M = 0\right] \\
=& E\left[Y^{A0} | L_0, A, L_1\right]
\end{aligned}
$$

where the first equality holds by the definition of $q_A(L_0, A, L_1; \gamma)$, the second equality holds since q_A is a function of L_0, A, L_1, the third equality holds because $q_M(L_0, A, L_1, 0; \gamma) = 0$, the fourth and fifth are simply rewriting, the sixth by consistency, and the last by $Y^{am} \perp\!\!\!\perp M | L_0, A, L_1$.

Then, take expectation of both sides conditional on (L_0, A), we get

$$
E\left[Y - q_M(L_0, A, L_1, M; \gamma) | L_0, A\right] = E\left[Y^{A0} | L_0, A\right].
$$

Lastly, by $Y^{am} \perp\!\!\!\perp A | L_0$, we have

$$
E\left[Y - q_M(L_0, A, L_1, M; \gamma) | L_0, A\right] = E\left[Y^{A0} | L_0\right].
$$

□

B.2 Estimation procedures and interpretations for the sequential g-estimator in Model 1 and in a general setup with up to three-way interactions in $E(Y|L_0, A, L_1, M)$

We discuss Model 1 and Model 5. Model 1 is the simplest specification, so it is used as an example to illustrate the idea. Model 5 is the most general specification, so the discussion

on Model 5 shows that the conclusion applies to all 5 models.

B.2.1 Model 1 (No interaction)

Model assumptions

1. The NPSEM-IE associated with DAG $\mathcal{G}$. (Thus consistency and sequential ignorability hold.)

2. Structural Nested Mean Model:

$$E(Y^{am} - Y^{0m}|l_0) = \varphi_A a. \tag{B.2}$$

3. Conditional mean of the outcome:

$$E(Y|L_0, A, L_1, M) = \gamma_0 + \gamma_{L_0} L_0 + \gamma_A A + \gamma_{L_1} L_1 + \gamma_M M, \tag{B.3}$$

   so that

$$\begin{aligned} q_A(L_0, A, L_1; \gamma) &= \gamma_0 + \gamma_{L_0} L_0 + \gamma_A A + \gamma_{L_1} L_1, \\ q_M(L_0, A, L_1, M; \gamma) &= \gamma_M M. \end{aligned} \tag{B.4}$$

4. Conditional mean of the post-treatment confounder at $A = 0$:

$$f(l_0, 0) \equiv E(L_1|L_0 = l_0, A = 0) = \pi_0 + \pi_{L_0} l_0. \tag{B.5}$$

Estimation procedure

1. Regress Y on $(1, L_0, A, L_1, M)$ and obtain the OLS estimator $\hat{\gamma}_M$ for γ_M. Generate $\hat{Y}_{-M} \equiv Y - \hat{\gamma}_M M$.

2. Regress $\hat{Y}_{-M}$ on $(1, L_0, A)$. Denote by $\hat{\varphi}_A$ the OLS estimator for the coefficient of A. Then $\hat{\varphi}_A$ is a consistent estimator for φ_A. Hence, $\hat{CDE}(m) = \hat{\varphi}_A$.

Validity of the second-step regression (i.e. the consistency of $\hat{\varphi}_A$ for φ_A)

Proof. (i) First of all, we show

$$E(Y^{00}|L_0 = l_0) = \gamma_0 + \gamma_{L_0} l_0 + \gamma_{L_1} f(l_0, 0). \tag{B.6}$$

Under sequential ignorability, consistency, and the specification in (B.3), we have

$$\begin{aligned} &E(Y^{a0}|L_0 = l_0, A = a, L_1 = l_1) \\ =&E(Y^{a0}|L_0 = l_0, A = a, L_1 = l_1, M = 0) \\ =&E(Y|L_0 = l_0, A = a, L_1 = l_1, M = 0) \\ =&\gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{L_1} l_1. \end{aligned} \tag{B.7}$$

The equality holds for any (l_0, a, l_1), and therefore we can write

$$E(Y^{A0}|L_0, A, L_1) = \gamma_0 + \gamma_A A + \gamma_{L_0} L_0 + \gamma_{L_1} L_1 + \gamma_{AL_1} AL_1.$$

Take expectation of both sides conditional on (L_0, A), we get

$$\begin{aligned} &E(Y^{A0}|L_0, A) \\ =&\gamma_0 + \gamma_A A + \gamma_{L_0} L_0 + \gamma_{AL_0} AL_0 + \gamma_{L_1} E(L_1|L_0, A) + \gamma_{AL_1} AE(L_1|L_0, A) \\ =&\gamma_0 + \gamma_A A + \gamma_{L_0} L_0 + \gamma_{AL_0} AL_0 + \gamma_{L_1} f(L_0, A) + \gamma_{AL_1} Af(L_0, A), \end{aligned}$$

i.e. for any (l_0, a),

$$\begin{aligned} &E(Y^{a0}|L_0 = l_0, A = a) \\ =&\gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{AL_0} al_0 + \gamma_{L_1} f(l_0, a) + \gamma_{AL_1} af(l_0, a). \end{aligned}$$

Then the first part of the sequential ignorability implies

$$\begin{aligned} &E(Y^{a0}|L_0 = l_0) \\ =&\gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{AL_0} al_0 + \gamma_{L_1} f(l_0, a) + \gamma_{AL_1} af(l_0, a). \end{aligned}$$

Set $a = 0$, we get

$$E(Y^{00}|L_0 = l_0) = \gamma_0 + \gamma_{L_0} l_0 + \gamma_{L_1} f(l_0, 0).$$

(ii) Secondly, we show

$$\begin{aligned} &E(Y - q_m(L_0, A, L_1, M; \gamma)|L_0 = l_0, A = a) \\ =&\varphi_A a + \gamma_0 + \gamma_{L_0} l_0 + \gamma_{L_1} f(l_0, 0). \end{aligned} \tag{B.8}$$

By setting $m = 0$ in (B.2), we have

$$E(Y^{a0} - Y^{00}|l_0) = \varphi_A a.$$

Then, (B.1) and (B.6) imply

$$\begin{aligned} &E(Y - q_m(L_0, A, L_1, M; \gamma)|L_0 = l_0, A = a) \\ =&E(Y^{a0}|L_0 = l_0) \\ =&E(Y^{a0} - Y^{00}|L_0 = l_0) + E(Y^{00}|L_0 = l_0) \\ =&\varphi_A a + \gamma_0 + \gamma_{L_0} l_0 + \gamma_{L_1} f(l_0, 0). \end{aligned}$$

(iii) Lastly, given (B.5), we have

$$\begin{aligned} &E(Y - q_m(L_0, A, L_1, M; \gamma)|L_0 = l_0, A = a) \\ =&\varphi_A a + \tilde{\gamma}_0 + \tilde{\gamma}_{L_0} l_0, \end{aligned} \tag{B.9}$$

where $\tilde{\gamma}_0 = \gamma_0 + \gamma_{L_1}\pi_0$ and $\tilde{\gamma}_{L_0} = \gamma_{L_0} + \gamma_{L_1}\pi_{L_0}$. Therefore, the OLS estimator $\hat{\varphi}_A$ in the second-step regression is consistent for φ_A. □

Necessity and sufficiency of (B.5) given all the other model assumptions and that L_0 is not binary

Proof. The sufficiency of (B.5) given all the other model assumptions have been shown by (B.9).

To show its necessity, first we rewrite (B.8) as

$$Y_{-M} = \varphi_A A + \ddot{\gamma}_0 + \gamma_{L_0} L_0 + v \tag{B.10}$$

where

$$Y_{-M} \equiv Y - q_m(L_0, A, L_1, M; \gamma) = Y - \gamma_M M,$$

$$\ddot{\gamma}_0 \equiv \gamma_0 + \gamma_{L_1} E\left[f(L_0, 0) - L_0\right],$$

$$\bar{f}(L_0, 0) \equiv \left[f(L_0, 0) - L_0\right] - E\left[f(L_0, 0) - L_0\right],$$

$$\xi \equiv Y_{-M} - E\left[Y_{-M}|A, L_0\right],$$

$$v \equiv \gamma_{L_1} \bar{f}(L_0, 0) + \xi.$$

Since L_0 is not binary, equation (B.10) indicates that $f(l_0, 0)$ must be linear for $\hat{\varphi}_A$ to be consistent for φ_A. We prove this by contradiction. If $f(l_0, 0)$ is nonlinear in l_0, in general we have

$$Cov(L_0, v) = E(L_0 v) = \gamma_{L_1} E\left[L_0 \bar{f}(L_0, 0)\right] \neq 0$$

unless, e.g, $\bar{f}(L_0, 0) = L_0^{-1}$ and $E(L_0^{-1}) = 0$, or some other particular conditions hold by fluke. But then, unless a regression of Y_{-M} on $(1, A, L_0)$ would yield inconsistent estimators for all coefficients. □

B.2.2 General model with a three-way interaction in $E(Y|L_0, A, L_1, M)$

Model assumptions

1. The NPSEM-IE associated with DAG $\mathcal{G}$. (Thus consistency and sequential ignorability hold.)

2. Structural Nested Mean Model:

$$E(Y^{am} - Y^{0m}|l_0) = \varphi_A a + \varphi_{AM} am + \varphi_{AL_0} a l_0 + \varphi_{AML_0} a m l_0 \tag{B.11}$$

.

3. Conditional mean of the outcome:

$$\begin{aligned} & E(Y|L_0, A, L_1, M) \\ =&\gamma_0 + \gamma_A A + \gamma_M M + \gamma_{AM} AM \\ &+ \gamma_{L_0} L_0 + \gamma_{AL_0} AL_0 + \gamma_{ML_0} ML_0 + \gamma_{AML_0} AML_0 \\ &+ \gamma_{L_1} L_1 + \gamma_{AL_1} AL_1 + \gamma_{ML_1} ML_1 + \gamma_{AML_1} AML_1, \end{aligned} \tag{B.12}$$

so that

$$\begin{aligned} & q_A(L_0, A, L_1; \gamma) \\ =&\gamma_0 + \gamma_A A + \gamma_{L_0} L_0 + \gamma_{AL_0} AL_0 + + \gamma_{L_1} L_1 + \gamma_{AL_1} AL_1 \end{aligned}$$

$$\begin{aligned} & q_M(L_0, A, L_1, M; \gamma) \\ =&\gamma_M M + \gamma_{AM} AM + \gamma_{ML_0} ML_0 + \gamma_{AML_0} AML_0 \\ &+ \gamma_{ML_1} ML_1 + \gamma_{AML_1} AML_1 \end{aligned}$$

4. Conditional mean of the post-treatment confounder, which is given by the following equation

$$f(l_0, a) \equiv E(L_1|l_0, a) = \pi_0 + \pi_{L_0} l_0 + \pi_A a + \pi_{AL_0} a l_0. \tag{B.13}$$

Estimation procedure

1. Regress Y on $(1, A, M, AM, L_0, AL_0, ML_0, AML_0, L_1, AL_1, ML_1, AML_1)$ and obtain the OLS estimator $\hat{\gamma}$ for γ. Generate

$$\begin{aligned} \hat{Y}_{-M} \equiv\ & Y - \hat{\gamma}_M M - \hat{\gamma}_{AM} AM - \hat{\gamma}_{ML_0} ML_0 - \hat{\gamma}_{AML_0} AML_0 \\ & - \hat{\gamma}_{ML_1} ML_1 - \hat{\gamma}_{AML_1} AML_1. \end{aligned}$$

2. Regress $\hat{Y}_{-M}$ on $(1, L_0, A, AL_0)$ and obtain the OLS estimators, $\hat{\varphi}_A$ and $\hat{\varphi}_{AL_0}$, for the coefficients of A and AL_0, respectively.

3. Regress L_1 on $(1, L_0, A, AL_0)$ and obtain the OLS estimator $\hat{\pi}$ for π. Then the sequential g-formula estimator for the controlled direct effect at $M = m$ is

$$\widehat{E(Y^{1m} - Y^{0m})} = \hat{\varphi}_A + \hat{\varphi}_{AM} m + \hat{\varphi}_{AL_0} \bar{L}_0 + \hat{\varphi}_{AML_0} m \bar{L}_0,$$

where

$$\hat{\varphi}_{AM} = \hat{\gamma}_{AM} + \hat{\gamma}_{ML_1} \hat{\pi}_A + \hat{\gamma}_{AML_1} (\hat{\pi}_0 + \hat{\pi}_A),$$

$$\hat{\varphi}_{AML_0} = \hat{\gamma}_{AML_0} + \hat{\gamma}_{ML_1} \hat{\pi}_{AL_0} + \hat{\gamma}_{AML_1} \left(\hat{\pi}_{L_0} + \hat{\pi}_{AL_0} \right).$$

Validity of the second step (i.e. the consistency of $\hat{\varphi}_A$ and $\hat{\varphi}_{AL_0}$ for φ_A and φ_{AL_0})

Proof. (i) First of all, we show

$$E(Y^{00} | L_0 = l_0) = \gamma_0 + \gamma_{L_0} l_0 + \gamma_{L_1} f(l_0, 0). \tag{B.14}$$

which is exactly the same as (B.6).

Under sequential ignorability, consistency, and the specification in (B.12), and using a similar argument to that for (B.7), we have

$$\begin{aligned}
&E(Y^{a0} | L_0 = l_0, A = a, L_1 = l_1) \\
=&E(Y^{a0} | L_0 = l_0, A = a, L_1 = l_1, M = 0) \\
=&E(Y | L_0 = l_0, A = a, L_1 = l_1, M = 0) \\
=&\gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{AL_0} a l_0 + \gamma_{L_1} l_1 + \gamma_{AL_1} a l_1.
\end{aligned} \tag{B.15}$$

The equality holds for any (l_0, a, l_1), and therefore we can write

$$\begin{aligned}
&E(Y^{A0} | L_0, A, L_1) \\
=&\gamma_0 + \gamma_A A + \gamma_{L_0} L_0 + \gamma_{AL_0} A L_0 + \gamma_{L_1} L_1 + \gamma_{AL_1} A L_1
\end{aligned}$$

Take expectation of both sides conditional on (L_0, A), we get

$$\begin{aligned}
&E(Y^{A0}|L_0, A)\\
=&\gamma_0 + \gamma_A A + \gamma_{L_0} L_0 + \gamma_{AL_0} AL_0 + \gamma_{L_1} E(L_1|L_0, A) + \gamma_{AL_1} AE(L_1|L_0, A)\\
=&\gamma_0 + \gamma_A A + \gamma_{L_0} L_0 + \gamma_{AL_0} AL_0 + \gamma_{L_1} f(L_0, A) + \gamma_{AL_1} Af(L_0, A),
\end{aligned}$$

i.e. for any (l_0, a),

$$\begin{aligned}
&E(Y^{a0}|L_0 = l_0, A = a)\\
=&\gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{AL_0} al_0 + \gamma_{L_1} f(l_0, a) + \gamma_{AL_1} af(l_0, a).
\end{aligned}$$

Then the first part of the sequential ignorability implies

$$\begin{aligned}
&E(Y^{a0}|L_0 = l_0)\\
=&\gamma_0 + \gamma_A a + \gamma_{L_0} l_0 + \gamma_{AL_0} al_0 + \gamma_{L_1} f(l_0, a) + \gamma_{AL_1} af(l_0, a).
\end{aligned}$$

Set $a = 0$, we get

$$E(Y^{00}|L_0 = l_0) = \gamma_0 + \gamma_{L_0} l_0 + \gamma_{L_1} f(l_0, 0).$$

(ii) Secondly, we show

$$\begin{aligned}
&E(Y - q_m(L_0, A, L_1, M; \gamma)|L_0 = l_0, A = a)\\
=&\varphi_A a + \gamma_0 + \gamma_{L_0} l_0 + \gamma_{L_1} f(l_0, 0). \qquad \text{(B.16)}
\end{aligned}$$

By setting $m = 0$ in (B.11), we have

$$E(Y^{a0} - Y^{00}|l_0) = \varphi_A a + \varphi_{AL_0} al_0.$$

Then, (B.1) and (B.14) imply

$$\begin{aligned}
&E(Y - q_m(L_0, A, L_1, M; \gamma)|L_0 = l_0, A = a)\\
=&E(Y^{a0}|L_0 = l_0)\\
=&E(Y^{a0} - Y^{00}|L_0 = l_0) + E(Y^{00}|L_0 = l_0)\\
=&\varphi_A a + \varphi_{AL_0} al_0 + \gamma_0 + \gamma_{L_0} l_0 + \gamma_{L_1} f(l_0, 0).
\end{aligned}$$

(iii) Lastly, given (B.5), we have

$$\begin{aligned} & E(Y - q_m(L_0, A, L_1, M; \gamma)|L_0 = l_0, A = a) \\ =& \varphi_A a + \varphi_{AL_0} a l_0 + \tilde{\gamma}_0 + \tilde{\gamma}_{L_0} l_0, \end{aligned} \tag{B.17}$$

where $\tilde{\gamma}_0 = \gamma_0 + \gamma_{L_1}\pi_0$ and $\tilde{\gamma}_{L_0} = \gamma_{L_0} + \gamma_{L_1}\pi_{L_0}$. Therefore, in the second-step regression the OLS estimators $\hat{\varphi}_A$ and $\hat{\varphi}_{AL_0}$ are consistent for φ_A and φ_{AL_0}, respectively. □

Necessity and sufficiency of (B.5) for the validity of the second step, given all the other model assumptions and that L_0 is not binary

The same argument as that in D.1 can be used.

Validity of the third step (i.e. the consistency of $\hat{\varphi}_{AM}$ and $\hat{\varphi}_{AML_0}$ for φ_{AM} and φ_{AML_0})

Proof. Given equation (B.12) and $Y^{am} \perp\!\!\!\perp M|L_0, A, L_1$, we have

$$\begin{aligned} & E(Y^{am}|L_0, a, L_1) \\ =& \gamma_0 + \gamma_A a + \gamma_M m + \gamma_{AM} am \\ & + \gamma_{L_0} L_0 + \gamma_{AL_0} a L_0 + \gamma_{ML_0} m L_0 + \gamma_{AML_0} am L_0 \\ & + \gamma_{L_1} L_1 + \gamma_{AL_1} a L_1 + \gamma_{ML_1} m L_1 + \gamma_{AML_1} am L_1. \end{aligned}$$

Take expectation conditional on (L_0, a), and notice $Y^{am} \perp\!\!\!\perp A|L_0$, we have

$$\begin{aligned} & E(Y^{am}|L_0) \\ =& \gamma_0 + \gamma_A a + \gamma_M m + \gamma_{AM} am \\ & + \gamma_{L_0} L_0 + \gamma_{AL_0} a L_0 + \gamma_{ML_0} m L_0 + \gamma_{AML_0} am L_0 \\ & + \gamma_{L_1} E(L_1|L_0, a) + \gamma_{AL_1} a E(L_1|L_0, a) \\ & + \gamma_{ML_1} m E(L_1|L_0, a) + \gamma_{AML_1} am E(L_1|L_0, a). \end{aligned}$$

Hence,

$$
\begin{aligned}
& E(Y^{am} - Y^{0m}|L_0) \\
=& \gamma_A a + \gamma_{AM} am + \gamma_{AL_0} aL_0 + \gamma_{AML_0} amL_0 \\
& + \gamma_{L_1} \left[E(L_1|L_0, a) - E(L_1|L_0, a = 0)\right] + \gamma_{AL_1} aE(L_1|L_0, a) \\
& + \gamma_{ML_1} m \left[E(L_1|L_0, a) - E(L_1|L_0, a = 0)\right] + \gamma_{AML_1} amE(L_1|L_0, a). \qquad \text{(B.18)}
\end{aligned}
$$

Assume $E(L_1|L_0, A) = \pi_0 + \pi_{L_0} L_0 + \pi_A A + \pi_{AL_0} AL_0$, we have

$$
\begin{aligned}
& E(Y^{am} - Y^{0m}|L_0) \\
=& \gamma_A a + \gamma_{AM} am + \gamma_{AL_0} aL_0 + \gamma_{AML_0} amL_0 \\
& + \gamma_{L_1} \left(\pi_A a + \pi_{AL_0} aL_0\right) + \gamma_{AL_1} a \left(\pi_0 + \pi_{L_0} L_0 + \pi_A a + \pi_{AL_0} aL_0\right) \\
& + \gamma_{ML_1} m \left(\pi_A a + \pi_{AL_0} aL_0\right) + \gamma_{AML_1} am \left(\pi_0 + \pi_{L_0} L_0 + \pi_A a + \pi_{AL_0} aL_0\right) \\
=& \left[\gamma_A + \gamma_{L_1} \pi_A + \gamma_{AL_1} (\pi_0 + \pi_A)\right] a \\
& + \left[\gamma_{AM} + \gamma_{ML_1} \pi_A + \gamma_{AML_1} (\pi_0 + \pi_A)\right] am \\
& + \left[\gamma_{AL_0} + \gamma_{L_1} \pi_{AL_0} + \gamma_{AL_1} \left(\pi_{L_0} + \pi_{AL_0}\right)\right] aL_0 \\
& + \left[\gamma_{AML_0} + \gamma_{ML_1} \pi_{AL_0} + \gamma_{AML_1} \left(\pi_{L_0} + \pi_{AL_0}\right)\right] amL_0.
\end{aligned}
$$

Compare the last equation with B.11, we see that

$$\varphi_A = \gamma_A + \gamma_{L_1} \pi_A + \gamma_{AL_1} (\pi_0 + \pi_A) \qquad \text{(B.19)}$$

$$\varphi_{AM} = \gamma_{AM} + \gamma_{ML_1} \pi_A + \gamma_{AML_1} (\pi_0 + \pi_A) \qquad \text{(B.20)}$$

$$\varphi_{AL_0} = \gamma_{AL_0} + \gamma_{L_1} \pi_{AL_0} + \gamma_{AL_1} \left(\pi_{L_0} + \pi_{AL_0}\right) \qquad \text{(B.21)}$$

$$\varphi_{AML_0} = \gamma_{AML_0} + \gamma_{ML_1} \pi_{AL_0} + \gamma_{AML_1} \left(\pi_{L_0} + \pi_{AL_0}\right) \qquad \text{(B.22)}$$

Therefore, we can estimate φ_{AM} and φ_{AML_0} consistently by

$$\hat{\varphi}_{AM} = \hat{\gamma}_{AM} + \hat{\gamma}_{ML_1} \hat{\pi}_A + \hat{\gamma}_{AML_1} (\hat{\pi}_0 + \hat{\pi}_A),$$

$$\hat{\varphi}_{AML_0} = \hat{\gamma}_{AML_0} + \hat{\gamma}_{ML_1} \hat{\pi}_{AL_0} + \hat{\gamma}_{AML_1} \left(\hat{\pi}_{L_0} + \hat{\pi}_{AL_0}\right),$$

where $\hat{\gamma}$ is from the first step regression, and $\hat{\pi}$ is from the third step regression. □

APPENDIX C

NUMERICAL EQUIVALENCE

We prove the numerical equivalence for Model 5 which is the most flexible model in all the five models considered. The proof for the other four models can be obtained in a similar fashion.

To present the results rigorously, we first need to express the two estimators using the same set of notation. Let $X_1 = (1, A, L_0, AL_0)$, $X_2 = (L_1, AL_1)$, $W = (M, AM, ML_0, ML_1)$, $Z = (X, W)$ and $e = Y - E(Y|Z)$. Let

$$\boldsymbol{\gamma} = (\gamma_0, \gamma_A, \gamma_{L_0}, \gamma_{AL_0}, \gamma_{L_1}, \gamma_{AL_1}, \gamma_M, \gamma_{AM}, \gamma_{ML_0}, \gamma_{ML_1})'.$$

Assume the sample size is n. Stack all observations in the matrix denoted by the corresponding bold letters. Under these notations, Model 5 says that

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}.$$

Denote the OLS estimator for $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}$. Then

$$\mathbf{Y} = \mathbf{Z}\hat{\boldsymbol{\gamma}} + \hat{\mathbf{e}}, \ \textit{where} \ \mathbf{Z}'\hat{\mathbf{e}} = \mathbf{0}.$$

Both the flexible plug-in estimator and the SG estimator use the above linear regression in the first step. In addition, the sequential g-formula estimator generates the fitted outcome

$$\begin{aligned} \hat{Y}_{-M} &= Y - \hat{\gamma}_M M - \hat{\gamma}_{AM} AM - \hat{\gamma}_{ML_0} ML_0 - \hat{\gamma}_{ML_1} ML_1 \\ &= Y - X_3 \hat{\gamma}_W \end{aligned}$$

which is free of the effect of M.

Let $\boldsymbol{\beta}_1 = (\beta_0, \beta_A, \beta_{L_0}, \beta_{AL_0})'$ and $\boldsymbol{\beta}_2 = (\beta_{L_1}, \beta_{AL_1})'$. Combine them in $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. Let $\boldsymbol{\pi} = (\pi_0, \pi_A, \pi_{L_0}, \pi_{AL_0})'$. Define $\varepsilon = \hat{Y}_{-M} - E(\hat{Y}_{-M}|X)$, $u = \hat{Y}_{-M} - E(\hat{Y}_{-M}|X_1)$, and

$v = L_1 - E(L_1|X_1)$. Then

$$\hat{\mathbf{Y}}_{-M} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\hat{\mathbf{Y}}_{-M} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u},$$

$$\mathbf{L}_1 = \mathbf{X}_1\boldsymbol{\pi} + \boldsymbol{v}.$$

Denote the linear projection coefficients in the above three equations by $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\pi}}$, respectively. Then

$$\hat{\mathbf{Y}}_{-M} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}, \qquad \mathbf{X}'\hat{\boldsymbol{\varepsilon}} = 0 \tag{C.1}$$

$$\hat{\mathbf{Y}}_{-M} = \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 + \tilde{\boldsymbol{u}}, \qquad \mathbf{X}_1'\tilde{\boldsymbol{u}} = 0 \tag{C.2}$$

$$\mathbf{L}_1 = \mathbf{X}_1\hat{\boldsymbol{\pi}} + \hat{\boldsymbol{v}}. \qquad \mathbf{X}_1'\hat{\boldsymbol{v}} = 0 \tag{C.3}$$

Note that (C.2) represents the second-step regression of the sequential g-estimation.

Finally, let $\boldsymbol{\varphi} = (\varphi_A, \varphi_{AL_0}, \varphi_{AM}, \varphi_{AML_0})'$. Let $\hat{\boldsymbol{\varphi}}$ denote the sequential g-formula estimator, and $\check{\boldsymbol{\varphi}}$ the flexible plug-in estimator. The two estimators are the same in estimating φ_{AM} and φ_{AML_0}:

$$\hat{\varphi}_{AM} = \check{\varphi}_{AM} = \hat{\gamma}_{AM} + \hat{\gamma}_{ML_1}\hat{\pi}_A, \tag{C.4}$$

$$\hat{\varphi}_{AML_0} = \check{\varphi}_{AML_0} = \hat{\gamma}_{AML_0} + \hat{\gamma}_{ML_1}\hat{\pi}_{AL_0} + \hat{\gamma}_{AML_1}\left(\hat{\pi}_{L_0} + \hat{\pi}_{AL_0}\right), \tag{C.5}$$

but they differ in estimating φ_A and φ_{AL_0}:

$$\hat{\varphi}_A = \tilde{\beta}_A, \tag{C.6}$$

$$\hat{\varphi}_{AL_0} = \tilde{\beta}_{AL_0}. \tag{C.7}$$

$$\check{\varphi}_A = \hat{\gamma}_A + \hat{\gamma}_{L_1}\hat{\pi}_A + \hat{\gamma}_{AL_1}\left(\hat{\pi}_0 + \hat{\pi}_A\right), \tag{C.8}$$

$$\check{\varphi}_{AL_0} = \hat{\gamma}_{AL_0} + \hat{\gamma}_{L_1}\hat{\pi}_{AL_0} + \hat{\gamma}_{AL_1}\left(\hat{\pi}_{L_0}\right). \tag{C.9}$$

Theorem 8. *(Numerical Equivalence in Model 5) The flexible plug-in g-formula and the sequential g-estimation are numerically equivalent in Model 5. That is*

$$\hat{\varphi}_A = \check{\varphi}_A,$$

$$\hat{\varphi}_{AL_0} = \check{\varphi}_{AL_0}.$$

Proof. Regarding those linear projection coefficients, the first fact is that

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\gamma}}_X \tag{C.10}$$

where $\hat{\boldsymbol{\gamma}}_X$ is the sub-vector of $\hat{\boldsymbol{\gamma}}$ associated with X. To see why, let $\hat{\boldsymbol{\gamma}}_W$ be the sub-vector of $\hat{\boldsymbol{\gamma}}$ associated with W. Then the regression of $\hat{Y}_{-M}$ on X can equivalently be cast as the restricted regression of Y on $Z = (X, W)$ with the constraint $\boldsymbol{\gamma}_W = \hat{\boldsymbol{\gamma}}_W$, and the later restricted regression is known to yield $\hat{\boldsymbol{\gamma}}_X$.

The second fact is that the orthogonality condition following each of the three equations (C.1), (C.2) and (C.3) is definitional for the corresponding linear projection coefficient vector. Note that the rank condition always hold in practice unless by fluke. Therefore, given a data set, $\hat{\beta}$, $\tilde{\beta}_1$ and $\hat{\pi}$ are uniquely defined by their respective orthogonality conditions.

Now we are ready to prove the equivalence. Plug equation (C.3) into (C.1), we have

$$\begin{aligned}
\hat{\mathbf{Y}}_{-M} &= \mathbf{X}\hat{\beta} + \hat{\boldsymbol{\varepsilon}} \\
&= \hat{\beta}_0 + \hat{\beta}_A\mathbf{A} + \hat{\beta}_{L_0}\mathbf{L}_0 + \hat{\beta}_{AL_0}\mathbf{A}\mathbf{L}_0 + \hat{\beta}_{L_1}\mathbf{L}_1 + \hat{\beta}_{AL_1}\mathbf{A}\mathbf{L}_1 + \hat{\boldsymbol{\varepsilon}} \\
&= \hat{\beta}_0 + \hat{\beta}_A\mathbf{A} + \hat{\beta}_{L_0}\mathbf{L}_0 + \hat{\beta}_{AL_0}\mathbf{A}\mathbf{L}_0 + \hat{\beta}_{L_1}(\hat{\pi}_0 + \hat{\pi}_A\mathbf{A} + \hat{\pi}_{L_0}\mathbf{L}_0 + \\
&\quad \hat{\pi}_{AL_0}\mathbf{A}\mathbf{L}_0 + \hat{\boldsymbol{v}}) + \hat{\beta}_{AL_1}\mathbf{A}(\hat{\pi}_0 + \hat{\pi}_A\mathbf{A} + \hat{\pi}_{L_0}\mathbf{L}_0 + \hat{\pi}_{AL_0}\mathbf{A}\mathbf{L}_0 + \hat{\boldsymbol{v}}) + \hat{\boldsymbol{\varepsilon}} \\
&= (\hat{\beta}_0 + \hat{\beta}_{L_1}\hat{\pi}_0) + \left[\hat{\beta}_A + \hat{\beta}_{L_1}\hat{\pi}_A + \hat{\beta}_{AL_1}(\hat{\pi}_0 + \hat{\pi}_A)\right]\mathbf{A} + (\hat{\beta}_{L_0} + \hat{\beta}_{L_1}\hat{\pi}_{L_0})\mathbf{L}_0 \\
&\quad + \left[\hat{\beta}_{AL_0} + \hat{\beta}_{L_1}\hat{\pi}_{AL_0} + \hat{\beta}_{AL_1}(\hat{\pi}_{L_0} + \hat{\pi}_{AL_0})\right]\mathbf{A}\mathbf{L}_0 + \left[(\hat{\beta}_{L_1} + \hat{\beta}_{AL_1})\hat{\boldsymbol{v}} + \hat{\boldsymbol{\varepsilon}}\right] \\
&= \hat{\beta}_0^* + \hat{\beta}_A^*\mathbf{A} + \hat{\beta}_{L_0}^*\mathbf{L}_0 + \hat{\beta}_{AL_0}^*\mathbf{A}\mathbf{L}_0 + \hat{\boldsymbol{\varepsilon}}^* \\
&= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^* + \hat{\boldsymbol{\varepsilon}}^*
\end{aligned}$$

where we define

$$\hat{\beta}_0^* = \hat{\beta}_0 + \hat{\beta}_{L_1}\hat{\pi}_0,$$

$$\hat{\beta}_A^* = \hat{\beta}_A + \hat{\beta}_{L_1}\hat{\pi}_A + \hat{\beta}_{AL_1}(\hat{\pi}_0 + \hat{\pi}_A), \tag{C.11}$$

$$\hat{\beta}_{L_0}^* = \hat{\beta}_{L_0} + \hat{\beta}_{L_1}\hat{\pi}_{L_0}$$

$$\hat{\beta}_{AL_0}^* = \hat{\beta}_{AL_0} + \hat{\beta}_{L_1}\hat{\pi}_{AL_0} + \hat{\beta}_{AL_1}(\hat{\pi}_{L_0} + \hat{\pi}_{AL_0}), \tag{C.12}$$

$$\hat{\boldsymbol{\beta}}_1^* = (\beta_0^*, \beta_A^*, \hat{\beta}_{L_0}^*, \hat{\beta}_{AL_0}^*)',$$

$$\hat{\boldsymbol{\varepsilon}}^* = (\hat{\beta}_{L_1} + \hat{\beta}_{AL_1})\hat{\boldsymbol{v}} + \hat{\boldsymbol{\varepsilon}}.$$

But (C.3) and (C.2) imply that

$$\begin{aligned} \mathbf{X}_1'\hat{\boldsymbol{\varepsilon}}^* &= \mathbf{X}_1' \left[(\hat{\beta}_{L_1} + \hat{\beta}_{AL_1})\hat{\boldsymbol{v}} + \hat{\boldsymbol{\varepsilon}} \right] \\ &= (\hat{\beta}_{L_1} + \hat{\beta}_{AL_1})\mathbf{X}_1'\hat{\boldsymbol{v}} + \mathbf{X}_1'\hat{\boldsymbol{\varepsilon}} \\ &= 0. \end{aligned} \tag{C.13}$$

Hence, by definition and uniqueness, $\tilde{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1^*$. In particular, we have

$$\tilde{\beta}_A = \hat{\beta}_A^*,$$

$$\tilde{\beta}_{AL_0} = \hat{\beta}_{AL_0}^*.$$

It follows immediately from (C.6), (C.7), (C.10), (C.11) and (C.12) that

$$\hat{\varphi}_A = \check{\varphi}_A,$$

$$\hat{\varphi}_{AL_0} = \check{\varphi}_{AL_0}.$$

□

Remark 1: In proving $\mathbf{X}_1'\hat{\boldsymbol{\varepsilon}}^* = 0$, it shows why we need the regressors be the same in the model of L_1 as in the model of the second-step regression.

Remark 2: If the statistical model for $E(L_1|A, L_0)$ is of the same level of flexibility as the second step regression of the SG estimator, we say the model is properly chosen. For example, in Model 1, the second step of the SG estimator is to regress $\hat{Y}_{-M,i}$ on $(1, A_i, L_{0i})$, and the proper model for $E(L_1|A, L_0)$ is $E(L_1|A, L_0) = \pi_C + \pi_A A + \pi_{L_0} L_0$. In Model 5, the second step of the SG estimator is to regress $\hat{Y}_{-M,i}$ on $(1, A_i, L_{0i}, AL_{0i})$, and the proper model for $E(L_1|A, L_0)$ should be $E(L_1|A, L_0) = \pi_C + \pi_A A + \pi_{L_0} L_0 + \pi_{AL_0} AL_0$.

If the second-step regression of the SG estimation contains AL_0, but the linear model for $E(L_1|A, L_0)$ excludes AL_0, then the numerical equivalence does not hold any more. The

intuition is that when π_{AL_0} is forced to be zero, it will alter the the estimates for π_0, π_A and π_{L_0}, which can be viewed as some restricted estimates. The math is explained in (C.13): $\mathbf{X}_1^{'}\hat{\boldsymbol{v}}$ is not zero anymore.

APPENDIX D

SENSITIVITY ANALYSIS

In this section we provide the derivation details for the sensitivity analysis in section 5 of main text. Specifically, we want to show that

$$\boldsymbol{\gamma}_Z = \boldsymbol{\gamma}_Z^{OLS} - \left[E(\mathbf{Z}'\mathbf{Z})\right]^{-1} \begin{bmatrix} \mathbf{0} \\ \sigma_{\varepsilon_M \varepsilon_Y} \end{bmatrix}.$$

Proof. Recall that

$$Y = \mathbf{Z}\boldsymbol{\gamma}_Z + \varepsilon_Y, \tag{D.1}$$

$$\mathbf{Z} = (1, L_0, A, L_1, M),$$

$$\sigma_{\varepsilon_M \varepsilon_Y} = Cov(\varepsilon_M, \varepsilon_Y) \neq 0.$$

Premultiply both sides of equation (D.1) by X', take expectation,, we have

$$E(\mathbf{Z}'Y) \;\; = \;\; E(\mathbf{Z}'\mathbf{Z})\boldsymbol{\gamma}_Z + E(\mathbf{Z}'\varepsilon_Y).$$

Assume $E(\mathbf{Z}'Y)$ has full rank, then

$$\boldsymbol{\gamma}_Z = \left[E(\mathbf{Z}'\mathbf{Z})\right]^{-1} E(\mathbf{Z}'Y) - \left[E(\mathbf{Z}'\mathbf{Z})\right]^{-1} E(\mathbf{Z}'\varepsilon_Y)$$

Now notice that the three additional assumptions in Section 5 imply

$$E[(1, L_0, A, L_1)'\varepsilon_Y] = \mathbf{0}. \tag{D.2}$$

where $\mathbf{0}$ is a 4×1 zero vector. Meanwhile recall that $\boldsymbol{\gamma}_Z^{OLS} = \left[E(\mathbf{Z}'\mathbf{Z})\right]^{-1} E(\mathbf{Z}'Y)$. It follows immediately that

$$\boldsymbol{\gamma}_Z = \boldsymbol{\gamma}_Z^{OLS} - \left[E(\mathbf{Z}'\mathbf{Z})\right]^{-1} \begin{bmatrix} \mathbf{0} \\ \sigma_{\varepsilon_M \varepsilon_Y} \end{bmatrix}.$$

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Beery, Keith E. 1989. *Developmental test of visual-motor integration: Administration and scoring manual.* Cleveland, OH:Modern Curriculum Pr.

Breslau, Naomi, Eric O Johnson, and Victoria C Lucia. 2001. "Academic achievement of low birthweight children at age 11: the role of cognitive abilities at school entry." *Journal of Abnormal Child Psychology*, 29(4): 273–279.

Breslau, Naomi, Nigel S Paneth, and Victoria C Lucia. 2004. "The lingering academic deficits of low birth weight children." *Pediatrics*, 114(4): 1035–1040.

Danaei, Goodarz, An Pan, Frank B Hu, and Miguel a Hernán. 2013. "Hypothetical midlife interventions in women and risk of type 2 diabetes." *Epidemiology*, 24(1): 122–8.

Daniel, Rhian M, Bianca L De Stavola, and Simon N Cousens. 2011. "gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula." *Stata Journal*, 11(4): 479.

Garcia-Aymerich, Judith, Raphaëlle Varraso, Goodarz Danaei, Carlos A. Camargo, and Miguel A. Hernán. 2014. "Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity: An application of the parametric G-Formula." *American Journal of Epidemiology*, 179(1): 20–26.

Hernan, Miguel A, and ames M Robins. 2015. *Causal Inference.* Boca Raton:Chapman & Hall/CRC.

Hicks, Raymond, and Dustin Tingley. 2011. "Causal mediation analysis." *Stata Journal*, 11(4): 605.

Horvitz, Daniel G, and Donovan J Thompson. 1952. "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association*, 47(260): 663–685.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A general approach to causal mediation analysis." *Psychological Methods*, 15(4): 309.

Lajous, Martin, Walter C. Willett, James Robins, Jessica G. Young, Eric Rimm, Dariush Mozaffarian, and Miguel A. Hernán. 2013. "Changes in fish consumption in midlife and the risk of coronary heart disease in men and women." *American Journal of Epidemiology*, 178(3): 382–391.

Luo, Zhehui, Joshua Breslau, Joseph C Gardiner, Qiaoling Chen, and Naomi Breslau. 2014. "Assessing interchangeability at cluster levels with multiple-informant data." *Statistics in medicine*, 33(3): 361–375.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference.* Cambridge University Press.

Pearl, Judea. 2014. "Interpretation and identification of causal mediation." *Psychological methods*, 19(4): 459.

Pearl, Judea, and James Robins. 1995. "Probabilistic evaluation of sequential plans from causal models with hidden variables." 444–453, Morgan Kaufmann Publishers Inc.

Richardson, Thomas S, and James M Robins. 2013. "Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality." *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128.

Robins, James M. 1986. "A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect." *Mathematical Modelling*, 7(9-12): 1393–1512.

Robins, James M. 1989. "The control of confounding by intermediate variables." *Statistics in Medicine*, 8(6): 679–701.

Robins, James M. 1994. "Correcting for non-compliance in randomized trials using structural nested mean models." *Communications in Statistics-Theory and methods*, 23(8): 2379–2412.

Robins, James M. 1998. "Structural nested failure time models." *Encyclopedia of Biostatistics.*

Robins, James M. 2000. "Marginal structural models versus structural nested models as tools for causal inference." In *Statistical models in epidemiology, the environment, and clinical trials*. 95–133. Springer.

Robins, James M, and Sander Greenland. 1992. "Identifiability and exchangeability for direct and indirect effects." *Epidemiology*, 143–155.

Robins, James M, and Thomas Richardson. 2010. "Alternative graphical causal models and the identification of direct effects." *Causality and psychopathology: Finding the determinants of disorders and their cures*, 103–158.

Robins, J M, M A Hernán, and U W E SiEBERT. 2004. "Effects of multiple interventions." *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*, 1: 2191–2230.

Rosner, Jerome, and Dorothea P Simon. 1971. "The Auditory Analysis Test An Initial Report." *Journal of Learning Disabilities*, 4(7): 384–392.

Taubman, Sarah L, James M Robins, Murray A Mittleman, and Miguel A Hernán. 2009. "Intervening on risk factors for coronary heart disease: an application of the parametric g-formula." *International Journal of Epidemiology*, 38(6): 1599–1611.

Valeri, Linda, and Tyler J Vanderweele. 2013. "Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros." *Psychological Methods*, 18(2): 137–50.

Van der Wal, W M, M Prins, B Lumbreras, and R B Geskus. 2009. "A simple G-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease." *Statistics in Medicine*, 28(18): 2325–2337.

Vansteelandt, Stijn. 2009. "Estimating direct effects in cohort and case-control studies." *Epidemiology*, 20(6): 851–860.

Westreich, Daniel, Stephen R Cole, Jessica G Young, Frank Palella, Phyllis C Tien, Lawrence Kingsley, Stephen J Gange, and Miguel A Hernán. 2012. "The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death." *Statistics in Medicine*, 31(18): 2000–2009.

Woodcock, Richard W, M Bonner Johnson, and Nancy Mather. 1990. *Woodcock-Johnson psycho-educational battery–Revised.* DLM Teaching Resources.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* . 2nd ed., Boston MA:MIT Press.

Young, Jessica G, Lauren E Cain, James M Robins, Eilis J O'Reilly, and Miguel A Hernán. 2011. "Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula." *Statistics in biosciences*, 3(1): 119–143.