UC Davis
UC Davis Electronic Theses and Dissertations

Title
Effect Decomposition and Heterogeneity in Development Economics and Policy Evaluation

Permalink
https://escholarship.org/uc/item/9758g02q

Author
Luo, Xiaoman

Publication Date
2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Effect Decomposition and Heterogeneity in Development Economics and Policy Evaluation

By

XIAOMAN LUO

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

AGRICULTURAL AND RESOURCE ECONOMICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Ashish Shenoy, Chair

J. Edward Taylor

Stephen R. Boucher

Committee in Charge

2021

© Xiaoman Luo, 2021. All rights reserved.

To my family and friends.

Contents

List of Figures

List of Tables

Abstract

This dissertation covers various topics ranging from migration to wealth measures in developing countries and to policy analysis in general. In Chapter 1, I study the internal migration in China and aim to understand its effect on the human capital accumulation of the later generations. In Chapter 2, I investigate novel wealth measures and look for ways to understand the low-cost wealth metrics. In Chapter 3, I discuss a doubly-robust method to estimate causal effects for panel data in the presence of effect heterogeneity.

In Chapter 1, I investigate the mechanisms through which parental migration affects the schooling outcomes of children left behind in rural China. This issue affects 61 million children. Previous literature on this topic focuses on estimating the net effect of migration, whereas this paper disentangles the net effect into different mechanisms of policy interests. I establish a theoretical framework to incorporate three essential and widely-studied mechanisms that migration could affect left-behind children's school performance: parental accompaniment, child's study time, and investment in children. Motivated by the theoretical model's solution, I apply the structural equation model to estimate the influence through different mechanisms. I propose an identification strategy based on instrumental variables and the Heckman selection model. Using the model on rural household survey data from nine provinces, I find that the effects through parental absence and investment are both significantly negative with large sizes. In contrast, the impact through the child's study time is insignificant with a negligible size. The surprising negative effect through investment is mainly driven by reduced nutrition investment by the de facto custodians, who may not have compatible incentives to allocate the remittances to the child. Through a refined subgroup analysis, I find that girls are suffering ten times more from the underinvestment than boys, revealing a shocking gender inequality in rural China.

In Chapter 2 (co-authored with Professor Ashish Shenoy), I aim to understand novel poverty measures. For under-developed countries, wealth measures are essential for measuring economic growth, policy design, and setting development goals. In particular, I focus on the use of nighttime

light data. Unlike standard wealth measures based on national accounts or expenditure surveys, nighttime light data has the advantages of high frequency, low cost, and precision over small geographic units. These advantages make it an ideal substitute for in-person surveys. Nighttime light data has been a popular wealth measure in recent years, and previous papers mainly argue that nighttime light intensity and gross domestic production levels are highly correlated globally. However, we analyze the relationship between night light growth and economic growth for 179 countries or regions and find heterogeneity in the correlation. To deal with the heterogeneity, we propose a weighted least squares estimator for the average correlation coefficient by properly re-weighting each country. We find a significant and positive average correlation among middle-income countries. Moving beyond the average association, we apply the LASSO regression to identify and estimate non-zero individual correlation coefficients. This is inspired by the sparsity of country-level associations observed in the preliminary analysis. We further apply the "knockoff" method to control the false discovery rate among the selected countries.

In Chapter 3 (co-authored with Professor Dmitry Arkhangelsky, Professor Guido W. Imbens, and Lihua Lei), we develop a novel method for causal inference with observational panel data, which overcomes the limitations of existing methods. Cross-sectional models account for treatment assignment using methods such as inverse probability weighting. We extend this approach to panel data. Taking the case of staggered adoption as an example, we model the adoption time with duration models such as the Cox hazards model. As long as the information about the assignment mechanism is accurate, our method works under substantially weaker assumptions than the traditional methods. As a byproduct, we characterize the class of experimental designs under which the conventional methods are guaranteed to produce consistent estimates of the causal effects. The method from our paper can be widely applied to empirical analysis, such as program evaluation.

Acknowledgments

First, it is a genuine pleasure to express my deep sense of thanks and gratitude to my Committee: Prof. Ashish Shenoy, Prof. J. Edward Taylor, and Prof. Stephen R. Boucher. Throughout my Ph.D. program, their invaluable guidance, unwavering support, and patience have encouraged me in my academic studies. I would especially like to thank Prof. Ashish Shenoy, the chair of my committee. He offered me the first opportunity to join his research, and always has the sharpness to discover the research questions and the vast expertise to tackle them. His approach to motivating research ideas has a significant impact on the way I conduct economic research. Working with Prof. J. Edward Taylor has been really enjoyable. His research scope and experiences have enlightened me tremendously. I am also thankful for Prof Stephen R. Boucher for his attention to detail in our discussion of my research projects. During our conversations, I gain a better knowledge of modeling.

I would also like to express my sincere gratitude to other professors, in particular, Prof. Michael R. Carter, Prof. Huang Chen, Prof. Xiaomeng Cui, Prof. Travis J. Lybbert, Prof. Pierre Mérel, Prof. Diana Moreira, Prof. Giovanni Peri, Prof.Arman Rezaee, Prof. Manisha Shah, Prof. Monica Singal, Prof. Stephen Vosti, Prof. Wesley Yin for their insightful comments on my research and career choices. I am also grateful to my coauthors Prof. Dmitry Arkhangelsky, Prof. Guido W. Imbens, Edward Whitney, and Heng Zhu. It was such a pleasure to work with them, and I learned a lot from them.

My gratitude extends to all the faculties and staff at our department and the Ph.D. program. In particular, I would like to thank Arnon Erba, Jeff Goettsch, Christy Hansen, and Laurie Warren for their assistance with any administrative concerns. I would like to thank all my friends and cohorts in the program. It is because of their generous assistance and companionship that my time in the program has been so enjoyable.

My students at Changjiang Village, Ganzhou, Jiangxi Province, China, deserve special recognition. I would not have studied development economics and investigated the topic of migrating parents and left-behind children if they hadn't shared with me their experiences ten years ago.

Finally, I would like to express my gratitude to my family. I wish to thank my parents. I would not have been able to complete my study without their unconditional support, encouragement, and understanding. Most importantly, I want to express my gratitude to my husband, Lihua Lei, who has been my best friend of all time. I am grateful that we are always learning from each other and making progress together.

CHAPTER 1

How Does Parental Out-migration Affect Left-behind Children's Schooling Outcomes?

1.1. Introduction

1.1.1. Background. Labor migrants represent a substantial share of the workforce in many developing countries. According to Human Development Report 2009 [Klugman, 2009], there are more than 740 million internal migrants who live and work outside their region of birth within their home country — approximately 2.6 times as many as international migrants, based on the estimates by World Migration Report 2020. Due to various kinds of mobility constraints, family members of migrants are often left behind, facing the potential adverse effects emotionally and physically [Waddington, 2003]. The separation could have non-negligible consequences on the education and health of left-behind family members.

In this paper, I focus on left-behind children of internal rural migrants in China. According to the 2010 Population Census of China, more than 61 million children are left behind in rural China by migrant parents, accounting for 37.7% of children in rural areas, and 21.88% of children in China overall. The massive number of left-behind children is a consequence of the *hukou* system, the household registration system in China. There are two types of *hukou* in China: rural and urban *hukou*, and it has been difficult to transfer from one type to the other. Prior to the 1970's, people with rural *hukou* were legally prohibited from migrating to urban areas. Since the late 1970's, to meet the huge labor demand in urban areas generated by the Chinese economic reform, the government has gradually relaxed the restriction on the *hukou* system to permit migration from rural to urban areas. Nevertheless, transfers of the *hukou* status remains highly restrictive — rural migrants and their families with rural *hukou* are generally excluded from the social benefits that urban citizens enjoy. In particular, children of rural migrants have limited access to free public schools, health care benefits, housing support, social security, and other resources. Children who migrate with their parents from rural to urban areas can either choose expensive private schools,

or less costly "migrant schools" run by local entrepreneurs, typically with unsatisfactory quality of education. As a consequence, most migrant parents choose to leave children behind in their hometown.

Considering the sizable population of left-behind children in rural China, the effect of parental migration on left-behind children's educational outcomes has considerable implications on the accumulation of human capital in China. Although the effect may not be reflected in the short-term household livelihood, it is highly indicative of the future human capital, which directly affects the poverty level. Therefore, it is of paramount importance to assess the impact and design policies to mitigate the negative effects and amplify the positive effects, if any.

Despite the rich literature on internal migration in different developing countries [e.g. Arnold and Shah, 1984, Booth, 1995, Battistella and Conaco, 1998, Ganepola, 2002, Afsar, 2003, Maruja and Baggio, 2003, Edwards and Ureta, 2003, Mendoza, 2004, Adams Jr and Page, 2005, Bryant, 2005, Gupta et al., 2009, Arguillas and Williams, 2010, Antman, 2011, McKenzie and Rapoport, 2011, Graham and Jordan, 2011, Antman, 2013], including China [He et al., 2012, Chang et al., 2011, Chen, 2013, Zhao et al., 2014, Sun et al., 2015], existing works almost exclusively focus on the net effect of parental migration on left-behind children's schooling outcomes. While the net effect is scientifically meaningful, it is less informative for policymakers — even if there is a large negative net effect, it is neither efficient nor ethical to directly impose restrictions on migration because education is not the only important factor for social welfare. A more realistic approach is to design policies targeting specific mechanisms that are more manipulable than the migration per se, through which the migration affects the left-behind children's educational outcomes. To achieve this, the first step is to disentangle the net effect into different causal channels to assess their respective importance [Huber, 2016].

1.1.2. Contributions. In line with the literature, I investigate three widely-studied mechanisms — parental absence, child's time allocation, and investment in the child [Démurger, 2015]. In Section 1.5.3, I further decompose the investment into two sub-mechanisms — nutritional spending and tuition spending. Each of these mechanism has been partially assessed in various countries. Nevertheless, previous studies in China and other developing countries almost exclusively focus on the net effect of each channel separately without taking their high correlation into account. A

significant net effect of an unimportant mechanism may be purely contributed by another important mechanism that is correlated but missing in the analysis. In principle, simultaneous analysis of different mechanisms mitigates the bias due to correlation, thereby providing a more convincing comparison of potential policy targets.

To lay the foundation for the simultaneous analysis, I establish a simple two-agent model to model the decision-making processes of the child and parents jointly. The child's decision variable is the study time; the parents' decision variable is migration; and both agents are maximizing a weighted average of utilities in the present and in the future. Under reasonable assumptions on the utility functions and production function of human capital, both the child and parents face a trade-off between the utilities in two periods, leading to a non-trivial equilibrium. The equilibrium solution reveals how the child's educational outcome is affected by migration directly through parental absence and indirectly through time allocation and monetary investment.

Beyond the lack of simultaneity, most of the existing studies have another important limitation. The effect through an intermediate variable has two components: the effect of migration on that variable, and the effect of that variable on the educational outcome. To the best of my knowledge, all previous works estimate one component only, which provides an incomplete answer. I summarize a selective set of works in Section 1.1.3. For instance, Chang et al. [2011] and Chen [2013] find that children of migrants tend to spend more time on housework and thus less time on studying. However, this conclusion on its own does not prove the effectiveness of a policy that increases child's study hours, unless one can further show that increased hours has a positive effect on the schooling performance, which may or may not be true depending on the relative effect size of stress or fatigue. Similarly, Kandel and Kao [2000] find that high remittances sent back by migrants may decrease the child's schooling performance. However, this is insufficient to guide policymakers because the negative effect can be either attributed to that the remittances are not allocated to the child by the de facto custodian or that the remittances increase child's desire to work and reduce their aspiration to study.

I overcome this limitation through the mediation analysis, which can provide estimates of both components, thereby providing a fuller description of the effect through different channels. Mediation analysis is a standard technique to decompose the net effect of a treatment into a direct effect and indirect effects through different causal mechanisms. It has been popular for decades in

psychology [Baron and Kenny, 1986], and was advocated recently in economics [e.g. Heckman et al., 2013, Heckman and Pinto, 2015, Huber, 2016, 2019, Celli, 2019]. In my problem, the treatment is the migration decision of parents, the direct effect is given by the absence of parents, and the indirect effects are given by the child's time allocation and monetary investment in the child. Notably, the equations involved in the mediation analysis coincide with the equilibrium solution of my two-agent model under certain functional specification, rendering the empirical analysis coherent with the theoretical analysis. In this problem, the mediation analysis is complicated by unmeasured confounders and non-random missing mediators, namely child's study time and monetary investment in the child. I propose a generic identification strategy to handle these two sources of endogeneity simultaneously.

Applying the mediation analysis on the Rural-Urban Migration in China (RUMiC) survey, I find significantly negative direct effects of parental migration on both the language and math scores of left-behind children as shown in the literature [Zhao et al., 2014, Meng and Yamauchi, 2015]. In particular, the direct effect is -0.524 ($p < 0.01$) for language scores and -0.453 ($p < 0.01$) for math scores, measured in standard deviations. The indirect effects through child's time allocation are insignificant and near zero, with size 0.003 ($p > 0.05$) for language scores and 0.002 ($p > 0.05$) for math scores in standard deviations, implying that intervention through this channel may not be effective for left-behind children in China. In contrast to many studies in other developing countries, I find that the indirect effects through investment in the child are significantly negative for both the language and math scores with larger magnitudes than the direct effects. Specifically, this indirect effect is -0.894 ($p < 0.001$) for language scores and -0.874 ($p < 0.001$) for math scores in standard deviations. The seemingly counterintuitive negativity is caused by underinvestment in the child despite the remittances sent by migrants. This could be driven by incompatible incentives of guardians in the absence of parents [Niimi et al., 2009, Chen, 2013]. The net effect of migration on child's schooling performance by adding up three mechanism-specific effects is -1.42 for language scores and -1.33 for math scores in standard deviations. Despite the large net effect, the decomposition into mechanism specific effects is clearly more informative for policy makers.

In addition, the gender subgroup analysis shows that girls' schooling performances are disproportionately affected by migration through underinvestment, revealing a shocking gender inequality in rural China. In fact, the indirect effect through investment for girls is more than 10

times larger than that for boys on both language and math scores. By further decomposing the investment into nutritional spending and educational spending, I find that the underinvestment is driven by both decreasing nutritional spending and decreasing educational spending, and the former has a substantially larger effect sizes. This suggests that a policy which compensates for the nutritional underinvestment tends to be effective in mitigating the negative effects of parental migration on left-behind children.

1.1.3. Related literature. As mentioned in the last subsection, existing works that I am aware of are almost exclusively focusing on the net effect of a single mechanism without detailed mediation analysis. In this subsection, I will review a selective set of relevant studies and highlight which net effect they estimate.

The first line of studies estimate the direct effect of migration due to the absence of parents [e.g. Graham and Jordan, 2011, He et al., 2012, Antman, 2013]. They find that left-behind children's schooling performances are negatively affected by higher levels of emotional disruption, stress, sadness [Ganepola, 2002, Mendoza, 2004], loneliness and abandonment [Battistella and Conaco, 1998, Maruja and Baggio, 2003], and lower self-esteem [Sun et al., 2015]. In addition, the absence of parents may disrupt the discipline of children [Arnold and Shah, 1984] and reduce their cognitive preparedness for school [Booth, 1995].

The second line of works focuses on the effect of migration on child's time allocation. It should be noted that it is still different from the effect of migration through child's time allocation since the former is only a component of the latter. Chen [2013] and Chang et al. [2011] examine the effect of children's labor substitution caused by parental migration, concluding that children of migrant households spend more time on housework and thus have less time for studying. Similar evidence has been found in Mexico [Antman, 2011, McKenzie and Rapoport, 2011], although there is no agreement on whether boys or girls suffer more from housework.

The third line of literature studies the effect of remittances sent back by migrants on children's schooling performance. Most studies find that the remittances relax the investment constraints in children, thereby improving children's living conditions, educational spending, and nutrition status [e.g. Afsar, 2003, Adams Jr and Page, 2005, Gupta et al., 2009]. The evidence has been found in Indonesia, Thailand [Bryant, 2005], Philippines [Bryant, 2005, Arguillas and Williams, 2010], Bangladesh [Afsar, 2003], Mexico [Hanson and Woodruff, 2003, Alcaraz et al., 2012], and

El Salvador [Edwards and Ureta, 2003]. Nevertheless, there are also contradictory findings that remittance sent home by migrants may not necessarily increase the investment in the child. Olwig [1999] shows that migrating parents usually leave their children with relatives such as grandparents or foster families. As a consequence, guardians may not have strong incentives to turn remittances into investment in the child due to the potential competition between elderlies and children, or between current and future consumption [Nguyen et al., 2006, Chen, 2013, Niimi et al., 2009, Knodel and Saengtienchai, 2002].

The rest of this paper is organized as follows. Section 1.2 describes the two-agent model that lays the foundation for empirical analyses and presents some descriptive results of the equilibrium. All mathematical derivations are relegated into Appendix A.1. Section 1.3 introduces the RUMiC survey data, as well as the definitions of the treatment, mediators, and outcome. In Section 1.4, I describe the identification strategy in detail, including the choices of instrumental variables. I present the main empirical findings in Section 1.5, with results on all samples and subgroups. Section 1.6 concludes and discusses future directions.

1.2. Theoretical Modeling Framework

1.2.1. A two-agent model. To understand the interaction between three mechanisms qualitatively, I consider a simple model with a household of one child and one parent with two time periods but without borrowing or savings. In the first period, I assume that the parent is at working age and the child is at school age, and the household consumption purely relies on the parent's income while the child's income from housework is negligible. In the second period, I assume that the child has grown up and fully entered the labor market while the parent has retired, so the household consumption solely relies on child's income. The child decides on how much time to spend on studying and the parent decides on migration, with both decisions made in the first period. The model is arguably over-simplified since it ignores different roles of the father and mother, behaviors of siblings and de facto custodians, irrationality of decision making from both sides, etc.. Nevertheless, it is complicated enough to reveal how the mechanisms of interest, namely the parental absence, child's study time, and investment in the child, interact and affect the child's schooling performance.

Let u_1^k and u_2^k denote the utility of the child [1] in period 1 and 2, respectively, and s be the share of time that the child spends studying. Therefore, $(1-s)$ denotes the share of time that the child spends on activities other than studying. Furthermore, I denote by h the human capital level of child in period 1, by h_0 the endowment of human capital, by $d \in [0,1]$ the proportion of days that parent migrate away and leave the child behind, by W_p the parent income from work as a function of d, by β_k the child's discount factor of the second-period utility, and by $f(\cdot)$ the production function of human capital, which takes the input of parent migration status, child study time, monetary investment in child, and the endowment in human capital. I assume the child's utility in the first period depends on s and child's consumption c_1^k, while the second-period utility depends solely on the household consumption c_2. Note that c_1^k does not necessarily increase as the parent income W_p increases, because $\gamma(d)$, the proportion of total income spent on the child, is not necessarily increased in d. Essentially, $\gamma(d)$ is the decision variable of the de facto custodian, who can decide how much of the remittances sent back by migrants will be spent on the child. If the custodian has full control of the spending when the parents migrate out, $\gamma(d)$ can be decreasing in d [Nguyen et al., 2006, Chen, 2013, Niimi et al., 2009, Knodel and Saengtienchai, 2002]. Due to the lack of data on guardians, I model $\gamma(d)$ as an exogenous factor to make the empirical analysis fully aligned with the theoretical model. Nonetheless, I briefly discuss how the guardian can be included into an extended three-agent model in Section 1.6 where $\gamma(d)$ is determined endogenously. For the second period, I denote by $g(h)$ the return to human capital h. Given all other variables, the child chooses an optimal study time s that maximizes the total utility from two periods, i.e.

$$
\begin{aligned}
\max_{s} \quad & u_1^k(s, c_1^k) + \beta_k u_2^k(c_2), \\
s.t. \quad & c_1^k = \gamma(d) W_p(d), \\
& c_2 = g(h), \\
& h = f(d, s, c_1^k, h_0).
\end{aligned} \tag{1.2.1}
$$

Similarly, for the parent, let u_1^p, u_2^p be the utility of the parent in period 1 and 2, respectively, and β_p be parent's discounting factor. I assume that parent consumption in the first period c_1^p is a fixed proportion γ_p of parent income with $0 < \gamma_p < 1$ because the parents can decide on their spending

[1] We use the letter k for "kid" instead of c for "child" to avoid similarity with consumption.

regardless of the migration status. Note that $\gamma(d) + \gamma_p \leq 1$ because when parents migrate away, the de facto custodian might not spend all the remittances on the child. The parent maximizes total utility by choosing the optimal migration status d^*, i.e.

$$
\begin{aligned}
(1.2.2) \qquad \max_{d} \quad & u_1^p(c_1^p) + \beta_p u_2^p(c_2), \\
s.t. \quad & c_1^p = \gamma_p W_p(d), \\
& c_2 = g(h), \\
& h = f(d, s, c_1^k, h_0).
\end{aligned}
$$

To derive the equilibrium, I make the following assumptions.

- $\frac{\partial u_j^i}{\partial c_j^i} > 0$ and $\frac{\partial^2 u_j^i}{\partial (c_j^i)^2} < 0$, where $i \in \{k, p\}$ and $j \in \{1, 2\}$, implying that the utility in each period increases while the marginal utility decreases in consumption in that period.
- $\frac{\partial u_1^k}{\partial s} < 0$ and $\frac{\partial^2 u_1^k}{\partial s^2} < 0$, implying a fatiguing effect of studying that is marginally increasing.
- $\frac{\partial f}{\partial s} \geq 0$, $\frac{\partial f}{\partial c_1^k} \geq 0$, $\frac{\partial^2 f}{\partial s^2} \leq 0$, $\frac{\partial^2 f}{\partial (c_1^k)^2} \leq 0$, implying that the study time and consumption weakly increase the production of human capital, but with decreasing marginal return.
- $\frac{\partial f}{\partial d} < 0$, implying that parental absence worsens child' human capital.
- $\frac{\partial g}{\partial h} \geq 0$ and $\frac{\partial^2 g}{\partial h^2} \leq 0$, implying that higher human capital of the child leads to higher income in the future, though with a decreasing marginal return.
- $\frac{\partial W_p}{\partial d} \geq 0$, implying the existence of monetary incentives to migrate.

For the rest of this section, I will provide qualitative analyses of the equilibrium solution from both the child and parent side. Formal mathematical derivations are relegated into Appendix A.1. In particular, I derive the closed-form solution for the equilibrium assuming specific functional forms of the utility functions in Appendix A.1.3, which yields the structural equation model that will be used for empirical analysis in Section 1.4.

1.2.2. Optimal decision of the child. For child utility maximization, there is a trade-off between current and future utility. Holding parent migration status d fixed, if study time s increases, the first-period utility decreases due to the fatiguing effect of studying, while the second-period utility increases because the child's human capital will increase due to increased study time, resulting in a higher future consumption c_2.

Intuitively, the child's optimal study time should be at the intersection of the marginal effect of study time on current utility ($\text{MU}_1^k = -\frac{\partial u_1^k}{\partial s}$) and its marginal effect on future utility ($\text{MU}_2^k = \frac{\partial u_2^k}{\partial s}$). The marginal effect of study time on current utility only depends on the level of study time. Holding the study time fixed, if the parent increases the proportion of days of migration, the marginal effect of study time on first-period utility is not affected, but the increased migration status will have a negative direct effect on human capital, and the indirect effect on human capital is undetermined since the change of investment in child is undetermined. Suppose the indirect effect of migration on human capital through investment in child is positive, the final effect on child's human capital still depends on the relative sizes of these direct and indirect effects. If the negative direct effect of migration dominates, then child's human capital worsens, leading to less income and consumption in the second period, so that the marginal utility from future consumption increases [2]. Graphically, the curve MU_1^k remains unchanged, while the curve MU_2^k shifts up, as shown in Figure 1.1. In general, the optimal child study time is increasing in migration status if the negative direct effect of migration dominates and is decreasing in migration status otherwise.

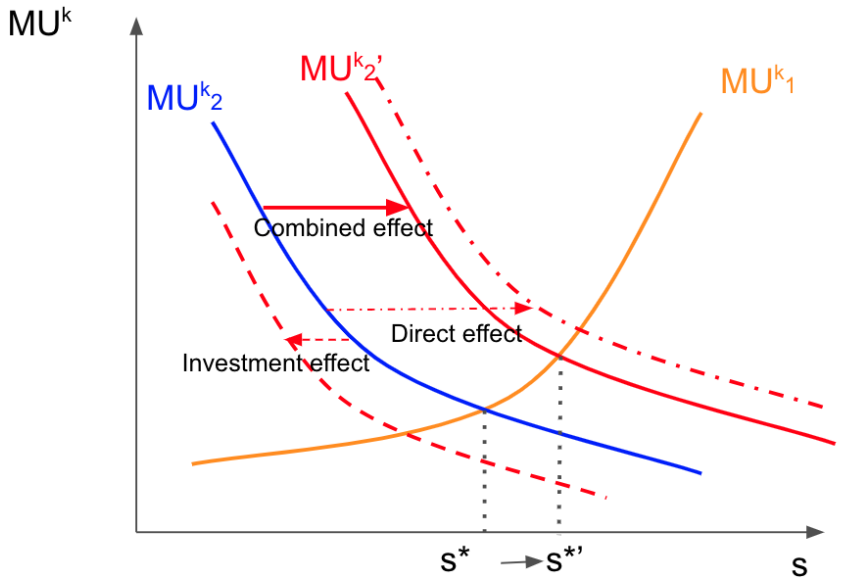


FIGURE 1.1. Trade-off between current and future utility of the child
The optimal study time s is determined by the intersection of two marginal utility curves. When the parental migration status d increases, the MU_1^k curve remains unchanged while the MU_2^k curve shifts up or down, depending on the relative sizes of the direct and indirect effect through investment in the child.

[2] See Appendix A.1.1 for mathematical details.

Due to the trade-off, under mild assumptions, we can find an interior solution $s^*(d)$. In Appendix A.1.3, I derive the closed-form solution of $s^*(d)$ assuming certain functional forms of the utility function. In general, assuming the existence of the interior optimal solution, I derive that [3]

$$\frac{\partial s^*}{\partial d} \propto -\left(\overbrace{\frac{\partial f}{\partial c_1^k} \frac{\partial W_k(d)}{\partial d}}^{\text{Investment effect}} + \overbrace{\frac{\partial f}{\partial d}}^{\text{Direct effect}} \right), \tag{1.2.3}$$

where $W_k(d) = \gamma(d) W_p(d)$, denoting the investment in the child, and $\propto$ denotes "proportional to", which hides a positive multiplicative factor. By the chain rule, $\frac{\partial W_k(d)}{\partial d} = \gamma(d) \frac{\partial W_p(d)}{\partial d} + W_p(d) \frac{\partial \gamma(d)}{\partial d}$. The decomposition of $\frac{\partial s^*}{\partial d}$ shows how the left-behind child's optimal study time changes when the parent migration status changes. $\frac{\partial f}{\partial c_1^k} \frac{\partial W_k(d)}{\partial d}$ represents the effect of migration on the child's human capital through investment in the child. $\frac{\partial f}{\partial d}$ measures the effect of parent absence on child's human capital. Since the sign of $\frac{\partial W_k(d)}{\partial d}$ is undetermined, equation (1.2.3) shows that the sign of $(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k} \frac{\partial W_k}{\partial d})$ is undetermined as well. If the indirect effect through investment in the child is positive, and the negative direct effect of being left-behind is greater in size, then $\frac{\partial s^*}{\partial d} \geq 0$, suggesting that the child will increase study time to compensate for worse human capital, and vice versa. This is consistent with the graphical illustration in Figure 1.1.

1.2.3. Optimal decision of the parent. For parent utility maximization, there is also a trade-off between current and future consumption. Holding the child's study time s fixed, if the parental migration status d increases, then the parent's first-period utility increases due to the increased consumption c_1^p, while the second-period utility decreases since the child's human capital will decrease due to the lack of parent accompaniment, resulting in a lower future consumption c_2.

Intuitively, the optimal parental migration status is at the intersection of the marginal effect of migration on current utility ($\text{MU}_1^p = \frac{\partial u_1^p}{\partial d}$) and its marginal effect on future utility ($\text{MU}_2^p = -\frac{\partial u_2^p}{\partial d}$). Parent utility in period 1 only depends on consumption levels. Holding the parental migration status constant, if the child increases the study time, it will not affect parent consumption or utility in period 1, but will decrease the marginal utility in period 2. This is because the increased study time will lead to higher human capital, which translates into higher income and higher consumption in period 2, so that the marginal utility from future consumption shifts down [4].

[3] See Appendix A.1.1 for the mathematical derivation.
[4] See Appendix A.1.2 for mathematical details.

Graphically, the curve for marginal effect of migration on current utility remains unchanged, and the curve for marginal effect of migration on future utility shifts down, as shown in Figure 1.2. That is, the optimal migration decision is increasing in child study time. Unlike the child optimal decision process, the MU_2^p curve always shifts down as the proportion of days of migration increases.

Due to the trade-off, under mild assumptions, we can find an interior solution $d^*(s)$. In Appendix A.1.3, I derive the closed-form solution of $d^*(s)$ assuming certain functional forms of the utility function. In general, assuming the existence of the interior optimal solution, I derive that [5]

$$\frac{\partial d^*}{\partial s} \propto -\left(\overbrace{\frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d}}^{\text{Investment effect}} + \overbrace{\frac{\partial f}{\partial d}}^{\text{Direct effect}} \right), \tag{1.2.4}$$

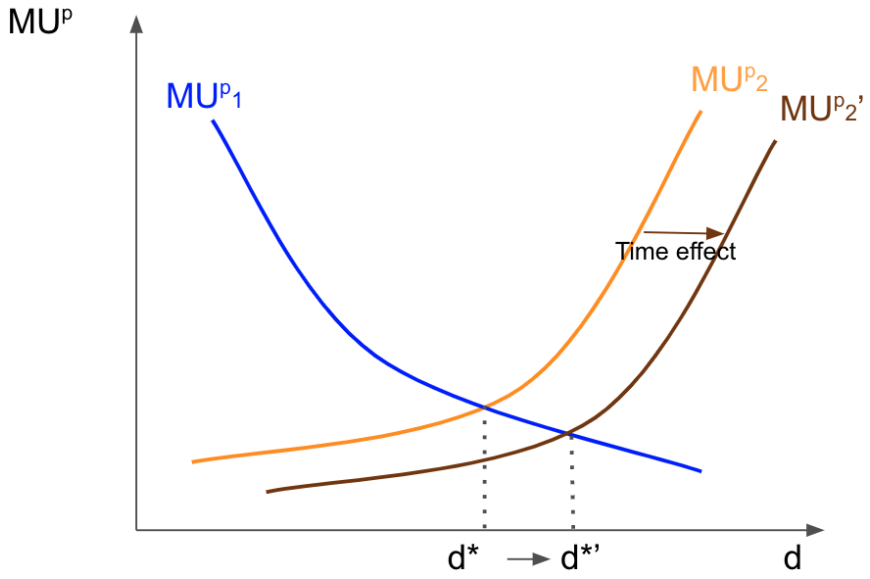


FIGURE 1.2. Trade-off between current and future utility of the parent

The optimal migration status d^* is determined by the intersection of two marginal utility curves. When the child's study time s increases, the MU_1^p curve remains unchanged while the MU_2^p curve shifts down.

The decomposition of $\frac{\partial d^*}{\partial s}$ shows how parent's optimal migration decision changes as child study time changes. The meaning of each part of $\frac{\partial d^*}{\partial s}$ is the same as in Equation (1.2.3). The marginal effect of parental migration on current utility is $\frac{\partial u_1^p}{\partial c_1^p}\frac{\partial c_1^p}{\partial d}$, and its marginal effect on future

[5] See Appendix A.1.2 for the mathematical derivation.

utility is $-\beta_p \frac{\partial u_2^p}{\partial h}(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d})$. To guarantee an interior solution, we need the future marginal effect to be nonnegative, that is, $\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d} \leq 0$. Therefore, $\frac{\partial d^*}{\partial s} \geq 0$. This confirms our intuition that $\frac{\partial d^*}{\partial s}$ has a definite sign, unlike $\frac{\partial s^*}{\partial d}$. This is consistent to the graphical illustration in Figure 1.2.

1.2.4. Equilibrium solution. In Section 1.2.2 and Section 1.2.3, I show that the child's optimal decision on study time is a function of the parental migration status d, and the parent's optimal decision on migration is a function of the child study time s. Solving both equations will lead to the equilibrium. Under specific functional forms to the utility function, human capital production function, and wage function, I show that there is only one unique equilibrium solution [6]. Figure 1.3 is an illustration of the equilibrium solution.

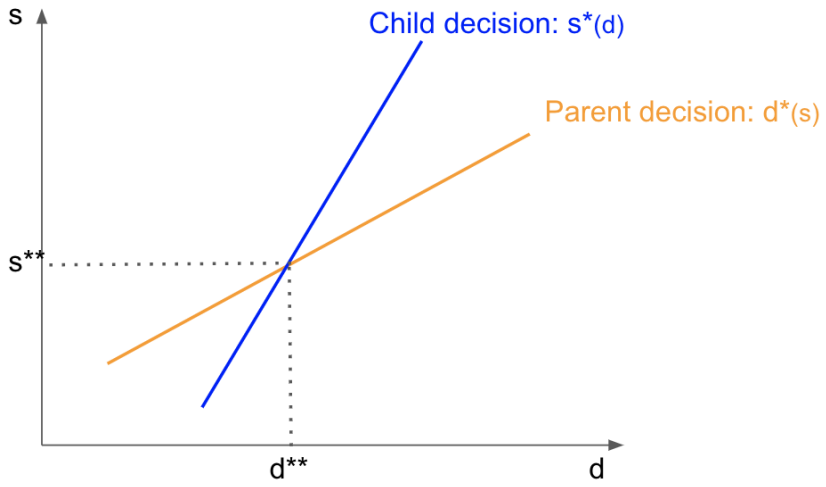


FIGURE 1.3. The equilibrium study time and migration status

It is determined by the intersection of the child's optimal study time $s^*(d)$ as a function of d and the parent's optimal migration status $d^*(s)$ as a function of s.

Given a migrant family, the theoretical model induces the following relationship among the observables – child's human capital h, investment in child W_k, the proportion of time that child's spent on studying s, the proportion of days of parent migration d, as well as other covariates X that account for the heterogeneity of families:

[6] See Appendix A.1.3 for the mathematical details.

$$
\begin{aligned}
h &= f(d, s, W_k; X) \\
W_k &= \gamma(d) W_p(d; X) \\
s &= s^*(d; X) \\
d &= d^*(s; X)
\end{aligned}
\tag{1.2.5}
$$

The main goal of this work is to study the effects of migration on child's human capital, i.e. $\frac{\partial h}{\partial d}$. By definition,

$$
\frac{\partial h}{\partial d} = \underbrace{\frac{\partial f}{\partial d}}_{\text{Direct effect}} + \underbrace{\frac{\partial f}{\partial s}\frac{\partial s}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k}{\partial d}}_{\text{Indirect effects}}.
$$

This yields the decomposition of the total effect into the direct effect and indirect effects. All parameters in the decomposition are necessary for answering my research question, and thus the first three equations in (1.2.5) must be estimated. By contrast, the effect of s on d, i.e. $\frac{\partial d}{\partial s}$, does not have to be known. As a result, we can simplify the last equation by computing $d^{**}(X)$ as the solution of $d = d^*(s^*(d, X), X)$ for each X, thereby facilitating the system. In Appendix A.1.3, I illustrate this step under a specific functional specification. To summarize, I focus on the following system:

$$
\begin{aligned}
h &= f(d, s, W_k; X) \\
W_k &= \gamma(d) W_p(d; X) \\
s &= s^*(d; X) \\
d &= d^{**}(X)
\end{aligned}
\tag{1.2.6}
$$

It is worth emphasizing that the way to simplify (1.2.5) hinges on the research question and whether the simplification works depends on whether an effective identification strategy exists. For (1.2.6), I find a promising identification strategy as detailed in Section 1.4. In principle, we can also study how child's human capital affects parent's migration decision, which can be answered by (1.2.5) in theory. However, it is arguably more challenging to find a convincing identification strategy.

1.3. Data

1.3.1. Data source. The data used in this paper is collected by the Rural-Urban Migration in China (RUMiC) Project, which is a longitudinal survey [Institute of Labor Economics (IZA) et al., 2014]. This project is a joint effort by the Australian University, University of Queensland, Beijing Normal University, and Institute for the Study of Labor (IZA). Starting in 2008, the project covers 9 provinces or province-level municipalities that are major sending or receiving areas of out-migration: Anhui, Chongqing, Guangdong, Hebei, Henan, Hubei, Jiangsu, Sichuan, and Zhejiang. The RUMiC survey includes 8,000 samples in rural household survey (RHS), 5,000 in urban household survey (UHS), and 5,000 in migrant household survey (MHS). Subjects in each category are randomly selected in each province. For detailed information on sampling design and tracking , see Gong et al. [2008], Meng et al. [2010], Kong [2010].

Although survey documents and data for both 2008 and 2009 are available, the 2008 data does not include important outcome variables such as children's exam scores or study time. For this reason, I mainly focus on the cross-sectional data in 2009 survey in this paper, and use the 2008 data for auxiliary purpose.

Since this paper focuses on rural migrants, data from RHS and MHS can both be used for analysis in principle. However, for the purpose of this paper, data from the RHS is preferable for several reasons. First, my paper is to compare left-behind children with children whose parent do not migrate. The RHS involves both groups, while the MHS only involves children of migrants. Second, RHS has substantially higher quality than MHS in terms of both the sample size and attrition rate (0.4% v.s. 58.4% attrition at the individual level, and 0.1% v.s. 63.6% at the household level, according to Akgüç et al. [2014]). Although the main analysis is based on 2009 data, the 2008 data is useful to impute for the missing values of demographic information, thereby increasing the effective sample size. As a result, RHS is more suitable for my analysis in terms of efficiency. Finally, this paper is focused on rural households. The RHS draws random samples from the annual household income and expenditure surveys carried out in rural villages, and tracks subjects having permanent living addresses. This makes the RHS a representative survey for my purpose.

The raw data has 6899 children in 4843 households. To make the analysis meaningful, I only include school-age children (6-15 years old) who have never married and with parents older than 16 years old. After filtering, 2666 children in 2112 households are left in the data. I further

exclude households for which the migration status is unreported, resulting in 1971 children in 1593 households. The parents in the data for my analysis come from 68 cities in 9 provinces, and their migration destinations spread over 137 cities in 29 provinces.

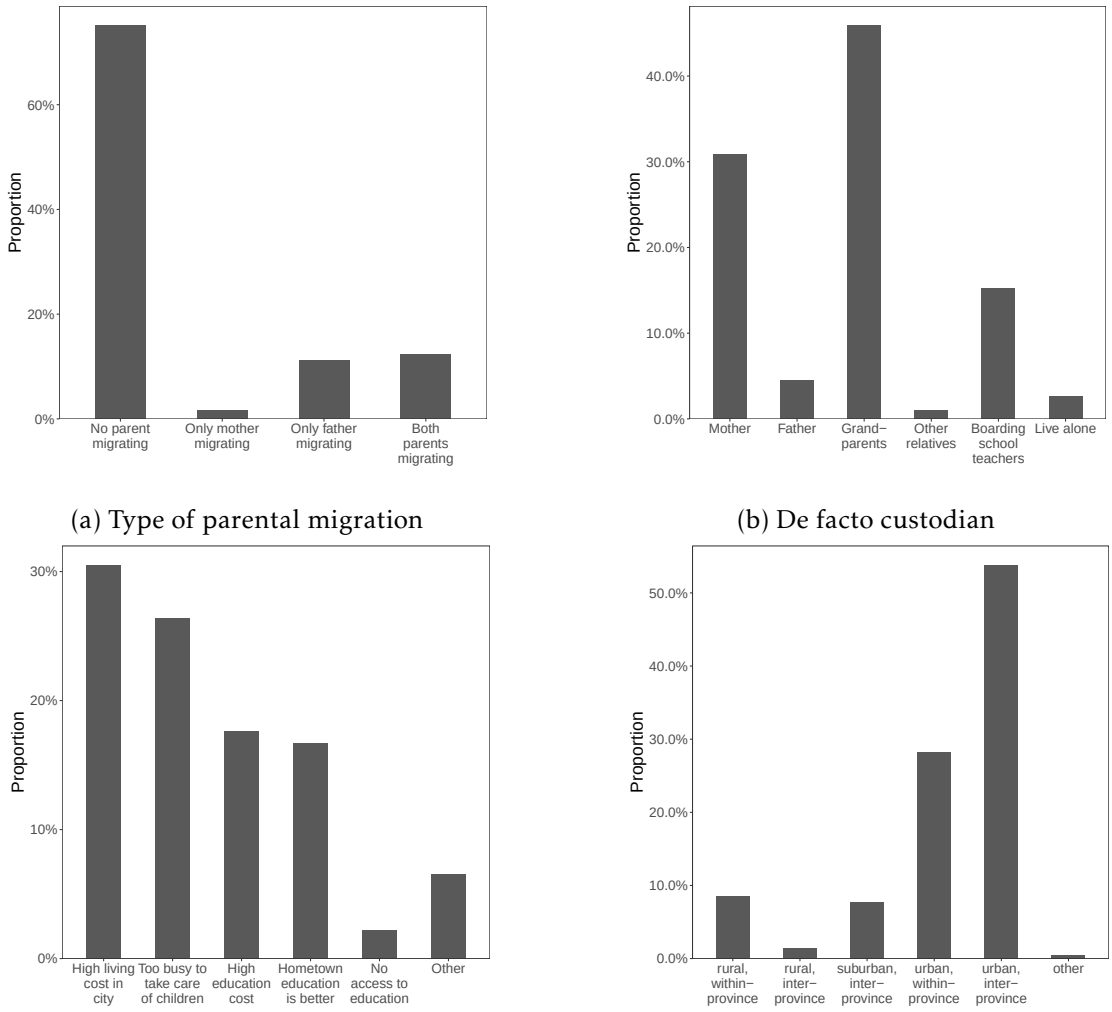
1.3.2. Descriptive statistics. In this section, I provide some basic information of the data. Figure 1.4a shows the fractions of children with both parents migrating, with father migrating only, with mother migrating only and with neither parents migrating. In this figure, a person is counted as a migrant if she/he migrates for more than 90 days in the past year; see Section 1.3.3 for a more detailed description and explanation. Adding up the proportions of the first three categories, left-behind children account for roughly 30% of children in rural China.

Figure 1.4b shows the fractions of different guardians whom the left-behind children live with. When both parents migrate out, grandparents are most common de facto custodians. This shows a potential source of incentive compatibility on monetary investment since grandparents may not want to allocate most remittances on children's education or nutrition, or they may not have a good sense on the appropriate amount of money spent on children. The second most common guardians are boarding school teachers, who may not have strong incentives to take care of any single child.

Figure 1.4c presents different reasons why parents do not bring children when migrating to work in cities. High living and education cost in cities appears to be the driving force to leaving children behind. This is partly because of the *hukou* restriction discussed in Section 1.1.1 that children with rural *hukou* cannot enjoy the social benefits such as education and housing. The lack of access to the social welfare system increases their living and education cost if they migrate with their parents. Another important motivation to leave children behind is that parents are too busy to take care of their children if they were brought along. This motivation is particularly strong when other family members who can play the role of caregivers, such as grandparents, are unable to migrate together.

Figure 1.4d shows the types of migration destinations. We can see that a majority of migrants move from rural to urban areas. Among rural-to-urban migrants, around two thirds of them move to an urban city in a different province.

1.3.3. Treatment variable: migration status. According to Meng and Yamauchi [2015], a good indicator for parental migration is based on very recent migration experience. Thus, I will define



(a) Type of parental migration

(b) De facto custodian

(c) Reason for leaving children behind

(d) Migration destination (work location)

FIGURE 1.4. Descriptive statistics of data

the migration status, the treatment variable in my problem, based on the duration of migration in the past year. In principle, I can define the migration status as the fraction of days migrating out in the past year as in the Section 1.2. However, this measure is not accurate since it is self-reported. In fact, I find that a large fraction of reported measures are multiples of 50, indicating potential lack of reliability of the measurement. Moreover, the estimated effects are harder to interpret and less transparent based on continuous treatment variables.

For these reasons, I will not use a continuous measure of migration but instead define a dummy variable D to represent the migration status where $D = 1$ if at least one parent of the child migrates out and leaves the child behind for at least 90 days in the past year. This measure is more accurate

since there are questions that explicitly use 90 days as the threshold [7]. In addition, since the control group in this paper is children in rural areas with non-migrant parents, rather than children who migrate with their migrant parents, I exclude the children of the latter kind from the analysis. Figure 1.5 provides the diagram definition of D. Note that the dummy variable D is essentially $I(d \geq 90)$ where d is the decision variable of the parent in Section 1.2.

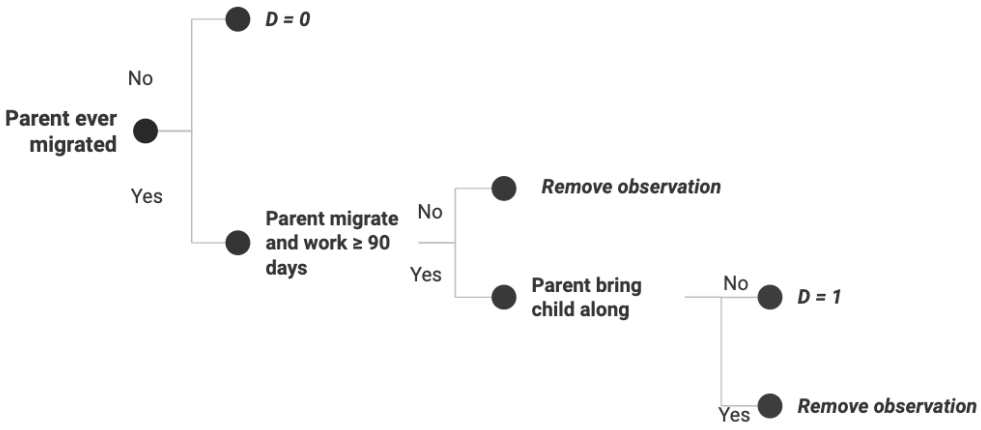


FIGURE 1.5. Definition of the treatment variable D

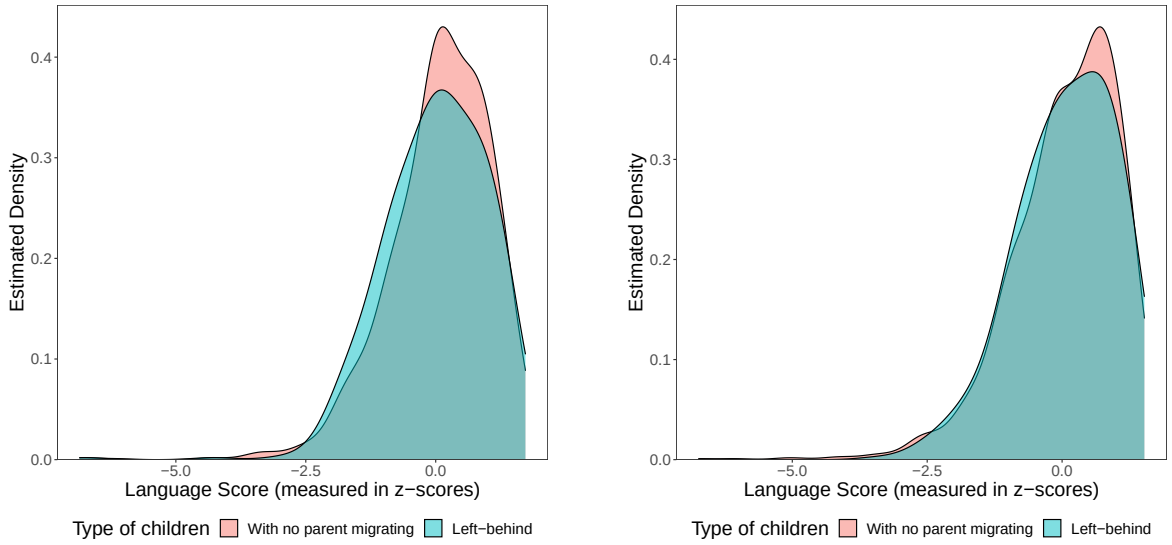
1.3.4. Mediation variables: study time and investment in the child. I use the weekly study time in hours reported by their guardians as a measure of the child's study time, denoted by T to be distinguished from the share of study time s used in Section 1.2. For investment in the child that corresponds to c_1^k in the theoretical model, I use the sum of the child's tuition at school, supplemental classes inside and outside of school, and food expenditure in 2008 reported by guardians. To stabilize the variance of the variable, I will use the logarithmic transformation of the total investment measured in Chinese Yuan as the mediation variable and denote it by W.

1.3.5. Dependent variables: standardized language and math scores. To measure the schooling performance, I choose the child exam scores, which are measures of the child human capital h in the model. In particular, I consider the final exam scores in language and math reported by parents or other guardians, who are informed of children's scores during parental meetings at school every semester. In addition, they would receive the hard copy of children's score reports from school at the end of every semester. Therefore, the reported scores are reliable. The test scores

[7] For example, question C07_4 states that how many days did you work outside your hometown in 2008? (If none, please fill in '0', if ≥ 90, skip to C07_6)

are also comparable across children in the sample since 7 out of 9 provinces use the same version of textbooks, while only a few villages in the remaining 2 provinces use another two versions of textbook. All of the three versions of textbooks and exams are designed closely following the curriculum standards designed by the Ministry of Education of China. Particularly, the materials are highly consistent for core subjects such as language and math. To make the scale of scores and estimated effects more interpretable, I convert test scores into z-scores by subtracting the average and dividing the standard error of the sample.

Figure 1.6 displays the distribution of exam scores. We could see that for left-behind children, the distribution of language scores is more left skewed, suggesting that these children perform worse in language exams on average. But the difference in math score distribution for left-behind children and other children is less pronounced. The marginal averages in two groups are reported in Table 1.1. We can see that left-behind children perform worse than children with non-migrant parents in language exams and slightly better in math exams, though neither of the differences is statistically significant at the 5% significance level. It is worth emphasizing that the marginal difference does reflect the effect of any kind because of the endogeneity and non-random missing values.



(a) Language score

(b) Math score

FIGURE 1.6. Distribution of testing scores

1.3.6. Control variables. As for other covariates, I include the personal characteristics of the child, such as the age, gender, height, weight, and birth weight. I also include parent-level characteristics such as the years of education of their parents. For those parents with missing values in these attributes in 2009, I impute them using the values reported in 2008 survey if available. For those parents who have records in both 2008 and 2009 surveys, if the measurements are inconsistent, I will choose the higher one. Other potentially important variables are excluded because they have a large fraction of missing values in the data for reasons that are hard to pinpoint.

Table 1.1 shows the summary statistics of a set of important control variables. We can see that left-behind children are significantly lighter and shorter than their counterparts. The difference in parent education levels in two groups is not statistically significant. In the empirical analysis, I control for covariates that are significantly different across treatment and control groups, and also include covariates that do not differ significantly to increase estimation efficiency.

TABLE 1.1. Summary statistics

Variable	Migrant Parents	Non-migrant Parents	Difference (P-value)
Dependent Variables			
Language score	0.01	0.08	0.14
Math score	0.08	0.06	0.73
Covariates: Child			
Male	0.53	0.55	0.52
Age	11.27	11.52	0.07
Height	135.92	142.44	$< 0.001^{***}$
Weight	38.96	41.46	$< 0.001^{***}$
Birthweight	32.50	32.51	0.94
Covariates: Parents			
Mother edu year	7.50	7.36	0.21
Father edu year	8.23	8.19	0.70
Note:	*p<0.05; **p<0.01; ***p<0.001		

1.4. Empirical Framework

1.4.1. Structural equation model (SEM) for mediation analysis. Under specific functional forms as in Appendix A.1.3, I show that the system of equations (1.2.6) have the following linear forms:

$$P_i = \gamma_0 + \gamma_T \cdot T_i + \gamma_W \cdot W_i + \gamma_D \cdot D_i + \xi_P \cdot X_i + \epsilon_{Pi}, \tag{1.4.1}$$

$$T_i = a_T + b_T \cdot D_i + \xi_T \cdot X_i + \epsilon_{Ti}, \tag{1.4.2}$$

$$W_i = a_W + b_W \cdot D_i + \xi_W \cdot X_i + \epsilon_{Wi}, \tag{1.4.3}$$

$$D_i = \mathbb{1}(a_D + \xi_D \cdot X_i + \epsilon_{Di} \geq 0), \tag{1.4.4}$$

where P_i denotes the schooling performance of child i, measured by normalized final exam scores in language and mathematics as described in Section 1.3.5, T_i denotes the weekly study time in hours, and W_i denotes the logarithmic transformation of monetary investment and D_i denotes the dummy variable of the parental migration, as defined in Section 1.3. Recalling Section 1.3.3 that D is essentially $I(d \geq 90$, the last equation has a different form compared to the other three. To account for individual heterogeneity, other covariates and error terms are included. X_i is the set of control variables, including characteristics of children and parents introduced in Section 1.3.6. The error terms ϵ_{Pi}, ϵ_{Ti}, ϵ_{Wi}, and ϵ_{Di} are random errors, and we assume that they are correlated due to unobserved confounders in order to account for the endogeneity. When ϵ_{Di} is normal, the last equation is equivalent to a Probit model.

With an identified model, if we define δ to be the total effect of migration on children's schooling outcomes, then the total effect can be decomposed into the following three part:

$$\delta = \underbrace{\gamma_D(\text{parental absence})}_{\text{Direct effect}} + \underbrace{\gamma_T b_T(\text{study time}) + \gamma_W b_W(\text{investment})}_{\text{Indirect effects}}, \tag{1.4.5}$$

where γ_D captures the direct effect of migration, $\gamma_T b_T$ captures the indirect effect of migration through the child's study time, and $\gamma_W b_W$ captures the indirect effect of migration through investment in the child. Figure 1.7 illustrates the decomposition.

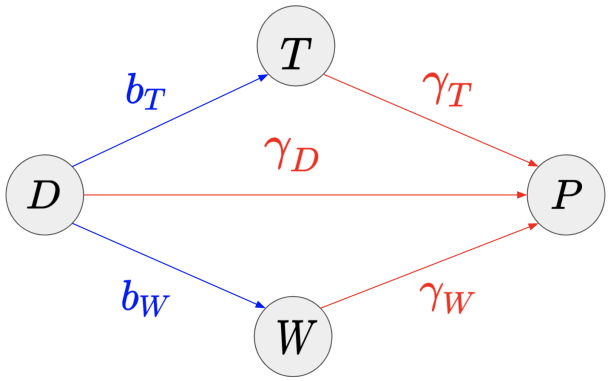


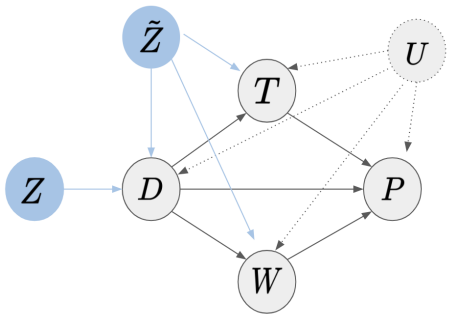
FIGURE 1.7. Decomposition of total effect into direct and indirect effects

1.4.2. Identification of coefficients via order condition. Since there are many unobserved factors that affect the parental migration decision, two mediators, and children's school performance simultaneously, the migration decision and mediators are all endogenous, rendering the standard mediation analysis, which assumes mutual independence between the errors $(\epsilon_{Pi}, \epsilon_{Ti}, \epsilon_{Wi}, \epsilon_{Di})$, implausible due to the omitted variable bias. For instance, variables such as child's self-control ability and parent's attitude toward child's education maybe correlated with both the parent migration decision and child's school performance, so the omission of such variables might lead to considerable bias.

To remove the confounding bias, I resort to an instrumental variable (IV) approach. Unlike the usual IV regression with a single endogenous variable and without mediators, the identification is more complicated for structural equation models with multiple endogenous variables. A necessary condition for identification is the order condition [e.g. Wooldridge, 2010], that is, for each equation in the system, the number of excluded exogenous variables, which includes both instrumental variables and other control covariates, should be larger than or equal to the number of included endogenous variables minus one. Although the order condition is only necessary but not sufficient, it is a simple and transparent condition to decide the necessary structure for identification. In the next subsection I will justify the rank condition, which is sufficient and necessary for identification.

Suppose we are able to find two sets of instrumental variables: Z that can only affect P through D, and $\tilde{Z}$ that can affect P through D or T or W. Both Z and $\tilde{Z}$ may involve multiple variables and we denote by $|Z|$ and $|\tilde{Z}|$ their sizes. Unlike Z which is required to be exogenous to both the outcome and mediators, $\tilde{Z}$ only needs to satisfy a weaker exclusion restriction. For instance, $\tilde{Z}$ is allowed to have direct effects on the mediators.

The permissible causal paths are illustrated in Figure 1.8a. The path diagram is a schematic representation of a structural equation model, where each node represents a variable and each directed edge from variable A to variable B represents the inclusion of A into the equation with B being the outcome. Note that the absence of an edge encodes an exclusion restriction while the presence of an edge imposes no constraint. Besides the outcome variable P, mediators T and W, and the treatment variable D, I also include the unmeasured confounders U, characterized by a dashed circle, as well as the instrumental variables Z and $\tilde{Z}$, characterized by blue circles. For simplicity I ignore other control covariates X that are allowed to link to all observed variables in



(a) Path diagram representation

$$\text{Eq. 1} : P \sim D + T + W$$
$$\text{Eq. 2} : T \sim D + \tilde{Z}$$
$$\text{Eq. 3} : W \sim D + \tilde{Z}$$
$$\text{Eq. 4} : D \sim Z + \tilde{Z}$$

(b) Algebraic representation

FIGURE 1.8. Representations of the SEM for mediation analysis with instrumental variables

the diagram. The four equations in Figure 1.8b are the algebraic representation of Figure 1.8a. More precisely, Z should be mean-independent of all residuals $(\epsilon_P, \epsilon_T, \epsilon_W, \epsilon_D)$ while $\tilde{Z}$ only needs to be mean-independent of ϵ_P.

The order condition can be read off from Figure 1.8b. For instance, for the equation of P, the number of endogenous variables is 4, given by P, T, W, D, and the number of excluded exogenous variables is $|Z| + |\tilde{Z}|$ since none of those instrumental variables are included in this equation. By contrast, for the equation of T, the number of endogenous variables is 2, given by T and D, while the number of excluded exogenous variables is only $|Z|$ since $\tilde{Z}$ is included. Table 1.2 summarizes these two quantities for each equation.

TABLE 1.2. Order condition for identification of the SEM in Figure 1.8b

Outcome variable	# Excluded Exogenous	# Included Endogenous - 1
P	$\lvert Z\rvert + \lvert\tilde{Z}\rvert$	3
T	$\lvert Z\rvert$	1
W	$\lvert Z\rvert$	1
D	0	0

Therefore, to identify all coefficients in the SEM, the order condition requires

$$|Z| + |\tilde{Z}| \geq 3, \quad |Z| \geq 1. \tag{1.4.6}$$

As a consequence, it suffices to find at least one IV in Z and at least two IVs in $\tilde{Z}$.

1.4.3. Identification of coefficients via rank condition. Rank condition is sufficient and necessary for identification. For the usual IV regression with one endogenous treatment variable, the order condition is equivalent to the exclusion restriction while the rank condition is equivalent to the exclusion restriction plus the relevance condition that requires the IV to be correlated with the treatment.

For general SEMs, the rank condition is much more complicated. Nonetheless, the SEM in Figure 1.8b is triangular, enabling a more transparent check of the rank condition. Recall that only five parameters need to be identified: $\gamma_T, \gamma_W, \gamma_D$ from the equation of P, b_T from the equation of T, and b_W from the equation of W. To identify b_T, it is sufficient to focus on the equations of T and D, which form an usual IV regression model with Z being the instrumental variable and $\tilde{Z}$ being a control covariate. The exclusion restriction is guaranteed by the order condition (1.4.6). As mentioned above, the rank condition still requires the relevance. In this case, it can be simply verified by the commonly-used F-test that tests if all coefficients of Z in the equation of D are zero. The same strategy can be applied to identify b_W. An asymptotically equivalent test in this case is the Anderson's canonical correlation test, which is designed for multiple endogenous variables and will be introduced in the next paragraph.

To identify γ_T, γ_W and γ_D, we can view T, W, D as three endogenous treatments for P. Then we can regard $Z, \tilde{Z}$ as instrumental variables for T, W, D because $Z, \tilde{Z}$ only affects P through T, W, D by definition. The well-known rule-of-thumb requires no fewer IVs than endogenous treatments. This is essentially the order condition and is guaranteed by (1.4.6) in my SEM. However, the relevance condition for multiple endogenous treatments is more involved. In my SEM, it requires the $3 \times (|Z| + |\tilde{Z}|)$ coefficient matrix by regressing (T, W, D) on $(Z, \tilde{Z})$, as well as other control covariates, to be full-rank (i.e. with rank 3). The usual F-test that borrows the heuristics from the single treatment case is flawed since it tests the wrong null hypothesis that the rank is zero instead of the correct null hypothesis that the rank is below 3. A more rigorous test is using the Anderson's canonical correlation LM statistic [Anderson, 1951], which is based on the smallest singular value of the coefficient matrix and thus testing the correct null. It is referred to as an underidentification test in the literature and is the default test in STATA `ivreg2` command [Baum et al., 2007].

In a nutshell, the rank condition can be justified by the order condition plus three underidentification tests, all based on Anderson's canonical correlation LM test. The triangular structure enables a clean interpretation of the identification strategy – when the rank condition holds, b_T, b_W are identified by Z and $\gamma_T, \gamma_W, \gamma_D$ are identified jointly by Z and $\tilde{Z}$. The relevance part of the rank condition is empirically testable so in the next subsection we will focus on justifying the exclusion restriction, which is generally untestable.

1.4.4. Choice of instrumental variables. I start with the choice of the instrumental variables Z. By definition, Z should be an IV that affects child performance only through the parental migration status. Some popular candidates for Z in the literature are religious preference uncommon in urban locations, dummy variable indicating whether the householder's first occupation was as a farmer, distance from home village to provincial capital, and average migration rate in the village [Xiang et al., 2016, Meng and Yamauchi, 2015]. However, these are not appropriate in my case. First, the religious tendencies are generally low in China and uncommon religious preferences are even rarer. Second, the householder's first occupation as a farmer is inappropriate since the share of farmers is predominantly high in rural China, implying a low variation of the occupation indicator. Third, the distance from home village to provincial capital suffers from lack of relevance for migrants in rural China because most people migrate to other provinces and thus the distance is less of a concern in deciding migration. The exclusion restriction of this variable is also likely violated because the general education facilities in regions closer to provincial capital tend to be better, leading to better schooling outcomes. Last, the average migration rate would not only influence the migration decision of each household, but also influence tax revenues and educational investment in that region, thereby influencing the schooling outcomes of children and violating the exclusion restriction.

In this paper, I use the method of Bartik [1991] to construct shift-share instrumental variables. Bartik instruments are widely used in migration literature. They are correlated with migration decision, but are arguably exogenous in the equations of schooling performance, study time, and investment in the child, which makes them appealing valid IVs [Goldsmith-Pinkham et al., 2020]. The Bartik-style instrument combines migrants' destination-industry information with changes in employment rate at destination by industry. The migration information is generated based on migrant's origin city, destination city, and the industry they work for using data from China

1% National Population Sample Survey 2005. The employment information is extracted from Urban Statistical Yearbook of China. The change in employment rate is generated using 2007 and 2008 employment data of each industry in all cities in China. These years are chosen such that there is sufficient time for migration flow to change as employment changes, but not too early so that the correlation between migration and employment would fade away. Specifically, the Bartik instrument is generated as below:

$$Z_{\text{ori},2008} = \frac{\sum_{\text{des}} \sum_{\text{ind}} (\text{Mig}_{\text{ori,des,ind},2005} \cdot \triangle\text{Employment}_{\text{des,ind,2007-2008}})}{\sum_{\text{des}} \sum_{\text{ind}} \text{Mig}_{\text{ori,des,ind},2005}},$$

where ori denotes the origin city that the migrant is from, des denotes the destination city that the migrant moves to, ind denotes the industry that the migrant works in, $\text{Mig}_{ori,des,ind,2005}$ denotes the total number of migrant workers from city ori to city des that work in industry ind in 2005, and $\triangle\text{Employment}_{\text{des,ind,2007-2008}}$ denotes the growth rate of employment in industry ind in destination des from 2007 to 2008. Considering the difference between inter-city ($des \neq ori$) and within-city ($des = ori$) migration, I generate Bartik instruments for each, where the inter-city Bartik instrument takes the sum of d over all cities other than o while the within-city Bartik instrument solely takes $d = o$ into account. Note that having two variables in Z does not violate the order condition (1.4.6). Later in the empirical analysis, I will conduct Sargan test [Wooldridge, 2010] to test for overidentification.

As for $\tilde{Z}$, I choose three variables – rainfall shocks in previous years, the inherited gold and silver accessories, and the birth order among siblings. Abnormally high or low precipitation is detrimental to agricultural production, which is the main economic activity in rural China. Such weather shocks push rural residents to migrate away in search of more stable job opportunities, and also affect the time allocation and wealth of rural households. Apart from through these channels, weather itself can hardly affect the children's school performance directly. More specifically, the rainfall shock is measured based on the 19-year (between 1991 and 2009) annual precipitation data at the city level obtained from China's National Meteorological Information Center. For each city, data from 1991 to 2003 is used to find the mean and standard deviation of precipitation. For each year in 2004 to 2009, if the annual precipitation is over 1.5 standard deviations below or above the city's historical mean, I define the city to have experienced a rainfall shock. Then I calculate the total number of rainfall shocks experienced in each city during 2004 to 2009. A

similar instrumental variable is constructed by Meng and Yamauchi [2015], although they treat high and low levels of precipitation asymmetrically while I treat them equally. The second variable, namely the inherited gold and silver accessories, measures the wealth level of older generations. It is thus unlikely confounded with the treatment, mediators, and outcome in the system. On the other hand, the variable may affect the current wealth level, through which affect the outcome, but it does not hurt identification since $\tilde{Z}$ is allowed to have such effects as shown in Figure 1.8a. The relaxed exclusion restriction, compared to that for Z, renders it a valid IV in this case. The third variable, namely the birth order among siblings, is random within a household and thus exogenous. Although it may affect the child's schooling performance through the investment due to the potentially unequal allocation between elder and younger children, this doesn't violate the requirement of $\tilde{Z}$ as specified in Figure 1.8a. Other than through migration, study time, and investment, it's unlikely that birth order among siblings can affect child's schooling performance. This justifies the relaxed exclusion restriction and renders it a valid IV. Finally, the order condition (1.4.6) only requires $|\tilde{Z}| \geq 1$ given that Z contains two variables, I choose more than one IV for $\tilde{Z}$. As with Z, I will conduct Sargan overidentification test in empirical analyses for $\tilde{Z}$.

1.4.5. Nonrandom missing patterns. The above subsections address one common source of endogeneity in variables of interest. In this subsection, I will focus on another source of endogeneity that originates from nonrandom missing patterns in the study time and investment, especially the former. It is unlikely that these two variables are missing at random conditionally on observed covariates because less caring parents may not know the child's education well and thus fail to report the information.

Previous studies simply remove observations with missing values in empirical analysis without accounting for nonrandom missing patterns. However, simply removing the observations with missing values in these variables may yield underestimation or overestimation of the negative effect of migration. Instead, I assume that the guardians reports the study time or investment in child only when a certain utility is above zero. When the utility is a linear function of the covariates with normal errors, this is precisely a Heckman model. In principle, the Heckman model can be added into the structural equation model directly and estimated using methods maximum likelihood. However, the non-standard form will significantly complicate the structure, making the estimation overly challenging. Therefore I apply a two-step procedure in which I first estimate

the Heckman model for the study time and investment separately to impute the missing values, and then estimate the SEM using the imputed data.

1.5. Empirical Results

1.5.1. Main results on all samples. Since my goal is to investigate the effect of migration on both language and math scores, I consider the extension of (1.4.1) - (1.4.4) that includes two scores simultaneously:

$$
\begin{aligned}
P_\ell &= \gamma_{0,\ell} + \gamma_{T,\ell} \cdot T + \gamma_{W,\ell} \cdot W + \gamma_{D,\ell} \cdot D + \xi_{P,\ell} \cdot X + \epsilon_{P,\ell}, \\
P_m &= \gamma_{0,m} + \gamma_{T,m} \cdot T + \gamma_{W,m} \cdot W + \gamma_{D,m} \cdot D + \xi_{P,m} \cdot X + \epsilon_{P,m}, \\
T &= a_T + b_T \cdot D + \xi_T \cdot X + \epsilon_T, \\
W &= a_W + b_W \cdot D + \xi_W \cdot X + \epsilon_W, \\
D &= \mathbb{1}(a_D + \xi_D \cdot X + \epsilon_D \geq 0),
\end{aligned}
$$

where ℓ and m in the subscripts are short for "language" and "math", and the subscript i for each unit is suppressed for notational convenience. Recall that P_ℓ and P_m are standardized z-scores, T measures the study time in hours, W is the logarithmic transformation of spending measured in Chinese Yuan, D is the binary migration decision. Therefore, $\gamma_{D,\ell}, \gamma_{D,m}, b_T, b_W$ measure the average difference of P_ℓ, P_m, T, W between left-behind children and non-left-behind children, $\gamma_{T,\ell}, \gamma_{T,m}$ measure the improvement of P_ℓ, P_m when the study time increases by one hour, and $\gamma_{W,\ell}, \gamma_{W,m}$ measure the improvement of P_ℓ, P_m multiplied by 100 when the investment increases by 1%.

Under this specification, the direct and indirect effects are $(\gamma_{D,\ell}, \gamma_{T,\ell} b_T, \gamma_{W,\ell} b_W)$ and $(\gamma_{D,m}, \gamma_{T,m} b_T, \gamma_{W,m} b_W)$ for language and math scores, respectively. Here I allow the error terms $\epsilon_{P,\ell}$ and $\epsilon_{P,m}$ to be correlated. This expanded SEM can capture the high correlation between the language and math scores. It is easy to see that the order condition remains the same as (1.4.6), and the rank condition can be tested exactly in the same way as in Section 1.4.3. Later on I will suppress the subscripts ℓ and m when no confusion can arise.

The top panel of Table 1.3 shows the direct and indirect effects of migration using all samples. The two columns report the effect on normalized language scores and math scores estimated

with the strategy introduced in Section 1.4. In the bottom panel, I report the underidentification, overidentification, and endogeneity tests separately for study time, investment, language score, and math score. Note that the underidentification results are the same for study time and investment because they are modelled simultaneously, and the same reasoning applies for the underidentification results for language and math scores. All underidentification tests reject the null at the 0.1% significance level, suggesting strong evidence that the rank condition holds. The test results for overidentification suggest that no evidence has been found against the null hypothesis that over-identifying restrictions are valid. As for the endogeneity test, although there is no evidence against the exogeneity of migration decision in the equation for study time, there is strong evidence against it in the equation for investment. In addition, there is strong evidence against the exogeneity of migration, study time, and investment jointly in the equations for exam scores. This marks the importance of accounting for the endogeneity.

For indirect effects, we report p-values from the joint significance test. Although this test cannot be inverted to a confidence interval as opposed to Sobel test, it is valid for testing the null effect. In addition, it is found to be powerful compared to alternative methods [e.g. Fritz and MacKinnon, 2007, Hayes and Scharkow, 2013].

Table 1.3. Effect of parental migration on child schooling outcomes (all sample)

			(1) Language	(2) Math
Direct Effect				
Parental Accompany			-0.524**	-0.453**
			(0.002)	(0.006)
Indirect Effect				
Study time			0.003	0.002
			(0.096)	(0.406)
Investment in children			-0.894***	-0.874***
			(0.000)	(0.001)
Sepecification Tests	*(1) Study time*	*(2) Investment*	*(3) Language*	*(4) Math*
Underidentification test				
(Anderson canon. corr. LM statistic)	29.868***	29.868***	29.859***	29.859***
	(0.000)	(0.000)	(0.000)	(0.000)
Overidentification test (Sargan statistic)	2.752	0.607	0.360	0.306
	(0.097)	(0.436)	(0.835)	(0.858)
Endogeneity test	2.366	10.199***	26.587***	30.409***
	(0.124)	(0.001)	(0.000)	(0.000)
Obs.		1971		

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Recall that the net effect of migration is the summation of the following three effects: direct effect of migration γ_D, which is the effect of parent absence; indirect effect of migration through

the child's study time $\gamma_T b_T$; and the indirect effect through investment in the child $\gamma_W b_W$. Adding up the direct and indirect effects, left-behind children perform 1.42 standard deviations worse than children with non-migrating parents in language, and 1.33 standard deviations worse in math.

From the first column of Table 1.3, the direct effect of migration on language score is -0.52 standard deviations and significant at the 1% level. From the second column of Table 1.3, the direct effect on math score is -0.45 standard deviations and significant at the 1% level.

As for the indirect effects shown in Table 1.3, the effect of migration on exam scores through the child's study time is almost zero and not significant. This might be because migration has little effect on child's study time, or because the child's study time doesn't change test scores by a lot. To further investigate the cause, I decompose this indirect effect into γ_T and b_T in the top panel of Table 1.4. The first column shows the estimate of b_T, namely the effect of migration on child's study time. The second and third columns show the estimate of γ_T, namely the effect of study time on child's language and math scores respectively. The coefficient b_T shows that left-behind children spend less time studying ($b_T = -0.29$ with $p = 0.083$) than children with non-migrant parents on average, but the difference is not statistically significant. The effect of study time on test scores is neither large nor significant ($|\gamma_T| \leq 0.012$ with $p \geq 0.096$ for both language and math scores).

Table 1.3 also show that parental migration has significant (at 0.1% significance level) and large negative indirect effects on both scores through investment, which are almost doubled in size compared to the corresponding direct effects. This finding is perhaps surprising since most of existing works conclude that remittances have positive effects on the child's educational outcomes. To dig into it, I show the decomposition of this indirect effect into γ_W and b_W in the bottom panel of Table 1.4. The first column shows the estimate of b_W, namely the effect of migration on investment in the child, and the second and third columns show the estimates of γ_W, namely the effects of investment on child's language and math scores respectively. It turns out that the investment in left-behind children is significantly lower ($b_W = -2.38$ with $p < 0.001$) than that in children who are not left behind, despite that investment is beneficial to child's school performance ($\gamma_W \geq 0.37$ with $p \leq 0.001$ for both language and math scores). This implies that the negative indirect effect through investment is driven by underinvestment.

The results in Table 1.3 are based on a careful account of both the endogeneity of migration/-mediators and non-random missing mediators as discussed in Section 1.4. In Appendix A.2, I

TABLE 1.4. Decomposition of indirect effects of migration (all sample)

	(1) Mediator	(2) Language	(3) Math
Through Study Time			
Migration (b_T)	-0.294		
	(0.083)		
Study time (γ_T)		-0.012	-0.006
		(0.096)	(0.406)
Through Investment			
Migration (b_W)	-2.375***		
	(0.000)		
Investment (γ_W)		0.376***	0.368***
		(0.000)	(0.001)

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

present the results without accounting for one of them or both of them as sanity checks, showing that failure of addressing these issues tends to underestimate the direct effect and indirect effect through investment drastically, despite that all analyses are consistent in the signs of the effects.

1.5.2. Exploring heterogeneous treatment effects. In this subsection, I investigate the heterogeneous treatment effects in different subgroups. In particular, I am interested in subgroups stratified by gender and child birth order. The different attitude of guardians toward boys and girls, as well as the role that the eldest and younger children play in multiple-children families will probably lead to heterogeneous treatment effects when parents migrate away. For each subgroup, I estimate the same SEM and present the results in Table 1.5 and Table 1.7.

Table 1.5 shows the effect of parental migration on left-behind boys and girls separately. Note that for migration in the equations with study time and investment being the outcome and Z being the IVs, the underidentification test statistic is marginally significant ($p = 0.052$) for girls, suggesting a reasonably strong evidence for the rank condition. The other underidentification tests all show strong evidence for the rank condition. The overidentification tests and endogeneity tests yield qualitatively the same results as those for all samples.

In terms of direct effects, left-behind boys are more negatively affected in language scores, which is consistent with the finding for all samples in Table 1.3. By contrast, left-behind girls are almost equally affected in language scores and math scores with significant and large negative effect sizes.

In terms of indirect effects, neither left-behind boys nor girls are largely or significantly affected through study time. The decomposition of this indirect effect is presented in the top panel of Table 1.6. Left-behind girls experience large and significant reductions in study time, while left-behind

TABLE 1.5. Effect of parental migration on child schooling outcomes (subgroup by gender)

	Girl		Boy	
	(1) Language	(2) Math	(3) Language	(4) Math
Direct Effect				
Parental Accompany	-0.413*	-0.424*	-0.351**	-0.207
	(0.015)	(0.030)	(0.008)	(0.074)
Indirect Effect				
Study time	0.002	0.003	-0.006	0.000
	(0.602)	(0.474)	(0.340)	(0.974)
Investment in children	-1.393*	-1.621*	-0.124**	-0.115**
	(0.010)	(0.010)	(0.008)	(0.003)
Sepecification Tests	*(1) Study time*	*(2) Investment*	*(3) Language*	*(4) Math*
Girl				
Underidentification test				
(Anderson canon. corr. LM statistic)	5.904	5.904	13.599**	13.599**
	(0.052)	(0.052)	(0.004)	(0.004)
Overidentification test (Sargan statistic)	0.903	1.938	1.486	2.573
	(0.342)	(0.164)	(0.476)	(0.276)
Endogeneity test	0.800	9.201**	9.436*	15.019**
	(0.371)	(0.002)	(0.024)	(0.002)
Obs.		887		
Boy				
Underidentification test				
(Anderson canon. corr. LM statistic)	28.100***	28.100***	20.772***	20.772***
	(0.000)	(0.000)	(0.000)	(0.000)
Overidentification test (Sargan statistic)	2.622	0.001	1.794	2.798
	(0.105)	(0.981)	(0.408)	(0.247)
Endogeneity test	1.156	1.696	18.078***	12.723**
	(0.282)	(0.193)	(0.000)	(0.005)
Obs.		1084		

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

boys tend to spend more time studying, though the effect is insignificant and cannot be viewed as positive on this data. As with the analysis for all samples, study time has no significant effect on either test scores.

Unlike the effect through study time, left-behind boys and girls are all significantly affected through investment. Notably, the girls are suffering ten times more than boys through this mechanism, with huge effects that are more than 1 standard deviation in sizes. The decomposition of this indirect effect is presented in the bottom panel of Table 1.6. Compared with left-behind boys, left-behind girls are suffering from a much severer reduction in investment, and their scores are also more vulnerable to underinvestment. This finding reveals a shocking gender inequality in rural China, at least among the left-behind children.

These results suggest that in order to increase left-behind girls' school performances, policies targeted at increasing the monetary investment in them should in general be much more effective than policies targeting at increasing their study time.

Table 1.6. Decomposition of indirect effects of migration (subgroup by gender)

	Girl			Boy		
	(1) Mediator	(2) Language	(3) Math	(4) Mediator	(5) Language	(6) Math
Through Study Time						
Migration (b_T)	-0.524*			0.400		
	(0.022)			(0.340)		
Study time (γ_T)		-0.004	-0.006		-0.016	0.000
		(0.602)	(0.474)		(0.096)	(0.974)
Through Investment						
Migration (b_W)	-3.514*			-0.967***		
	(0.010)			(0.001)		
Investment (γ_W)		0.397**	0.461**		0.128**	0.119**
		(0.008)	(0.008)		(0.008)	(0.003)

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Similar to the gender subgroup analysis, I also investigate the heterogeneous treatment effects with respect to the birth order. In particular, I conduct separate mediation analyses for the eldest and younger children in households with more than one child. Since the birth order is used as an instrumental variable, I replace this IV with the number of regular residents in the household. The number of regular residents is relatively exogenous. On the other hand, it unlikely affects child's schooling performance through mechanisms other than migration, household wealth and child's study time. Thus it also meets the requirements for $\tilde{Z}$, as shown in Figure 1.8a.

For households with multiple children, usually the eldest child takes care of the younger children and provide their younger siblings with emotional support. As a consequence, I expect that the role of parent partially shifts to the eldest child when parents migrate away, so the subsequent children may suffer less than the eldest child due to parent absence. This is confirmed by the results in Table 1.7. Due to the absence of parents, the eldest children have approximately 0.7 standard deviations lower scores on average in language and math exams than their non-migrant counterparts, and these effects are significant at the 1% level approximately. They are much larger than the direct effects of migration on the subsequent children.

The indirect effects through study time are small in sizes for both subgroups. Even though the effect on subsequent children's math scores is significant, I will not over-interpret it due to the tiny effect size. The top panel of Table 1.8 shows the decomposition of this effect into b_T and γ_T. We can observe that the eldest child tend to spend more time studying, though the effect is insignificant and cannot be interpreted as positive, while the subsequent children experience significant and

TABLE 1.7. Effect of parental migration on child schooling outcomes (subgroup by birth order)

	First child		Subsequent children	
	(1) Language	(2) Math	(3) Language	(4) Math
Direct Effect				
Parental Accompany	-0.696*	-0.696*	-0.289	-0.052
	(0.012)	(0.013)	(0.056)	(0.696)
Indirect Effect				
Study time	-0.014	-0.013	-0.001	-0.011*
	(0.399)	(0.399)	(0.779)	(0.019)
Investment in children	-0.295	-0.359*	-0.470*	-0.205
	(0.092)	(0.047)	(0.043)	(0.310)
Sepecification Tests	*(1) Study time*	*(2) Investment*	*(3) Language*	*(4) Math*
First child				
Underidentification test				
(Anderson canon. corr. LM statistic)	18.535***	18.535***	7.175*	7.175*
	(0.000)	(0.000)	(0.028)	(0.028)
Overidentification test (Sargan statistic)	0.050	2.671	0.037	0.612
	(0.823)	(0.102)	(0.848)	(0.434)
Endogeneity test	2.011	0.959	12.934**	16.820***
	(0.156)	(0.327)	(0.005)	(0.001)
Obs.		891		
Subsequent children				
Underidentification test				
(Anderson canon. corr. LM statistic)	7.183*	7.183*	7.598*	7.598*
	(0.028)	(0.028)	(0.022)	(0.022)
Overidentification test (Sargan statistic)	0.619	1.179	0.077	0.247
	(0.431)	(0.278)	(0.781)	(0.619)
Endogeneity test	6.058*	9.804**	6.815	4.326
	(0.014)	(0.002)	(0.078)	(0.228)
Obs.		860		

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 1.8. Decomposition of indirect effects of migration (subgroup by birth order)

	First child			Subsequent children		
	(1) Mediator	(2) Language	(3) Math	(4) Mediator	(5) Language	(6) Math
Through Study Time						
Migration (b_T)	0.528			-0.641**		
	(0.399)			(0.003)		
Study time (γ_T)		-0.027	-0.025		0.002	0.017*
		(0.110)	(0.167)		(0.779)	(0.019)
Through Investment						
Migration (b_W)	-1.226*			-2.251*		
	(0.044)			(0.012)		
Investment (γ_W)		0.241	0.293*		0.209*	0.091
		(0.092)	(0.047)		(0.043)	(0.310)

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

large reductions in study time. In line with the results for all samples, the study time has tiny effect on test scores.

As opposed to the effects through study time, the effects through investment are large in sizes for both subgroups. In particular, the effect on math scores is significant for the eldest child and the effect on language scores is significant for subsequent children. The decomposition of the indirect

effect through investment is presented in the bottom panel of Table 1.8. Compared with the eldest child, subsequent children suffer from a more severe reduction in investment. From the estimates of γ_W's, we see that the eldest child's math performance is more vulnerable to underinvestment, while the subsequent child's language performance is more vulnerable to underinvestment. This might be partially explained by the different cognitive development stages of these children.

1.5.3. Extended analysis. Previous results reveal notable effects of migration through investment in the child. To decouple the contributions of different types of investment, I further decompose it into nutrition spending and course-related spending, which includes expenditure on tuition and remedial classes at school and outside school. I refer to the latter as tuition spending for simplicity. In this case, three mediators are involved — the child's study time and two types of investment. The SEM to estimate becomes slightly more complicated:

$$
\begin{aligned}
P_\ell &= \gamma_{0,\ell} + \gamma_{T,\ell} \cdot T + \gamma_{W,\mathrm{tu},\ell} \cdot W_{\mathrm{tu}} + \gamma_{W,\mathrm{nu},\ell} \cdot W_{\mathrm{nu}} + \gamma_{D,\ell} \cdot D + \xi_{P,\ell} \cdot X + \epsilon_{P,\ell}, \\
P_m &= \gamma_{0,m} + \gamma_{T,m} \cdot T + \gamma_{W,\mathrm{tu},m} \cdot W_{\mathrm{tu}} + \gamma_{W,\mathrm{nu},m} \cdot W_{\mathrm{nu}} + \gamma_{D,m} \cdot D + \xi_{P,m} \cdot X + \epsilon_{P,m}, \\
T &= a_T + b_T \cdot D + \xi_T \cdot X + \epsilon_T, \\
W_{\mathrm{tu}} &= a_{W,\mathrm{tu}} + b_{W,\mathrm{tu}} \cdot D + \xi_{W,\mathrm{tu}} \cdot X + \epsilon_{W,\mathrm{tu}}, \\
W_{\mathrm{nu}} &= a_{W,\mathrm{nu}} + b_{W,\mathrm{nu}} \cdot D + \xi_{W,\mathrm{nu}} \cdot X + \epsilon_{W,\mathrm{nu}}, \\
D &= \mathbb{1}(a_D + \xi_D \cdot X + \epsilon_D \geq 0),
\end{aligned}
$$

where the subscript i for each unit is suppressed for notational convenience. Using a similar argument as in Section 1.4.2 see that the order condition becomes

$$
|Z| + |\tilde{Z}| \geq 4, \quad |Z| \geq 1.
$$

It is clear that the set of instrumental variables in Section 1.4.4 satisfies this condition. The rank condition can be tested using the same strategy as in Section 1.4.3.

The results are presented in Table 1.9. The model specification results, direct effects and indirect effects through the child's study time are all consistent with Table 1.3. Decomposition of the effects through study time is presented in the top panel of Table 1.10, which shows qualitatively the same results as previous analyses. As for the investment, the effect through both tuition

spending and nutrition spending are significant with large negative effect sizes, and the effect through nutrition spending is larger in sizes. Further decomposing the indirect effects into γ_W and b_W, as shown in the bottom panel of Table 1.10, the left-behind children are suffering from underinvestment in both tuition and nutrition, and the latter is more severe. Therefore, policies targeting at increasing investment in the child should put more weights on nutrition. For instance, conditional cash transfer programs that improve these children's food intakes and nutrition status could be more effective in increasing their school performances than tuition waivers.

Table 1.9. Effect of parental migration on child schooling outcomes (nutrition and tuition)

				(1) Language	(2) Math
Direct Effect					
Parental Accompany				-0.431**	-0.357*
				(0.003)	(0.011)
Indirect Effect					
Study time				0.004	0.002
				(0.092)	(0.426)
Tuition				-0.495**	-0.490**
				(0.007)	(0.007)
Nutrition				-0.950***	-0.895***
				(0.001)	(0.001)
Sepecification Tests	*(1) Study time*	*(2) Tuition*	*(3) Nutrition*	*(4) Language*	*(5) Math*
Underidentification test					
(Anderson canon. corr. LM statistic)	29.868***	29.868***	29.868***	14.696***	14.696***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Overidentification test (Sargan statistic)	2.752	1.189	0.559	0.287	0.713
	(0.097)	(0.276)	(0.454)	(0.592)	(0.398)
Endogeneity test	2.366	2.865	13.350***	26.117***	29.262***
	(0.124)	(0.091)	(0.000)	(0.000)	(0.000)
Obs.			1971		

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.10. Decomposition of indirect effects of migration (nutrition and tuition)

	(1) Study Time	(2) Tuition	(3) Nutrition	(4) Language	(5) Math
Through Study Time					
Migration (b_T)	-0.390*				
	(0.012)				
Study Time (γ_T)				-0.012	-0.005
				(0.092)	(0.426)
Through Investment					
Migration (b_W)		-2.472***	-4.128***		
		(0.001)	(0.001)		
Tuition ($\gamma_{W,\text{tu}}$)				0.200**	0.198**
				(0.007)	(0.007)
Nutrition ($\gamma_{W,\text{nu}}$)				0.230***	0.217***
				(0.001)	(0.001)

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.6. Conclusion and Discussions

In this paper, I disentangle the total effect of parental out-migration to the child's schooling performance into three mechanism-specific effects through parental absence, child's study time, and investment in the child via a mediation analysis. Using the RUMiC data on rural households from nine provinces in China, I find that the effects through parental absence and investment are both significantly negative with large sizes, while the effect through child's study time is insignificant with a negligible size. The surprising negative effect through investment is mainly driven by reduced nutrition investment by de facto custodians, who may not have compatible incentives to allocate the remittances on the child. The subgroup analysis reveals a shocking gender inequality that girls are suffering ten times more from the underinvestment than boys. These mechanism-specific effects show relative importance of each policy interventional target, thereby providing stronger policy implications than the net effect estimated in previous works. For example, the findings in this paper suggest that policies which compensate for underinvestment, especially for girls and younger children in the household, tend to be more effective in mitigating the negative effect of migration than other types of policies. In particular, policies that increase the nutrition spending on left-behind children also tends to be effective in improving the human capital of left-behind children.

There are a few extensions that are worth discussing. First, the negative effect through investment corroborates the importance to study the role of de facto custodians. As mentioned in Section 1.2.1, we can consider a three-agent model with the guardian included. Instead of taking $\gamma(d)$ as an exogeneous factor, we can view it as the decision variable of the guardian. Denote it by γ for simplicity. The first-period utility function of the guardian depends on the his/her own consumption, which is $(1-\gamma-\gamma_p)W_p(d)$ by definition. The equilibrium solution for γ is determined by how much the guardian care about the future human capital of the child, which can be characterized by the discount factor for the second-period utility. We can study whether the equilibrium proportion of household income invested in the child is increasing or decreasing in d. Empirically, one needs to collect more information about the de facto custodians in order to estimate this part of the model.

Second, the other facet of the problem, namely the effect of the child's schooling performance on parent migration decision, is also interesting and of no less policy importance, as pointed

out in Section 1.2.4. For this research question, the child's schooling performance becomes the major endogenous variable. However, compared to parental migration decision, it is much more challenging to find valid instrumental variables. For instance, the school quality may affect both the performance and study time, implying that school-level information is needed. However, such information is not available in the RUMiC survey, which is focused on adult migrants rather than their children. A well-designed child-centered survey is needed to address this question.

Finally, as shown in Section 1.5.3, it is straightforward to decompose the pathways or add other pathways into the empirical framework. For instance, one can also investigate the effect through the child's time spent on housework since this may have a negative effect due to fatigue or a positive effect due to the aspiration to leave rural areas. Furthermore, the methodology is quite flexible and can be applied to evaluate the effect of other types of parents' labor market participation on child education.

CHAPTER 2

What Do We See in the Lights? Lights at Night and Measures of National Growth

2.1. Introduction

Measuring national-level economic activities is fundamental and crucial to studying economic growth and poverty. While Gross Domestic Product (GDP) is the most popular metric, the quality by which it is measured has been a serious concern, especially for less-developed countries lacking strong government statistical infrastructure. A nation's ability to collect, analyze, and disseminate high-quality data about its population and economy is referred to as "statistical capacity". [1] According to the Penn World Table (PWT) Version 10.0 [Feenstra et al., 2015, Zeileis, 2021], the statistical capacity has a highly unequal distribution across countries and over time, which leads to the uncertainty in international comparisons of GDP [Deaton and Heston, 2010]. Recently, alternative approaches to measure economic activities have been proposed based on large-scale surveys. For example, Young [2012] uses the expenditure measures in the Demographic and Health Survey (DHS) data to estimate the growth rate of real consumption in sub-Saharan countries, which is higher than that indicated by official accounts. Despite the much higher quality, the availability is hindered by the high cost of national-level surveys, especially in developing countries.

To break the trade-off between quality and availability, creative proxy measures have been proposed. Among others, the use of nighttime light data from satellite images has become popular in recent years. A vast literature argues that the nighttime light luminosity and economic activity levels are highly correlated, ranging from the world' largest economy to some least developed countries [e.g. Croft, 1978, Elvidge et al., 1997, Sutton and Costanza, 2002, Ghosh et al., 2009, 2010, Henderson et al., 2018, Hu and Yao, 2021]. Unlike GDP, which has high availability but potentially low quality, or the survey-based measures, which have high quality but low availability, the nighttime light-based measures excel on both ends — advances in remote sensing and image

[1] Source: https://datatopics.worldbank.org/statisticalcapacity/

processing technologies have increased the quality of nighttime light measures which are publicly available at low cost. On top of that, the nighttime light-based measures are available at relatively high temporal frequency and spatial resolution — the raw data is collected on a daily basis and further processed by the National Oceanic and Atmospheric Administration (NOAA) into monthly imagery at a spatial resolution of 30 x 30 arc seconds, or approximately 1000 x 1000 meters at the equator.

Although nighttime light-based measures have unique advantages in assessing the level of economic activities, the extremely strong correlation with other measures such as the reported GDP found in the literature is suspicious. For example, Elvidge et al. [1997] found that the correlation between GDP and nighttime light luminosity (after log-transformation) is as high as 0.97 for 21 countries. Similarly high correlations have been found for other sub-populations [e.g. Sutton and Costanza, 2002, Chen and Nordhaus, 2011]. Taking a step back, if we believe that reported GDP is not reliable enough, the reliability of any other measure that has 80% – 90% correlation with GDP should be questionable as well. As a result, an extremely strong correlation with the reported GDP disqualifies the nighttime light measures. On the other hand, it is apparent that the nighttime could not reflect every aspect of economic activities and, thus, cannot be near-perfectly correlated with GDP.

In this paper we demonstrate that these high correlations are indeed unreasonable and the main reason is the failure to adjust for non-stationarity. The correlation between two non-stationary time series tends to be excessively large even if no such relationship is present in the data generating process. This phenomenon is called "spurious correlation" in macro-econometrics [Granger and Newbold, 1974]; see also standard textbooks [e.g. Stock et al., 2012, Wooldridge, 2015]. After adjusting for non-stationarity properly, the association declines drastically. The substantial drop also corroborates that the issue of spurious correlation cannot be neglected.

Another challenge is the heterogeneity in the light-GDP association. One contribution of this paper is to show that nighttime light is an appropriate measure of GDP in some contexts but not in others. The marginal impact of "true" GDP on nightlight intensity can differ across countries with different geographic or cultural characteristics. In addition, the relationship between "true" GDP and measured GDP varies depending on statistical capacity and other things. While most existing works implicitly assumes the homogeneity in their analyses, we explicitly build the heterogeneity

into the model. Under our model, we show that the usual ordinary least squares (OLS) and fixed-effect regression estimators can be misleading in that they weigh countries in a highly imbalanced way. In particular, they put very low weights on smaller or less-developed countries while the weights on other countries are unequal and uninterpretable.

To address the heterogeneity, we carefully define our inferential targets. The first class of targets are the average correlation coefficients (ACC), defined as a weighted average of the individual correlation coefficients (ICC) between nighttime light and GDP measures for each country with user-specified weights. In particular, we are interested in the ACC over all countries and ACC over a certain sub-population, such as the middle-income countries. We propose a weighted least squares (WLS) first-differencing regression estimator that is an unbiased estimator of the ACC without assuming stationarity. The p-value and confidence interval can be computed by standard software. The second class of inferential targets are non-zero ICCs. We cannot expect to estimate all ICCs to a reasonably high accuracy because the panel is short and data is scarce for each country. Fortunately, our preliminary analysis shows that the ICC is close to zero for most countries. We exploit this sparsity and apply the LASSO regression to identify and estimate non-zero ICCs. To control the fraction of false discoveries, we apply the advanced "knockoff" method [Barber and Candès, 2015] to control the false discovery rate (FDR) at level 0.2, meaning that the fraction of true positives is at least 80% on average. This identifies in 10 countries.

2.1.1. Data. The nighttime light data we analyze in this paper is provided by Proville et al. [2017] and collected by the Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) satellites, which provides global daily measurements of nocturnal light. Annual average composite images of nighttime light have been released by NOAA since 1992, including 30 x 30 arc-second grids covering -180 to 180 degrees longitude and -65 to 75 degrees latitude. In particular, we use the stable light measurements that measures persistent lighting, which dismiss fires as ephemeral events and remove background noises. Each pixel in the nighttime light image has a digital number (DN) representing light intensity, ranging from 0 to 63. The measure we use for light intensity is the area lit during 1992 to 2013. To calculate area lit, the first step is to calculate pixel count, which is the sum of DN for all pixels with $DN > 31$ (to remove weak light signals) in grids of 30 x 30 arc seconds. Next, the pixel counts are converted to the equivalent area

coverage in square kilometers. This is the area lit measure we use. On the other hand, we use the World Bank nominal GDP at the current US dollar levels.

2.1.2. Other related work. The nighttime light measures have been used as a proxy for the level of economic activity in different problems. Ghosh et al. [2009] estimated the light-GDP association in the United States, and then applied the estimated association in Mexico to predict Mexico's economic activity levels, concluding that Mexico's informal economy and remittances are much higher than reported in national accounts. Henderson et al. [2018] studied the association between trade and agricultural factors and the distribution of economic activities proxied by the nighttime light measure. The nighttime light measures are also used to construct better predictors for economic growth. For example, Henderson et al. [2012] developed a measure as a weighted average of the income growth from national accounts and the income growth predicted by the nighttime light measures. The weight reflects the statistical capacity and differs by countries.

Because of the high spatial resolution, the nighttime light measures are also used as proxies for economic activities at the sub-national level [e.g. Sutton and Costanza, 2002, Sutton et al., 2007, Hodler and Raschky, 2014, Harari, 2020], the region level [e.g. Doll et al., 2006], or even abstract areas such as ethnic homelands [Michalopoulos and Papaioannou, 2013].

The rest of this paper is structured as follows. We begin Section 2 by describing our model and comparing the limitation of popular methods in estimating model parameters. We then propose a weighted least square estimator in Section 3, which addresses the limitations of methods mentioned in Section 2. We also present the estimated relationship between nighttime light and GDP — the highest and strongest relationship exists only in middle-income countries. With these results of subgroup average association, we are curious about the individual-level association. In Section 4, we further estimate the correlation coefficients between nighttime light and GDP for each country. Inspired by the fact that many country-level associations observed in the preliminary analysis is close to zero, we apply the LASSO regression to identify and estimate non-zero individual correlation coefficients. We further apply the "knockoff" method to control the false discovery rate among the selected countries with non-zero coefficients, and the results from LASSO regression and "knockoff" method are very consistent.

2.2. Model

2.2.1. A panel data model with country-wise heterogeneity and non-stationarity. For country $i \in \{1,\dots,I\}$ in year $t \in \{1,\dots,T\}$, let L_{it} denote the logarithmic nighttime light measure, Y_{it} denote the reported measure of GDP, and Y^*_{it} denote the (unobserved) measure of actual level of economic activities. In this paper, we take Y_{it} as logarithm of reported annual GDP, and conceptualize Y^*_{it} as the true annual GDP. Intuitively, Y^*_{it} is associated with both Y_{it} and L_{it}, though not perfectly correlated. Their relationships can be abstracted through the following simplified model:

$$L_{it} = \eta^L_i + \gamma^L_i Y^*_{it} + \epsilon^L_{it}, \quad Y_{it} = \eta^Y_i + \gamma^Y_i Y^*_{it} + \epsilon^Y_{it}, \tag{2.2.1}$$

where $(\epsilon^L_{it}, \epsilon^Y_{it})$ are error terms that capture the effects of other economic variables. Here, we consider a country-specific intercept and slope in both equations in order to capture the country-wise heterogeneity. Indeed, the effect of actual economic activities on the nighttime light measure depends on the level of industrialization, environmental policies, lifestyles of the citizens, and so on, while the mechanisms through which each country misreports their GDP are affected by the statistical capacity, political ideology, technology, etc. It is thus unrealistic to impose a constant parameter for all countries.

Since Y^*_{it} is a latent variable, the parameters are unidentifiable without further structural or distributional assumptions. While some previous works attempted to estimate γ^L and γ_Y and recover Y^*_{it} [e.g. Henderson et al., 2012], this paper is focused on exploring the marginal association between L_{it} and Y_{it}. A reduced-form model yielded by (2.2.1) is

$$L_{it} = \alpha_i + \beta_i Y_{it} + \epsilon_{it},$$

where

$$\beta_i = \gamma^L_i/\gamma^Y_i, \quad \alpha_i = \eta^L_i - \eta^Y_i \gamma^L_i/\gamma^Y_i, \quad \epsilon_{it} = \epsilon^L_{it} - \epsilon^Y_{it}\gamma^L_i/\gamma^Y_i. \tag{2.2.2}$$

Under the above reduced-form model, the inferential targets are β_is, which measures the country-wise association between the nighttime light measure and the reported GDP.

Since we are analyzing a fairly long panel (with 17 years), we cannot neglect the non-stationarity of the nighttime light measure and reported GDP. It is widely accepted that GDP measures have a

unit root [e.g. McCallum, 1993, Libanio, 2005]. Unless the nighttime light measure is cointegrated with the reported GDP, ϵ_{it} should be a non-stationary process for each country i. The cointegration is arguably impossible since the nighttime light measure only captures part of economic activities. For this reason, we model the error terms as a unit-root process of order 1.

For reference, we summarize the model below:

$$L_{it} = \alpha_i + \beta_i Y_{it} + \epsilon_{it}, \quad \epsilon_{it} = \epsilon_{i(t-1)} + \nu_{it}, \quad \nu_{it} \text{ is a mean-zero stationary process for each } i. \tag{2.2.3}$$

The above model is similar to a non-stationary dynamic panel model because it can be equivalently formulated as

$$L_{it} = L_{i(t-1)} + \beta_i Y_{it} - \beta_i Y_{i(t-1)} + \nu_{it}. \tag{2.2.4}$$

Nevertheless, (2.2.3) is more complicated due to the country-wise heterogeneity, which introduces the incidental parameter problem.

2.2.2. Inferential targets: average and individual correlation coefficients. It would be ideal to estimate every individual correlation coefficient (ICC) β_i precisely. However, this can only be achieved with a sufficiently long panel because there are only T observations $(L_{it}, Y_{it})_{t=1}^{T}$ carrying information on β_i, and there are at least two more nuisance parameter: the intercept α_i and the variance of ν_{it}. In our case, $T = 22$ and thus only 20 degree-of-freedom can be used to estimate β_i. Therefore, it is impossible to estimate all β_i's with a desirable accuracy without further assumptions (i.e, under unrestricted heterogeneity).

In the presence of unrestricted heterogeneity, a routine approach to summarize the effects is to consider an average correlation coefficient (ACC), as an average of ICCs with *user-specified* weights. For example, we can consider the ACC over all countries

$$\beta_{\text{all}} \triangleq \frac{1}{I} \sum_{i=1}^{I} \beta_i. \tag{2.2.5}$$

Sometimes it could be more informative to investigate a weighted average of β_i:

$$\beta(v) = \frac{\sum_{i=1}^{I} v_i \beta_i}{\sum_{i=1}^{I} v_i}, \tag{2.2.6}$$

where the weights depend on, for example, the population size. Similarly, we can define the ACC over a sub-population, say, the middle-income countries. As we will show in Section 2.3, the effective sample size for ACC is much larger than T, and hence the variance of the ACC estimator is much lower than that of the ICC estimator.

The ICCs could also be of interest, when the goal is to find the countries with largest ICCs, or to explore how the ICC varies with other factors. As mentioned above, we need to impose additional plausible assumptions in order to ensure inferential reliability. In this paper, we exploit the sparsity of β_i's observed from the preliminary analysis. It may sound implausible because the overall correlation between the nighttime light measure and GDP has been found to be high in the literature [e.g. Elvidge et al., 1997, Sutton and Costanza, 2002, Chen and Nordhaus, 2011]. However, we will show in Section 2.2.3.1 that the high correlation is spurious due to the failure of adjusting for non-stationarity. After the correct adjustment, most ICCs are substantially reduced and close to zero. On the other hand, recalling that β_i is the ratio between dL_{it}/dY^*_{it} and dY_{it}/dY^*_{it}. it is large only when either the nighttime light captures most economic activities or there is substantial amount of under-reporting. Intuitively, neither of them is realistic. Therefore, we argue that it is reasonable to assume sparsity.

2.2.3. Challenges.

2.2.3.1. *Non-stationarity and spurious correlation.* It is known in the macro-econometric theory that the correlation coefficient between two non-stationary sequences is typically large and can sometimes be close to ± 1, even if the two sequences are independent [e.g. Phillips, 1998, Ernst et al., 2019]. This surprising phenomenon is dubbed "spurious correlation" in the literature. Failure to adjust for spurious correlation properly can yield misleading conclusions. In fact, for the countries under study in this paper, we test for unit root with ADF test (critical ADF p-value $>$ 0.1) [Cheung and Lai, 1995] and KPSS test (critical KPSS p-value $<$ 0.1) [Kwiatkowski et al., 1992]. We find that about 70% of the countries have non-stationarity in GDP, and 40% of countries have non-stationarity in nighttime light.

Unfortunately, this issue has not been addressed in the literature on nighttime light measures. Suppose the non-stationarity is absent and thus the spurious correlation does not exist, the correlation coefficient between L_{it} and Y_{it} (or R^2 equivalently) should be similar to that between $\Delta L_{it} = L_{it} - L_{i(t-1)}$ and $\Delta Y_{it} = Y_{it} - Y_{i(t-1)}$. As a sanity check, I perform the ADF test and KPSS

test again after taking the first difference for GDP and nighttime light, and the corresponding share of countries demonstrating non-stationarity drop to about 20% and less than 5% respectively. Note that we use the 10% significance level here, so there is 10% chance that we get reports of non-stationarity when there actually is not, Thus, 20% demonstrating non-stationarity is not quite a large number that we should worry about.

Using our data, if L_{it} and Y_{it} are given by the levels (without any transformation), the first correlation is 0.96 (95% confidence interval $[0.95, 0.97]$), coinciding with the observation by Elvidge et al. [1997], while the second correlation is 0.18 (95% confidence interval $[0.10, 0.26]$); if L_{it} and Y_{it} are given by the log-levels, as in our analyses, so that the first-order difference measures the growth rate, the first correlation is 0.31 (95% confidence interval $[0.23, 0.39]$), while the second correlation is 0.00 (95% confidence interval $[-0.08, 0.08]$). The confidence intervals are non-overlapping with a large gap in both cases, suggesting strong evidence for spurious correlation.

2.2.3.2. *Failure of OLS estimators on the pooled sample.* A popular approach is to run the OLS regression on the pooled panel data [e.g. Henderson et al., 2012]. However, this approach relies on three assumptions: homogeneity of ICCs ($\beta_i \equiv \beta$), homogeneity of intercepts ($\alpha_i \equiv \alpha$), and stationarity of error terms. Even if the last two assumptions both hold, the heterogeneity in ICCs can make the resulting estimator uninterpretable. To illustrate the failure of the OLS estimator, assume $\alpha_i = 0$ and ϵ_{it} are stationary for a moment. It is not hard to see that

$$\mathbb{E}[\hat{\beta}_{\mathrm{OLS}}] = \frac{\sum_{i=1}^{I} w_i^{\mathrm{OLS}} \beta_i}{\sum_{i=1}^{I} w_i^{\mathrm{OLS}}},$$

where

$$w_i^{\mathrm{OLS}} = \sum_{t=1}^{T} Y_{it}^2.$$

This limit is a weighted average of ICCs, where the weights are data-dependent. One implication is that the countries with larger levels of GDP have larger weights. Given the dramatic imbalance in country-level GDP, the weights for smaller or less developed countries are almost zero. Therefore, the OLS estimator essentially estimates a weighted average of a few large countries in the world, which is clearly misleading.

A more sophisticated approach is to add country fixed-effects into the regression [e.g. Henderson et al., 2012]. This relaxes the assumption of homogeneous intercepts. Nevertheless, the

heterogeneity in β_i's is still problematic even if the error terms are stationary. Through some tedious calculations, we can show that,

$$\mathbb{E}[\hat{\beta}_{\mathrm{FE}}] = \frac{\sum_{i=1}^{I} w_i^{\mathrm{FE}} \beta_i}{\sum_{i=1}^{I} w_i^{\mathrm{FE}}},$$

where

$$w_i^{\mathrm{FE}} = \sum_{t=1}^{T} (Y_{it} - \bar{Y}_i)^2, \quad \bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}.$$

Again, the limit of this estimator severely penalizes smaller or less-developed countries.

Another commonly used method in panel data analysis is the first-difference regression estimator, which regresses $\Delta L_{it} = L_{it} - L_{i(t-1)}$ on $\Delta Y_{it} = Y_{it} - Y_{i(t-1)}$, though we are not aware of the application in the literature on nighttime light measures. Under our model (2.2.3),

$$\Delta L_{it} = \beta_i \Delta Y_{it} + \nu_{it}.$$

Therefore, the first-difference estimator relaxes the assumptions on both the homogeneity of intercepts and the stationarity of error terms. Despite being more robust than the aforementioned two estimators, it still suffers from the same issue under the heterogeneity of ICCs. As with $\hat{\beta}_{\mathrm{OLS}}$, we can show that

$$\mathbb{E}[\hat{\beta}_{\mathrm{FD}}] = \frac{\sum_{i=1}^{I} w_i^{\mathrm{FD}} \beta_i}{\sum_{i=1}^{I} w_i^{\mathrm{FD}}},$$

where

$$w_i^{\mathrm{FD}} = \sum_{t=1}^{T} (\Delta Y_{it})^2.$$

When Y_{it} is given by the log-GDP, ΔY_{it} approximately measures GDP growth rate. While the variation of GDP growth rate is much smaller than that of GDP, the FD weights are still non-uniform and they are difficult to interpret.

TABLE 2.1. Methods for estimating model parameters

Method	Assumptions	Regression (in R syntax)
Ordinary Least Squares regression	$\alpha_i \equiv \alpha$, $\beta_i \equiv \beta$, ϵ_{it} is stationary	$\texttt{lm}(L_{it} \sim \alpha + \beta \cdot Y_{it})$
(Country) Fixed-Effects regression	$\beta_i \equiv \beta$, ϵ_{it} is stationary	$\texttt{lm}(L_{it} \sim \alpha_i + \beta \cdot Y_{it})$
First-Difference regression	$\beta_i \equiv \beta$	$\texttt{lm}(\Delta_t L_{it} \sim \beta \cdot \Delta_t Y_{it})$

2.3. Estimating Average Correlation Coefficients

2.3.1. Weighted first-difference regression estimator. As discussed in Section 2.2.3.2, the first-difference regression estimator requires neither homogeneity of intercepts nor stationarity of errors. In this section, we will propose an adjustment for the first-difference estimator to handle the heterogeneity of ICCs. To set the stage, we start by reformulating the model (2.2.3) based on the first-differenced quantities:

$$\Delta L_{it} = \beta_i \cdot \Delta Y_{it} + \nu_{it}. \tag{2.3.1}$$

A natural estimator for $\beta(v)$ is the aggregated OLS estimator, defined as

$$\hat{\beta}(v) = \frac{\sum_{i=1}^{I} v_i \hat{\beta}_{i,\mathrm{OLS}}}{\sum_{i=1}^{I} v_i},$$

where $\hat{\beta}_{i,\mathrm{OLS}}$ is obtained by the OLS regression for unit i (i.e., on $(L_{it}, Y_{it})_{t=1}^{T}$). Clearly, it is an unbiased estimator of $\beta(v)$ and

$$\mathrm{Var}\left[\hat{\beta}(v)\right] = \frac{\sum_{i=1}^{I} v_i^2 \mathrm{Var}[\hat{\beta}_{i,\mathrm{OLS}}]}{\left(\sum_{i=1}^{I} v_i\right)^2}.$$

Since $\hat{\beta}_{i,\mathrm{OLS}}$ is the best linear unbiased estimator (BLUE) for β_i, $\hat{\beta}(v)$ is BLUE for $\beta(v)$. For example, for β_{all}, the variance of the above estimator is

$$\frac{1}{I^2} \sum_{i=1}^{I} \mathrm{Var}[\hat{\beta}_{i,\mathrm{OLS}}].$$

When the variances of all individual OLS estimators are similar, the above variance can be I times lower than each, implying that the efficiency to estimate ACC is much higher than the efficiency to estimate an ICC.

In Section 2.2.3.2, we have seen that the OLS estimator, whose objective function treats every country equally, does not treats each β_i equally. Therefore, the estimator $\hat{\beta}(v)$, which treats every β_i equally, must treat each country differently. It turns out that $\hat{\beta}(v)$ can be equivalently formulated as a weighted least squares (WLS) regression estimator with properly chosen weights:

$$\hat{\beta}(v) = \operatorname{argmin}_\beta \frac{1}{IT} \sum_{i=1}^{I} \gamma_i \sum_{t=1}^{T} (\Delta L_{it} - \beta \Delta Y_{it})^2, \quad \text{where } \gamma_i = \frac{v_i}{\sum_{t=1}^{T} (\Delta Y_{it})^2}. \tag{2.3.2}$$

To see the equivalence, via the standard computation,

$$\hat{\beta}(v) = \frac{\sum_i \gamma_i \sum_t \Delta Y_{it} \Delta L_{it}}{\sum_i \gamma_i \sum_t (\Delta Y_{it})^2} = \frac{\sum_i \gamma_i \hat{\beta}_{i,\mathrm{OLS}} \sum_t (\Delta Y_{it})^2}{\sum_i \gamma_i \sum_t (\Delta Y_{it})^2} = \frac{\sum_{i=1}^{I} v_i \hat{\beta}_{i,\mathrm{OLS}}}{\sum_{i=1}^{I} v_i}.$$

Consider the ACC over all countries as an example, in which case γ_i is inversely proportional to $\sum_{t=1}^{T} (\Delta Y_{it})^2$. Then $\hat{\beta}(v)$ down-weights countries with larger growth rates while up-weights countries with smaller growth rates. This is unsurprising because the unweighted OLS estimator tends to up-weight the former.

Another advantage of the WLS formulation (2.3.2) is that the variance of $\hat{\beta}(v)$, and thus the p-value and confidence interval, can be computed by standard software directly. Comparisons of these methods are summarized in Table 2.1.

2.3.2. Results. First, we estimate β_{all}, the ACC over all countries defined in (2.2.5). The point estimate is 0.176 with the 95% confidence interval $[0.094, 0.258]$. Though it is significant at the 1% level, the magnitude is substantially smaller than what is found in previous works which did not adjust for spurious correlation [e.g. Chen and Nordhaus, 2011, Henderson et al., 2012]. The huge gap between our estimates and the previous ones suggests that the non-stationarity is non-negligible in studying the relationship between nighttime light measures and measures of wealth such as GDP. Put another way, both variables should be measured by growth rates instead of levels.

Next, we estimate the ACC over subpopulations. The association between the nighttime light measure and GDP is affected by the industrialization levels. As a result, we expect the ACCs to

vary with the income levels. We stratify the countries by income groups defined by World Bank. The results are summarized in Table 2.2. Again, the point estimates are not large in general. The estimates are significant at the 1% level for middle-income countries. This is partly attributed to the "capping" effect: for low-income countries, the luminosity is too low to be detected by the satellites, while for high-income countries, the luminosity is beyond the maximal detectable level. The latter is called the "saturation effect" in the remote sensing literature. Therefore, there is sufficient variation in the nighttime light measure only for middle income countries.

Table 2.2. WLS estimates by income group

	Income Group	Obs.	WLS estimate	p-value	95% C.I.
1	Low income	638	0.130	0.121	(-0.034, 0.295)
2	Upper middle income	1100	0.416***	<0.001	(0.312, 0.521)
3	High income: non-OECD	924	0.004	0.330	(-0.004, 0.011)
4	Lower middle income	968	0.297***	<0.001	(0.177, 0.416)
5	High income: OECD	704	0.111	0.126	(-0.031, 0.253)
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 2.3. WLS estimates by continent

	Continent	Obs.	WLS estimate	p-value	95% C.I.
1	South Asia	154	0.465**	0.023	(0.058, 0.872)
2	Europe & Central Asia	1188	0.017	0.664	(-0.058, 0.092)
3	Middle East & North Africa	440	0.301***	<0.001	(0.196, 0.407)
4	East Asia & Pacific	616	0.003	0.382	(-0.004, 0.01)
5	Sub-Saharan Africa	1012	0.112*	0.058	(-0.004, 0.227)
6	Latin America & Caribbean	858	0.459***	<0.001	(0.34, 0.578)
7	North America	66	0.154	0.617	(-0.446, 0.754)
Note:			*p<0.1; **p<0.05; ***p<0.01		

Lastly, we stratify the countries by geographical locations because the human lifestyles heavily depend on the geography and climate. In particular, we estimate the ACC over seven continents. The results are presented in Table 2.3. Interestingly, we found significant ACCs (at at least the 5% level) at South Asia, Middle East & North Africa, and Latin America & Caribbean, where most developing countries reside in. This corroborates the findings in Table 2.2.

2.4. Estimating Individual Correlation Coefficients

2.4.1. Naive estimates: country-wise regressions. As shown in Section 2.2.2, the ICCs can be estimated by running a first-difference regression for each country separately. Though this naive approach is inefficient due to data scarcity, we present the result here as a benchmark.

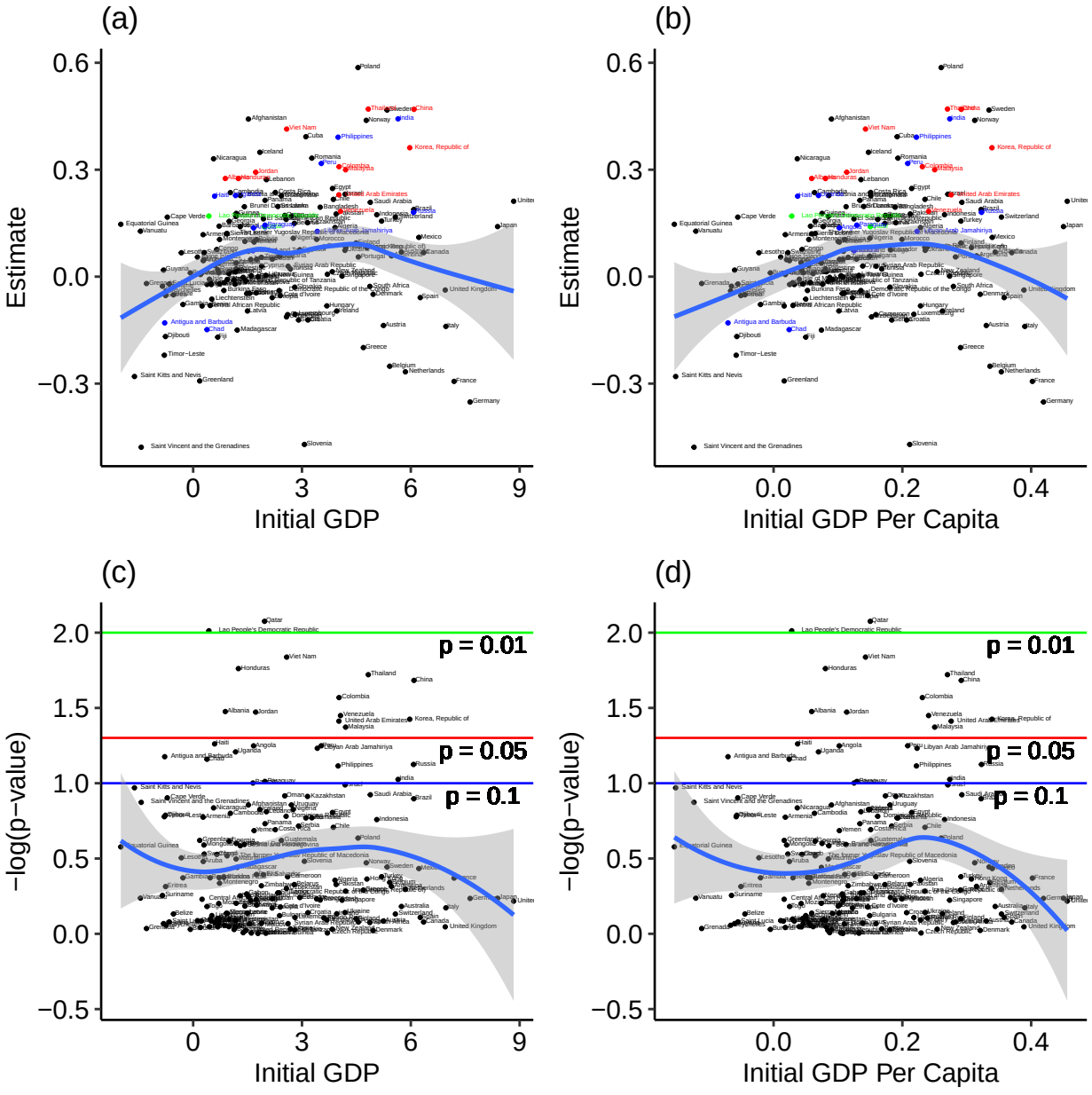


FIGURE 2.1. Estimated ICC over initial income levels and corresponding p-values

For visualization, we plot the estimated ICCs against initial income levels, measured by GDP or per capita GDP in 1992, the initial year when the nighttime light data is available. Figure 2.1(a) and 2.1(b) present the scatter-plots of ICC estimates versus GDP and per capita GDP, respectively. Both figures show the inverse-U shapes, motivating us to fit a quadratic regression:

$$\hat{\beta}_i = \alpha_0 + \alpha_1 Z_{i,1992} + \alpha_2 (Z_{i,1992})^2 + \zeta_i, \tag{2.4.1}$$

where $\hat{\beta}_i$ denotes the estimated ICC and $Z_{i,1992}$ denotes GDP or per capita GDP. Table 2.4 summarizes the regression results. As expected, the coefficient on $Z^2_{i,1992}$ is negative and statistically significant, and the coefficient on $Z_{i,1992}$ is positive and statistically significant. This suggests that countries distributed in the middle range of income tend to have higher positive ICCs, while countries distributed at the lowest or highest end tend to have near-zero ICCs.

Next, we compute the p-values for each ICC estimate. Figure 2.2 plots the histogram of -log(p-values) for ICCs. The vertical blue, red, and green lines correspond to the $p = 0.1, 0.05, 0.01$ thresholds. Only 35 out of the 179 countries or regions have a significant ICC estimate at the 10% level. This could be either due to the fact that the actual ICC is low or that the variance is too large. Furthermore, Figure 2.1(c) and 2.1(d) show the scatter-plots of -log(p-values) for each country versus GDP and per capita GDP in the initial year.

The p-values displayed above are only marginally valid for each country. Since there are 179 p-values, we must adjust for multiplicity to control the number of false discoveries, i.e., the countries with zero ICCs that are claimed to have significant ICCs. Simply thresholding the p-values at level 10% does not guarantee the fraction of false rejections to be controlled. In particular, we apply the Benjamini-Hochberg (BH) procedure [Benjamini and Hochberg, 1995] on these p-values to control the false discovery rate (FDR). Given a target FDR level α, the BH procedure sorts the p-values in ascending order, denoted as $p_{(1)} \leq \ldots \leq p_{(I)}$, and rejects all p-values less than or equal to $p_{(R)}$ where

$$R = \max \left\{ r : p_{(r)} \leq \frac{r\alpha}{n} \right\}.$$

Unfortunately, the BH procedure cannot reject any p-value even if the target FDR level $\alpha = 1$. This suggests that the naive estimates are so inefficient that the evidence is too weak to support any multiplicity adjustment.

TABLE 2.4. Relationship between β_i and initial wealth

	Dependent variable:	
	β_i	
	(1)	(2)
(Initial GDP)^2	-0.009*	
	(0.005)	
Initial GDP	0.056*	
	(0.033)	
(Initial GDP per capita)^2		-3.657**
		(1.431)
Initial GDP per capita		1.119**
		(0.519)
Constant	0.200***	0.208***
	(0.048)	(0.046)
R^2	0.023	0.040
Adjusted R^2	0.010	0.028
Residual Std. Error (df = 160)	0.350	0.347
F Statistic (df = 2; 160)	1.842	3.328**
Note:	*p<0.1; **p<0.05; ***p<0.01	

Next, we apply the BH procedure in each income group separately. This is easier than the previous task because each group involves far fewer countries. Again, no p-value is rejected from any group with $\alpha = 20\%$, which is a commonly considered level. If the target level is raised to 50%, only two countries are rejected: Lao People's Democratic Republic from the lower middle income group and Qatar from the high income non-OECD group. This further confirms the inefficiency of the naive estimates.

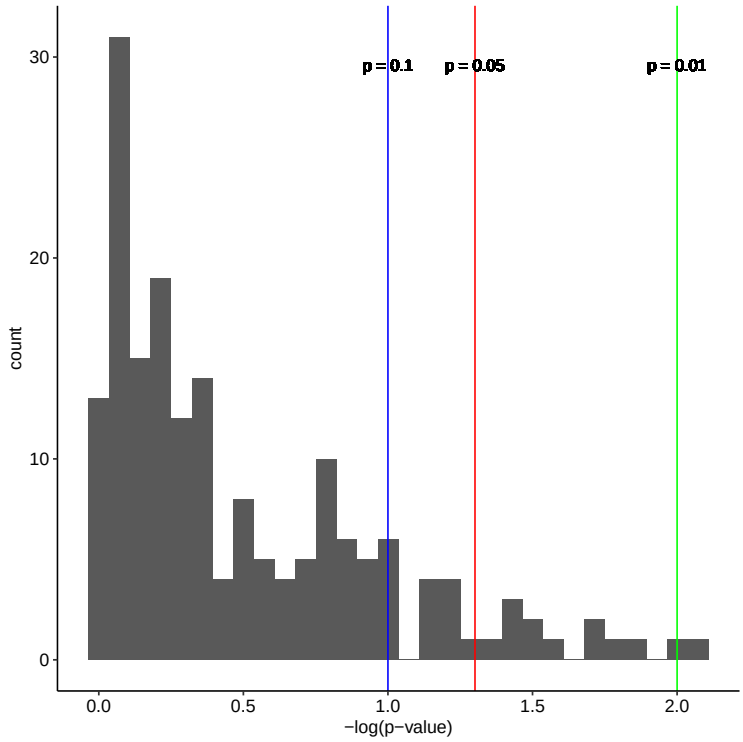


FIGURE 2.2. Histogram of -log(p-values) for 179 countries or regions

2.4.2. Exploiting the sparsity of ICCs via LASSO regression. In Section 2.4.1, we have seen that the naive estimates of β_i's are not only statistically insignificant but also have small sizes. Inspired by this observation, it is reasonable to assume the sparsity of ICCs, i.e., most β_i's are zero. The sparsity enables reliable estimation even when the number of observations used to estimate each the parameter is small [e.g. Bühlmann and Van De Geer, 2011].

Note that our model (2.3.1) can be reformulated as a generic linear model with coefficients $(\beta_1, \ldots, \beta_I)$:

$$\underbrace{\begin{bmatrix} \Delta \boldsymbol{L}_1 \\ \Delta \boldsymbol{L}_2 \\ \vdots \\ \Delta \boldsymbol{L}_I \end{bmatrix}}_{\Delta \boldsymbol{L}} = \underbrace{\begin{bmatrix} \Delta \boldsymbol{Y}_1 & & & \\ & \Delta \boldsymbol{Y}_2 & & \\ & & \ddots & \\ & & & \Delta \boldsymbol{Y}_I \end{bmatrix}}_{\Delta \boldsymbol{Y}} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\nu}_1 \\ \boldsymbol{\nu}_2 \\ \vdots \\ \boldsymbol{\nu}_I \end{bmatrix}, \tag{2.4.2}$$

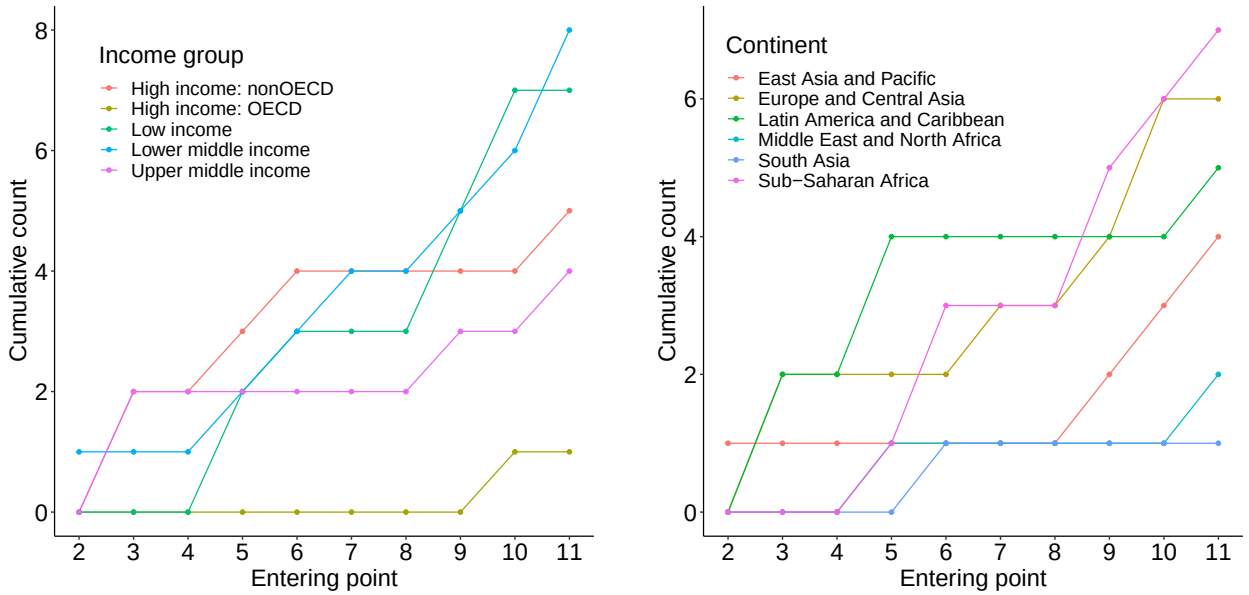
where

$$\Delta \boldsymbol{L}_i = \begin{bmatrix} L_{i1} \\ L_{i2} \\ \vdots \\ L_{iT} \end{bmatrix}, \quad \Delta \boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iT} \end{bmatrix}, \quad \boldsymbol{\nu}_i = \begin{bmatrix} \nu_{i1} \\ \nu_{i2} \\ \vdots \\ \nu_{iT} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_I \end{bmatrix}$$

As a result, we can apply the LASSO regression to estimate the sparse $\boldsymbol{\beta}$. The LASSO regression estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{IT} \sum_{j=1}^{IT} (\Delta \boldsymbol{L}_j - \Delta \boldsymbol{Y}_j \cdot \boldsymbol{\beta})^2 + \lambda \sum_{i=1}^{I} |\beta_i|, \tag{2.4.3}$$

where λ is the penalty level. When $\lambda = 0$, the LASSO estimator reduces to the OLS estimator. As λ grows, the regularization term has a greater effect and thus fewer variables will enter the model. One way to sort the ICCs is based on the largest λ at which each coefficient turns non-zero. Specifically, we choose a grid of penalty levels $\lambda_1 > \lambda_2 > \ldots > \lambda_N$, which are chosen by the `cv.glmnet` function in `R`, and then solve (2.4.3) to find the set of active countries with non-zero coefficients for each λ_k. Figure 2.3 plots the number of active countries in each strata as λ decreases until 25 countries or regions enter the model, where the x-axis represents the entering point, namely the minimal k such that the $\hat{\beta}_i$ turns non-zero with $\lambda = \lambda_k$.



(a) (b)

FIGURE 2.3. Coefficients from LASSO

From Figure 2.3, to order the countries, we apply the LARS algorithm [Efron et al., 2004] which computes the exact λ at which each coefficient turns non-zero. The first 25 countries or regions that enter the model are listed in Table 2.5.

TABLE 2.5. The first 25 countries or regions entering the model

Order	Country/Region	Continent	Income group
1	Rwanda	Sub-Saharan Africa	Low income
2	Grenada	Latin America & Caribbean	Upper middle income
3	Albania	Europe & Central Asia	Upper middle income
4	Greenland	Europe & Central Asia	High income: non-OECD
5	Bosnia and Herzegovina	Europe & Central Asia	Upper middle income
6	Vanuatu	East Asia & Pacific	Lower middle income
7	Fiji	East Asia & Pacific	Upper middle income
8	Gambia	Sub-Saharan Africa	Low income
9	Belize	Latin America & Caribbean	Upper middle income
10	Iceland	Europe & Central Asia	High income: OECD
11	Afghanistan	South Asia	Low income
12	Liechtenstein	Europe & Central Asia	High income: non-OECD
13	Equatorial Guinea	Sub-Saharan Africa	High income: non-OECD
14	Sierra Leone	Sub-Saharan Africa	Low income
15	Timor-Leste	East Asia & Pacific	Lower middle income
16	Cape Verde	Sub-Saharan Africa	Lower middle income
17	Slovenia	Europe & Central Asia	High income: OECD
18	Haiti	Latin America & Caribbean	Low income
19	Serbia	Europe & Central Asia	Upper middle income
20	Georgia	Europe & Central Asia	Lower middle income
21	Libyan Arab Jamahiriya	Middle East & North Africa	Upper middle income
22	Guyana	Latin America & Caribbean	Lower middle income
23	Viet Nam	East Asia & Pacific	Lower middle income
24	Madagascar	Sub-Saharan Africa	Low income
25	Armenia	Europe & Central Asia	Lower middle income

2.4.3. Controlling the fraction of false positives via the knockoff method. While the LASSO regression yields the ordering of the countries or regions that enter the model, it does not provide any statistically meaningful control of false positives. As mentioned in Section 2.4.1, it would be ideal to control the FDR for selected ones. Unfortunately, the BH procedure fails to reject any country when applied to the p-values obtained from the country-wise regressions because they are inefficient. Here, we apply a more advanced procedure called the "knockoff" method which can control FDR for linear models based on the more efficient LASSO estimates [Barber and Candès, 2015].

The knockoff method has gained tremendous popularity in statistics and biology since introduced. Roughly speaking, it constructs a "knockoff" variable as a negative control for each variable of interest and selects variables whose entry to the LASSO path is significantly earlier than its knockoff counterpart. The actual procedure is very complicated and we refer to the readers to Barber and Candès [2015] for a detailed description of the method. Theoretically, the knockoff method controls FDR in finite samples whenever the errors are homoscedastic and Gaussian. The performance is also shown to be robust to heteroscedasticity and non-normality.

Here, we set the target FDR level as 0.2, which is a commonly used level in multiple testing, and apply the `knockoff.filter` function from the `knockoff` package in R [Patterson and Sesia, 2018]. The knockoff method rejects the first 10 countries or regions in Table 2.5. An informal interpretation is that at least 80% of the selected ones are true discoveries on average. Table 2.6 reports the selected countries or regions where we think night lights are the most accurate measure of GDP. Among the ten that demonstrate selected by the knockoff method, six belong to the middle-income group (Grenada, Albania, Bosnia and Herzegovina, Vanuatu, Fiji, and Belize), two belong to the high income group (Greenland, and Iceland), and only two belong to the low income group (Rwanda, and Gambia). This is a warning that we should be very careful when using the nighttime light data as a proxy for GDP for low-income countries or regions.

TABLE 2.6. The 10 countries or regions selected by the knockoff method with FDR control at level 0.2

Order	Country/Regiondisc	Continent	Income group
1	Rwanda	Sub-Saharan Africa	Low income
2	Grenada	Latin America & Caribbean	Upper middle income
3	Albania	Europe & Central Asia	Upper middle income
4	Greenland	Europe & Central Asia	High income: non-OECD
5	Bosnia and Herzegovina	Europe & Central Asia	Upper middle income
6	Vanuatu	East Asia & Pacific	Lower middle income
7	Fiji	East Asia & Pacific	Upper middle income
8	Gambia	Sub-Saharan Africa	Low income
9	Belize	Latin America & Caribbean	Upper middle income
10	Iceland	Europe & Central Asia	High income: OECD

2.5. Conclusion

We investigate the association between the nighttime light measures and reported GDP for 179 countries or regions. Our analyses overcome major limitations in previous works by including

non-stationarity and heterogeneity explicitly in the model. To adjust for non-stationarity, we apply the first differencing technique and find a substantially smaller overall correlation than that found in the literature, suggesting that the latter might be attributed to "spurious correlation", a well-understood phenomenon in macro-econometrics. To deal with heterogeneity, we propose a weighted least square estimator for the average correlation coefficient by properly re-weighting each country, which resolves the issue of unequal weighting for the standard OLS and fixed-effect regression estimators. We find positive and significant average correlation among middle-income countries. Moving beyond the average association, we apply the LASSO regression to identify and estimate non-zero individual correlation coefficients. This is inspired by the sparsity of country-level associations observed in the preliminary analysis. We further apply the "knockoff" method to control the false discovery rate among the selected countries. We find that the majority of countries or regions that demonstrate a strong and significant association between nighttime light and GDP belongs to the middle-income group. This is a warning that the light-GDP association is not universally high, and we should be very careful when using the nighttime light data as a proxy for GDP for low-income countries or regions.

CHAPTER 3

Double Robust Two-Way-Fixed-Effects Regression For Panel Data

3.1. Introduction

We study estimation of causal effects of a binary treatment in a panel data setting with a large number of units, a modest (fixed) number of time periods, and general treatment patterns. Following much of the applied work we focus on least squares estimators with two-way fixed effects (TWFE). We augment this specification with unit-specific weights, leading to the following estimator:

$$\hat{\tau}(\gamma) = \arg\min_{\tau,\alpha_i,\lambda_t,\beta} \sum_{it} (Y_{it} - \alpha_i - \lambda_t - \beta^\top X_{it} - \tau W_{it})^2 \gamma_i \tag{3.1.1}$$

Here Y_{it} is the outcome variable of interest, W_{it} is a binary treatment, and X_{it} are observed exogenous characteristics. Unit-specific weighs γ_i are constructed using both attributes and realized assignment paths $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iT})$, but free of dependence on the outcomes.

We primarily focus on settings with sufficient cross-sectional variation in $\boldsymbol{W}_i$ to consider and estimate the assignment process – a model for W_{it} conditional on observed characteristics, past values of the treatment (but free of dependence on the outcomes). Motivated by the literature on double robust estimation of treatment effects in cross-section settings (Robins et al. [1994]), we use this assignment model to construct the weights $\gamma^\star$ that guarantee that $\hat{\tau}(\gamma^\star)$ converges to the average (equally over units and periods) treatment effect even if the TWFE regression model is misspecified. Perhaps surprisingly, and in contrast to the cross-section double robust literature, using (generalized) inverse propensity score weights (e.g., Rosenbaum and Rubin [1983], Hirano et al. [2003], Imbens and Rubin [2015]) does not work here. The intuition for the failure of the standard inverse propensity score weighting is that the TWFE regression model does not correspond to a consistently estimable conditional expectation because it includes unit fixed effects. In general, we characterize the limiting behavior of $\hat{\tau}(\gamma)$ for a large class of weighting functions

and provide an analytic correspondence between the choice of weights and the resulting causal estimand.

In controlled experiments the assignment process for $\boldsymbol{W}_i$ is known, and in Section 3.2 we show how to use this knowledge to construct $\gamma^\star$, and conduct design-based inference. Under correct specification of the assignment model our inference procedure is valid regardless of the underlying model for potential outcomes, and in particular we do not need to assume any version of parallel trends. Our results substantially generalize the properties established in Athey and Imbens [2018], in particular, allowing for arbitrary assignment process (subject to mild overlap restrictions). These results are then used as a building block in Section 3.3, where the assignment process is unknown, but can be estimated from the data.

After establishing design-based properties of $\hat{\tau}(\gamma^\star)$, we turn to the robustness–the behavior of the estimator in settings where the postulated assignment model is incorrect. At this point, we use the structure of the regression problem (3.1.1) to demonstrate that $\hat{\tau}(\gamma^\star)$ has a strong double-robustness property (Robins et al. [1994], Kang and Schafer [2007], Bang and Robins [2005], Chernozhukov et al. [2018]): it has a small bias whenever either the assignment or the regression model is approximately correct. We view these results as the primary motivation for using our estimator in practice, where we cannot expect the TWFE model or the assignment model to be fully correct.

To construct $\gamma^\star$, we need to solve a nonlinear equation that depends on the support of $\boldsymbol{W}_i$. Practically, this means that the construction varies across different types of designs. In Section 3.4 we provide solutions for several prominent examples, including staggered adoption, *i.e.*, a situation where units opt into treatment sequentially. Another input we need for $\gamma^\star$ is the probability distribution of $\boldsymbol{W}_i$ (generalized propensity score, Imbens [2000])). In Section 3.5 we use two empirical examples to show how to estimate this distribution for the staggered adoption design using duration models. This approach is connected to Shaikh and Toulis [2019] that uses a duration model to test a sharp null hypothesis that specifies that there are no treatment effects.

Our focus on TWFE regression (3.1.1) is motivated by its increased popularity in economics (see Currie, Kleven, and Zwiers [2020] for documentation on this). In applications, this model provides a parsimonious approximation for the baseline outcomes, allowing researchers to capture unobserved confounders and to improve the efficiency of the resulting estimator by reducing

noise. At the same time, recent research shows that regression estimators for average treatment effects based on TWFE models might have undesirable properties, in particular, negative weights for unit-time specific treatment effects. These concerns are particularly salient in settings with heterogeneity in treatment effects and general assignment patterns (*e.g.*, De Chaisemartin and d'Haultfoeuille [2020], Goodman-Bacon [2018], Abraham and Sun [2018], Callaway and Sant'Anna [2018], Borusyak and Jaravel [2017]). Our results show that some of the concerns raised in this literature regarding negative weights disappear once we properly reweight the observations.

Our analysis assumes that the treatment affects only contemporaneous outcomes, thus not allowing for dynamic effects. We make this choice to crystallize the connection between the TWFE regression model (3.1.1) and the assignment process. Importantly, we do not restrict heterogeneity in contemporaneous treatment effects that can vary over units and periods. To test for, or estimate, dynamic treatment effects, one has to compare units that receive treatment at different times. Such comparisons are justified only if we restrict individual heterogeneity in treatment effects or if we treat the assignment as random. Consequently, it is imperative to model both the assignment mechanism and the outcome model. In Bojinov et al. [2020a] the authors show how to use the assignment process to estimate dynamic treatment effects (see also Blackwell and Yamauchi [2021] for the related analysis in large-T setup). Our results suggest that researchers can construct robust estimators by combining Bojinov et al. [2020a] approach to estimation with more conventional dynamic panel regression models using the methods described in the current paper for the static case.

Our results are related to recent literature on doubly robust estimators with panel data. Conceptually the closest paper to us is Arkhangelsky and Imbens [2019] that also emphasizes the role of the assignment process in the same setting and shows double robustness. Our focus, however, is quite different. First, we restrict attention to a very particular and transparent class of estimators (3.1.1). Operationally, our estimator allows applied researchers to combine flexibility and simplicity of standard regression models with available knowledge about the assignment process while retaining statistical guarantees. Second, we show how to estimate a flexible class of average treatment effects with user-specified weights over units and time. The double robustness property in our paper is distinct from the one analyzed recently in the difference-in-difference setting (*e.g.*, Sant'Anna and

Zhao [2020]): our estimator is robust to arbitrary violations of parallel trends assumptions, as long as the assignment model is correctly specified.

We also connect to recent work on causal panel model with experimental data (*e.g.*, Athey and Imbens [2018], Bojinov et al. [2020a], Roth and Sant'Anna [2021]). Similar to these papers, we establish properties of regression estimators under design assumptions. Importantly, we consider a general setting without restricting our attention to staggered adoption design. Our contribution to this literature is the characterization of the behavior of $\hat{\tau}(\gamma)$ for a large class of weighting functions and general designs. By establishing a connection between weighting functions and limiting estimands, we allow users to construct consistent estimators for a pre-specified weighted average treatment effect of interest.

Finally, the form of our estimator (3.1.1) connects it to the Synthetic Difference in Differences (SDID) estimator introduced in Arkhangelsky et al. [2019]. The difference between these two procedures is in the way they construct the weights $\gamma^\star$. The SDID estimator uses pretreatment outcomes to build a synthetic control unit that follows the path of the average treated unit as closely as possible (up to an additive shift). This strategy is infeasible if W_{it} varies over time. However, precisely in situations with enough variation in $\boldsymbol{W}_i$, we can estimate the assignment process and use it to construct the weights $\gamma^\star$. As a result, the two estimators are complementary and can be used in applications with different assignment patterns.

Throughout the paper, we adopt the standard probability notation $O(\cdot), o(\cdot), O_{\mathbb{P}}(\cdot)$, $o_{\mathbb{P}}(\cdot)$. For any vector v, denote by $v^\top$ the transpose of v, $\|v\|_2$ the L_2 norm of v, and by $\text{diag}(v)$ the diagonal matrix with the coordinates of v being the diagonal elements. For a pair of vectors v_1, v_2, we write $\langle v_1, v_2 \rangle$ for their inner product $v_1^\top v_2$. Furthermore, let $[m]$ denote the set $\{1, \ldots, m\}$, I_m the $m \times m$ identity matrix, and $\mathbf{1}_m$ the m-dimensional vector with all entries 1. Finally, the support of a discrete distribution F is the set of elements with positive probabilities under F.

3.2. Reshaped IPW Estimator and Design-based Inference

In this section, we consider a pure design-based setting, i.e., assume that assignment paths $\boldsymbol{W}_i$ have known distributions. The results of this section are directly applicable to situations where $\boldsymbol{W}_i$ are assigned in a controlled experiment. They also serve as a building block for general non-experimental results discussed in Section 3.3.

3.2.1. Setup and Assumptions. We consider a setting with a finite population of n units. In particular, each unit is characterized by a set of fixed potential outcomes $\{Y_{it}(1), Y_{it}(0)\}_{t\in[T]}$. By writing the potential outcomes in this form we assume away any dynamic effects of past treatments on current outcomes (see Imai and Kim [2019] and Arkhangelsky and Imbens [2019]). Given the realized treatment assignment W_{it}, the observed outcomes are defined in the usual way:

$$Y_{it} = Y_{it}(1)W_{it} + Y_{it}(0)(1 - W_{it}). \tag{3.2.1}$$

Throughout this paper, we consider the asymptotic regime with n going to infinity, and fixed $T \geq 2$, i.e., $T = O(1)$.

We define the unit and time-specific treatment effect as:

$$\tau_{it} \triangleq Y_{it}(1) - Y_{it}(0). \tag{3.2.2}$$

For each time period t, we define the time-specific ATE as:

$$\tau_t \triangleq \frac{1}{n}\sum_{i=1}^{n} \tau_{it}, \tag{3.2.3}$$

and consider a broad class of weighted average of time-specific ATE:

$$\tau^*(\xi) \triangleq \sum_{t=1}^{T} \xi_t \tau_t \tag{3.2.4}$$

for some user-specified deterministic weights $\xi = (\xi_1, \ldots, \xi_T)^\top$ such that

$$\sum_{t=1}^{T} \xi_t = 1, \quad \xi_t \geq 0. \tag{3.2.5}$$

We refer to (3.2.4) as a doubly average treatment effect (DATE). For example, the weights $\xi_t = 1/T$ yield the usual ATE over units and time periods. In the difference-in-differences setting with two time periods, $\xi_t = \mathbf{1}_{t=2}$. In a particular application, one might also be interested in an effect with time discounting factor that puts more weight on initial periods, i.e. $\xi_t \propto \beta^t$ for some $\beta < 1$.

For each unit i and a possible assignment path $\boldsymbol{W}_i$ we define the generalized propensity score (Imbens [2000], Athey and Imbens [2018], Bojinov et al. [2020a,b]) – the marginal probability of

such path:

$$\boldsymbol{\pi}_i(\boldsymbol{w}) = \mathbb{P}[\boldsymbol{W}_i = \boldsymbol{w}], \quad \forall \boldsymbol{w} \in \{0,1\}^T. \tag{3.2.6}$$

Given our focus on design-based inference we treat $\boldsymbol{\pi}_i$ as known objects. These functions are unit-specific thus allowing for general experimental designs, for example, stratification based on observed unit characteristics. We impose minimal overlap restrictions on each $\boldsymbol{\pi}_i$:

Assumption 3.2.1. *There exists a universal constant $c > 0$ and a non-stochastic subset $\mathbb{S}^* \subset \{0,1\}^T$ with at least two elements and at least one element not in $\{\boldsymbol{0}_T, \boldsymbol{1}_T\}$, such that*

$$\boldsymbol{\pi}_i(\boldsymbol{w}) > c, \quad \forall \boldsymbol{w} \in \mathbb{S}^*, i \in [n], \quad \textit{almost surely}, \tag{3.2.7}$$

To capture different assignment processes we allow $\boldsymbol{W}_i$ to be dependent across units. Such dependence arises in applications, sometimes for technical reasons (e.g., in case of sampling without replacement as in Athey and Imbens [2018]), and sometimes by the nature of assignment process (spatial experiments). To quantify this dependence we follow Rényi [1959] and define the maximal correlation:

$$\rho_{ij} \triangleq \sup_{f,g} \left\{ \operatorname{corr}\Big(f(\boldsymbol{W}_i), g(\boldsymbol{W}_j) \Big) \right\} \tag{3.2.8}$$

In the main text we maintain a simplified restriction on $\{\rho_{ij}\}_{ij}$ leaving a more general one to Appendix B.1. The assumption is stated as follows:

Assumption 3.2.2. *There exists $q \in (0,1]$ such that as n approaches infinity the following holds:*

$$\frac{1}{n^2} \sum_{i,j=1}^{n} \rho_{ij} = O(n^{-q}). \tag{3.2.9}$$

Since by construction $\frac{1}{n} \leq (1/n^2) \sum_{i,j=1}^{n} \rho_{ij} \leq 1$, q measures the strength of correlation. When $\boldsymbol{W}_i$ are independent across units, (3.2.9) holds with $q = 1$. More generally, when $\{\boldsymbol{W}_i\}_{i=1}^n$ have a network dependency with $\rho_{ij} = 0$ if there is no edge between i and j, (3.2.9) is satisfied if the number of edges is $O(n^{2(1-q)})$. Note that it imposes no constraint on the maximum degree of the dependency graph. Even if the network is fully connected, it can still hold if the pairwise dependence is weak, e.g., sampling without replacement; see Appendix B.1.4. On the other hand, (3.2.9) excludes the

case where different units are perfectly correlated or equi-correlated with a positive maximal correlation that is bounded away from 0.

Our final assumption puts restrictions on the outcomes by requiring that they are bounded:

Assumption 3.2.3. *There exists* $M < \infty$ *such that* $\max_{i,t,w}\{|Y_{it}(w)|\} < M$.

It is presented here only for simplicity. We relax it substantially in Appendix B.1.

3.2.2. Reshaped IPW estimator. We consider a class of weighted TWFE regression estimators. We refer to them as reshaped inverse propensity weighted (RIPW) estimators, and formally define them as follows:

$$\hat{\tau}(\mathbf{\Pi}) \triangleq \underset{\tau,\mu,\sum_i \alpha_i=\sum_t \lambda_t=0}{\arg\min} \sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\mu-\alpha_i-\lambda_t-W_{it}\tau)^2\frac{\mathbf{\Pi}(\boldsymbol{W}_i)}{\boldsymbol{\pi}_i(\boldsymbol{W}_i)}, \tag{3.2.10}$$

where $\mathbf{\Pi}(\boldsymbol{w})$ is a density function on $\{0,1\}^T$, i.e.,

$$\sum_{\boldsymbol{w}\in\{0,1\}^T}\mathbf{\Pi}(\boldsymbol{w}) = 1. \tag{3.2.11}$$

We refer to the distribution $\mathbf{\Pi}$ as a reshaped distribution, and the weight $\mathbf{\Pi}(\boldsymbol{W}_i)/\boldsymbol{\pi}_i(\boldsymbol{W}_i)$ as a RIP weight. To ensure that the RIPW estimator is well-defined, we require $\mathbf{\Pi}$ to be absolutely continuous with respect to each $\boldsymbol{\pi}_i$, i.e.

$$\mathbf{\Pi}(\boldsymbol{w}) = 0 \text{ if } \boldsymbol{\pi}_i(\boldsymbol{w}) = 0 \text{ and } \boldsymbol{W}_i = \boldsymbol{w} \text{ for some } i \in [n]. \tag{3.2.12}$$

The estimator (3.2.10) is feasible for any such $\mathbf{\Pi}$ because $\boldsymbol{\pi}_i$ is assumed to be known.

The reshaped distribution $\mathbf{\Pi}$ can be interpreted as an experimental design. If $\boldsymbol{W}_i \sim \mathbf{\Pi}$, then $\boldsymbol{\pi}_i = \mathbf{\Pi}$ and (3.2.10) reduces to the standard unweighted TWFE regression. If this is not the case, then $\mathbf{\Pi}(\boldsymbol{W}_i)/\boldsymbol{\pi}_i(\boldsymbol{W}_i)$ acts like a likelihood ratio that changes the original design to one provided by $\mathbf{\Pi}$. For cross-sectional data, we would like to shift the distribution to uniform $\{0,1\}$, making the weights equal to $1/2\boldsymbol{\pi}_i(\boldsymbol{W}_i)$ if the fixed effects are not included. This would yield the standard IPW estimator. However, as we alluded to in the introduction, the situation is more complicated with the panel data, and shifting towards the uniform design might not deliver consistent estimators for the DATE of interest. We explore this formally in the next section where we characterize the set of $\mathbf{\Pi}$ that one can use. This interpretation of $\mathbf{\Pi}$ has one caveat: RIP weights only shift the marginal

distribution of $\boldsymbol{W}_i$ to $\boldsymbol{\Pi}$, but they do not say anything about the joint distribution of $\{\boldsymbol{W}_i\}_{i\in[n]}$ which can remain complicated.

3.2.3. DATE equation and consistency of RIPW estimators. We now derive sufficient conditions under which the RIPW estimator is a consistent estimator for a given DATE of interest. The following theorem presents a precise condition for consistency of $\hat{\tau}(\boldsymbol{\Pi})$ for $\tau^*(\xi)$:

THEOREM 3.2.1. *Let $J = I_T - \mathbf{1}_T\mathbf{1}_T^\top/T$ and $\boldsymbol{\tau}_i = (\tau_{i1},\ldots,\tau_{iT})^\top$; fix ξ that satisfies (3.2.5). Under Assumptions 3.2.1, 3.2.2, and 3.2.3, for any reshaped distribution $\boldsymbol{\Pi}$ with support $\mathbb{S}^*$ that satisfies Assumption 3.2.1, as n tends to infinity,*

$$\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi) = O_{\mathbb{P}}\left(\mathrm{Bias}_\tau(\xi)\right) + o_{\mathbb{P}}(1),$$

where

$$\mathrm{Bias}_\tau(\xi) = \left\langle \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\Big[(\mathrm{diag}(\boldsymbol{W}) - \xi\boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\Big], \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\tau}_i - \tau^*(\xi)\mathbf{1}_T)\right\rangle.$$

This result has two user-specified parameters: time weights ξ, and the reshaped distribution $\boldsymbol{\Pi}$. They are naturally connected: to guarantee consistency for $\tau^*(\xi)$ we can select $\boldsymbol{\Pi}$ such that the following holds:

$$\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\Big[(\mathrm{diag}(\boldsymbol{W}) - \xi\boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\Big] = 0. \tag{3.2.13}$$

Alternatively, for a given $\boldsymbol{\Pi}$ we can look for ξ such that (3.2.13) is satisfied. We call (3.2.13) the DATE equation hereafter. For a fixed ξ, it is a quadratic system with $\{\boldsymbol{\Pi}(\boldsymbol{w}) : \boldsymbol{w}\in\{0,1\}^T\}$ being the variables. Together with the density constraint (3.2.11) and the support constraint in Theorem 3.2.1 that $\boldsymbol{\Pi}(\boldsymbol{w}) = 0$ for $\boldsymbol{w}\notin\mathbb{S}^*$, there are $T + 1 + 2^T - |\mathbb{S}^*|$ equality constraints and $|\mathbb{S}^*|$ inequality constraints that impose the positivity of $\boldsymbol{\Pi}(\boldsymbol{w})$ for each $\boldsymbol{w}\in\mathbb{S}^*$. We will show in Section 3.4 that the DATE equation have closed-form solutions in various examples and provide a generic solver based on nonlinear programming in Appendix B.2.5.

Without further restrictions on $\boldsymbol{\tau}_i$, the DATE equation is a necessary condition for consistency of $\hat{\tau}(\boldsymbol{\Pi})$ for $\tau^*(\xi)$. To see this assume that

$$\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\Big[(\mathrm{diag}(\boldsymbol{W}) - \xi\boldsymbol{W}^\top)J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\Big] = z. \tag{3.2.14}$$

for some vector z that is not proportional to ξ. Because we can vary individual treatment effects without changing the average one we can find a set $\{\boldsymbol{\tau}_i : i \in [n]\}$ that yields the same DATE but $\left\langle z, (1/n)\sum_{i=1}^{n}(\boldsymbol{\tau}_i - \tau^*(\xi)\mathbf{1}_T)\right\rangle \neq 0$, leading to inconsistency. For $z = b\xi$ we get that the inner product of the LHS of (3.2.14) and $\mathbf{1}_T$ is 0 while that of the RHS and $\mathbf{1}_T$ is equal to b. This entails that $z = 0$, and thus the DATE equation.

Notably, when the DATE equation has a solution, our estimator is consistent without any restrictions on the potential outcomes, except Assumption 3.2.3. This is in sharp contrast to usual results about TWFE estimators which typically require the trends to be parallel, at least conditionally on observed covariates (e.g., Callaway and Sant'Anna [2018], Sant'Anna and Zhao [2020]). Theorem 3.2.1 shows that if the assignment process is known and DATE equation has a solution then we can correct the potentially misspecified TWFE regression model by simply reweighting the objective function.

Another interpretation of the DATE equation is through the effective estimand given by a fixed reshaped distribution. (3.2.13) can be rewritten as

$$\left(\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])]\right)\xi = \mathbb{E}[\mathrm{diag}(\boldsymbol{W})J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])]. \tag{3.2.15}$$

It is easy to see that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])] &= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])] \\ &= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\left\|\tilde{\boldsymbol{W}} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\tilde{\boldsymbol{W}}]\right\|^2\right],\end{aligned}$$

where $\tilde{\boldsymbol{W}} = J\boldsymbol{W}$. It is strictly positive since the support of $\boldsymbol{\Pi}$ involves a point $\boldsymbol{w} \notin \{\mathbf{0}_T, \mathbf{1}_T\}$, for which $\boldsymbol{w}' \neq 0$. Therefore, (3.2.15) implies that

$$\xi = \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\mathrm{diag}(\boldsymbol{W})J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])]}{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\left\|\tilde{\boldsymbol{W}} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\tilde{\boldsymbol{W}}]\right\|^2\right]}. \tag{3.2.16}$$

By Theorem 3.2.1, in a randomized experiment with $\boldsymbol{\pi}_i \equiv \boldsymbol{\Pi}$, the effective estimand of the unweighted TWFE regression is the DATE with weight vector ξ.

3.2.4. Design-based inference. To enable statistical inference of DATE, we first present an asymptotic expansion showing the asymptotic linearity of RIPW estimators.

THEOREM 3.2.2. *Let* $\boldsymbol{Y}_i^{\text{obs}}$ *be the vector* $(Y_{i1}^{\text{obs}},\ldots,Y_{iT}^{\text{obs}})$. *Further let* $\Theta_i = \boldsymbol{\Pi}(\boldsymbol{W}_i)/\boldsymbol{\pi}_i(\boldsymbol{W}_i)$, *and*

$$\Gamma_\theta \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i, \quad \Gamma_{ww} \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i, \quad \Gamma_{wy} \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i \boldsymbol{W}_i^\top J \boldsymbol{Y}_i^{\text{obs}},$$

and

$$\boldsymbol{\Gamma}_w \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i J \boldsymbol{W}_i, \quad \boldsymbol{\Gamma}_y \triangleq \frac{1}{n}\sum_{i=1}^n \Theta_i J \boldsymbol{Y}_i^{\text{obs}}.$$

Under the same settings as Theorem 3.2.1,

$$\mathcal{D}(\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi)) = \frac{1}{n}\sum_{i=1}^n (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + O_{\mathbb{P}}\left(n^{-2q}\right),$$

where $\mathcal{D} = \Gamma_{ww}\Gamma_\theta - \boldsymbol{\Gamma}_w^\top \boldsymbol{\Gamma}_w$, *and*

$$\begin{aligned}\mathcal{V}_i = \Theta_i \Big\{ &\left(\mathbb{E}[\Gamma_{wy}] - \tau^*(\xi)\mathbb{E}[\Gamma_{ww}]\right) - \left(\mathbb{E}[\boldsymbol{\Gamma}_y] - \tau^*(\xi)\mathbb{E}[\boldsymbol{\Gamma}_w]\right)^\top J \boldsymbol{W}_i \\ &+ \mathbb{E}[\Gamma_\theta]\boldsymbol{W}_i^\top J\left(\boldsymbol{Y}_i^{\text{obs}} - \tau^*(\xi)\boldsymbol{W}_i\right) - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top J\left(\boldsymbol{Y}_i^{\text{obs}} - \tau^*(\xi)\boldsymbol{W}_i\right)\Big\}\end{aligned}$$

Note that the asymptotic linear expansion holds under fairly general dependency structure in the treatment assignments. Below, we derive a valid confidence intervals for $\tau^*(\xi)$ when $\{\boldsymbol{W}_i : i \in [n]\}$ are independent. The general case is discussed in Appendix B.1.4. If $\{\mathcal{V}_i : i \in [n]\}$ are well-behaved, Theorem 3.2.2 implies that

$$\frac{\mathcal{D}\cdot\sqrt{n}(\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi))}{\sigma_n^*} \approx N(0,1), \quad \text{where } \sigma_n^{*2} = (1/n)\sum_{i=1}^n \text{Var}(\mathcal{V}_i),$$

where $\mathcal{D}$ is known by design. If $\{\mathcal{V}_i : i \in [n]\}$ were known, a natural estimator for σ_n^{*2} would be the empirical variance:

$$\hat{\sigma}_n^{*2} = \frac{1}{n-1}\sum_{i=1}^n (\mathcal{V}_i - \bar{\mathcal{V}})^2, \quad \text{where } \bar{\mathcal{V}} = \frac{1}{n}\sum_{i=1}^n \mathcal{V}_i.$$

We should not expect $\hat{\sigma}_n^*$ to converge to σ_n^* since $\mathbb{E}[\mathcal{V}_i]$ in general varies over i. Nonetheless, $\hat{\sigma}_n^*$ is an asymptotically conservative estimate of σ_n^* since

$$\mathbb{E}[\hat{\sigma}_n^{*2}] \approx \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left(\mathcal{V}_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i]\right)^2\right] \approx \sigma_n^{*2} + \underbrace{\frac{1}{n-1}\sum_{i=1}^n \left(\mathbb{E}[\mathcal{V}_i] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i]\right)^2}_{\text{empirical variance of } \mathbb{E}[\mathcal{V}_i]}, \tag{3.2.17}$$

where the second term measures the heterogeneity of $\mathbb{E}[\mathcal{V}_i]$ and is always non-negative, implying that $\hat{\sigma}_n^{*2}$ is a conservative estimator for σ_n^{*2}. This is unsurprising because even in the cross-section case, the asymptotic design-based variance is only partially identifiable due to the unknown correlation structure between two potential outcomes; see e.g. Neyman's variance formula [Neyman, 1923/1990, Rubin, 1974].

In general, $\mathcal{V}_i$ is unknown due to $\tau^*(\xi)$ and the expectation terms. Nonetheless, we can estimate $\mathcal{V}_i$ by replacing each expectation with the corresponding plug-in estimate, i.e.

$$\hat{\mathcal{V}}_i = \Theta_i \Bigg\{ \left(\Gamma_{wy} - \hat{\tau}\Gamma_{ww}\right) - \left(\mathbf{\Gamma}_y - \hat{\tau}\mathbf{\Gamma}_w\right)^\top J \boldsymbol{W}_i + \Gamma_\theta \boldsymbol{W}_i^\top J\left(\boldsymbol{Y}_i^{\text{obs}} - \hat{\tau}\boldsymbol{W}_i\right) - \mathbf{\Gamma}_w^\top J\left(\boldsymbol{Y}_i^{\text{obs}} - \hat{\tau}\boldsymbol{W}_i\right) \Bigg\}, \tag{3.2.18}$$

and use them to compute the variance:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (\hat{\mathcal{V}}_i - \bar{\hat{\mathcal{V}}})^2, \quad \text{where } \bar{\hat{\mathcal{V}}} = \frac{1}{n}\sum_{i=1}^n \hat{\mathcal{V}}_i. \tag{3.2.19}$$

This yields a Wald-type confidence interval for $\tau^*(\xi)$ as

$$\hat{C}_{1-\alpha} = [\hat{\tau}(\mathbf{\Pi}) - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}, \hat{\tau}(\mathbf{\Pi}) + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}], \tag{3.2.20}$$

where z_η is the η-th quantile of the standard normal distribution. Properties of this confidence interval are established in the next theorem.

Theorem 3.2.3. *Assume that* $\{\boldsymbol{W}_i : i \in [n]\}$ *are independent with*

$$\frac{1}{n}\sum_{i=1}^n \text{Var}(\mathcal{V}_i) \geq v_0, \quad \textit{for some constant } v_0 > 0. \tag{3.2.21}$$

Then under Assumptions 3.2.1 and 3.2.3, for any $\alpha \in (0,1)$,

$$\liminf_{n\to\infty} \mathbb{P}\left(\tau^*(\xi) \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha.$$

In Appendix B.1.4, we discuss a generic result for general dependent assignments (Theorem B.1.6), which covers completely randomized experiments, blocked and matched pair experiments, two-stage randomized experiments, and so on. We present a detailed result (Theorem B.1.7) for

completely randomized experiments where $\boldsymbol{W}_i$'s are sampled without replacement from a user-specified subset of $[0,1]^T$. This substantially generalizes the setting of Athey and Imbens [2018] and Roth and Sant'Anna [2021] where the assignments are sampled without replacement from the set of $T+1$ staggered assignments.

3.2.5. Discussion. Theorem 3.2.1 might appear counter-intuitive given well-understood problems of TWFE estimators (e.g., de Chaisemartin and d'Haultfoeuille [2019], Goodman-Bacon [2018], Abraham and Sun [2018]). To put our result in context we emphasize two important features of the setup. First, we restrict attention to static models, and second, we use the randomness that is coming from $\boldsymbol{W}_i$. Both of these restrictions play a key role in Theorem 3.2.1. Absence of dynamic effects implies that we can meaningfully average units with different histories of past treatments. A version of this assumption is inescapable if we want the method to work for general designs where controlling for past history is practically infeasible. As we explain below, randomness of assignments helps to resolve the issue that TWFE estimators put negative weights on some individual treatment effects.

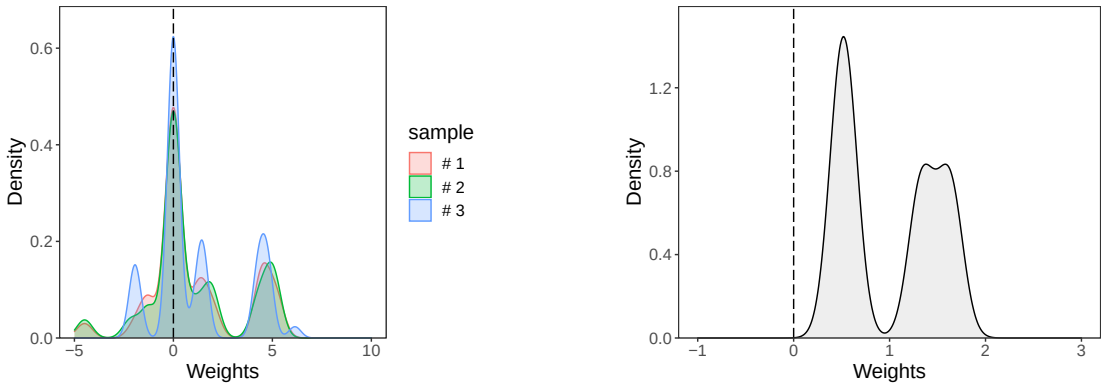
In de Chaisemartin and d'Haultfoeuille [2019], Goodman-Bacon [2018], Abraham and Sun [2018] the authors show that treated units are averaged with potentially negative weights, but these results are conditional on the assignments $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)$ being fixed. Let $\xi_{it}(\gamma; \boldsymbol{W})$ be these weights for the general weighted least squares estimator $\hat{\tau}(\gamma)$ defined in (3.1.1) such that

$$\mathbb{E}[\hat{\tau}(\gamma) \mid \boldsymbol{W}] = \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it}(\gamma; \boldsymbol{W}) \tau_{it},$$

where we now explicitly allow them to depend on $\boldsymbol{W}$. When the assignments are treated as random, the large sample limit of $\hat{\tau}(\gamma)$ is

$$\mathbb{E}[\hat{\tau}(\gamma)] = \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it}(\gamma) \tau_{it},$$

where $\xi_{it}(\gamma) = \mathbb{E}_{\boldsymbol{W}}[\xi_{it}(\gamma; \boldsymbol{W})]$. While $\{(i,t) : \xi_{it}(\gamma; \boldsymbol{W}) < 0\}$ is non-empty almost surely for every realization of $\boldsymbol{W}$, it is still possible that all $\xi_{it}(\gamma)$ are positive due to the averaging over $\boldsymbol{W}$. For illustration, we consider a simulation study with $n = 100, T = 4$ and other details specified in Section 3.5.1. We consider the conditional and unconditional weights induced by the unweighted and RIP weighted TWFE estimator in Figure 3.1 and Figure 3.2 respectively. We plot the histograms



(a) Histograms of $(nT)\cdot\xi_{it}(\gamma;\boldsymbol{W})$'s (b) Histogram of $(nT)\cdot\xi_{it}(\gamma)$'s

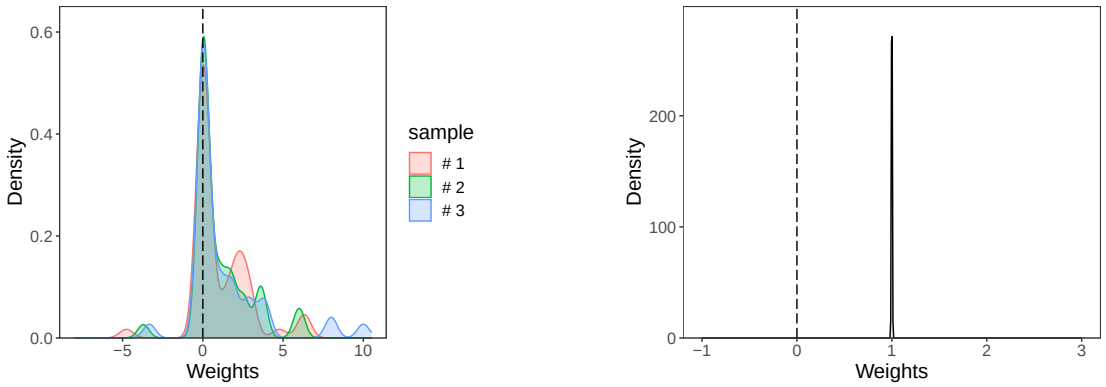
FIGURE 3.1. Effect weights for the unweighted TWFE estimator

of $\{(nT)\cdot\xi_{it}(\gamma;\boldsymbol{W}) : i\in[n], t\in[T]\}$ for three realizations of $\boldsymbol{W}$ and the histogram of $\{(nT)\cdot\xi_{it}(\gamma) : i\in[n], t\in[T]\}$, approximately by averaging over a million realizations of $\boldsymbol{W}$, where the multiplicative factor nT is chosen to normalize the weights into a more interpretable scale. Clearly, despite the large fraction of negative weights in each realization, their averages do not have any negatives. Therefore, the criticism on TWFE estimators does not apply in this case. Indeed, it never applies to the RIPW estimator because all weights are designed to be $1/nT > 0$ when $\boldsymbol{\Pi}$ is a solution of the DATE equation with $\xi = \mathbf{1}_T/T$, as shown in Figure 3.2(b), regardless of the data generating process.

The discussion above demonstrates that while for each cell (i,t) a particular realization of weights can be negative, this fact is not systematic, i.e., on average. If we use the RIPW estimator designed for the equally-weighted DATE, then all cells will receive the same weight. An alternative description of the same phenomenon is that once correctly weighted, the realized treatment paths $\boldsymbol{W}_i$ are uncorrelated with potential outcomes. This independence implies that there cannot be systematic differences in treatment effects among units with distinct assignment paths. The presence of such heterogeneity (together with dynamic treatment effects) is the main reason why negative weights arise in practice.

3.3. Doubly Robust Inference

In this section, we consider a non-experimental setting. In particular, we no longer assume that the distribution of $\boldsymbol{W}_i$ is known. Moreover, to incorporate the standard TWFE model, we allow outcomes to be random as well.



(a) Histograms of $(nT)\cdot\xi_{it}(\gamma;\boldsymbol{W})$'s (b) Histogram of $(nT)\cdot\xi_{it}(\gamma)$'s

FIGURE 3.2. Effect weights for our RIPW estimator

3.3.1. An extended causal framework. The finite population framework is insufficent to handle outcome modelling since the potential outcomes are assumed to be arbitrary fixed quantities. Therefore, we consider a more general framework that is suitable for both treatment and outcome modelling and includes the finite population framework as a special case. In particular, we assume that each unit is characterized by $\mathcal{Z}_i \triangleq \{(Y_{it}(1), Y_{it}(0), X_{it}, U_{it}, W_{it}) : t \in [T]\}$, where X_{it} is a vector of (potentially) time-varying observed confounders, U_{it} is a vector of (potentially) time-varying unobserved confounders. We further assume that $\mathcal{Z}_i$ are i.i.d. samples from a distribution. For notational convenience, we write $\boldsymbol{Y}_i(1)$ for $(Y_{i1}(1),\ldots,Y_{iT}(1))$, $\boldsymbol{Y}_i(0)$ for $(Y_{i1}(0),\ldots,Y_{iT}(0))$, $\boldsymbol{X}_i$ for $(X_{i1},\ldots,X_{iT})$, and $\boldsymbol{U}_i$ for $(U_{i1},\ldots,U_{iT})$. We assume latent mean ignorability:

ASSUMPTION 3.3.1. (LATENT MEAN IGNORABILITY)

$$\mathbb{E}[(\boldsymbol{Y}_i(1),\boldsymbol{Y}_i(0)) \mid \boldsymbol{W}_i,\boldsymbol{X}_i,\boldsymbol{U}_i] = \mathbb{E}[(\boldsymbol{Y}_i(1),\boldsymbol{Y}_i(0)) \mid \boldsymbol{X}_i,\boldsymbol{U}_i] \tag{3.3.1}$$

The definition of the individual treatment effect is modified as

$$\tau_{it} \triangleq \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid \boldsymbol{X}_i,\boldsymbol{U}_i]. \tag{3.3.2}$$

The time-specific ATE and DATE are defined as in (3.2.3) and (3.2.4), respectively. Throughout the rest of this section, we treat $\{(\boldsymbol{X}_i,\boldsymbol{U}_i) : i \in [n]\}$ as fixed. Put another way, the inferential claims, such as consistency and coverage, are conditional on all observed and unobserved confounders. Conceptually, the conditional estimand $\tau^*(\xi)$ is similar to the unconditional estimand $\mathbb{E}[\tau^*(\xi)]$;

indeed, $\tau^*(\xi) - \mathbb{E}[\tau^*(\xi)] = O_{\mathbb{P}}(1/\sqrt{n})$ because $(\boldsymbol{X}_i, \boldsymbol{U}_i)$ are i.i.d.. As a consequence, a conditionally consistent estimator for $\tau^*(\xi)$ is also unconditionally consistent for $\mathbb{E}[\tau^*(\xi)]$.

If we ignore X_{it} and set $U_{it} = (Y_{it}(1), Y_{it}(0))$, which mechanically satisfies Assumption 3.3.1, this setup can be reduced to the finite population framework considered in Section 3.2. Importantly, Assumpion 3.3.1 does not imply unconditional or conditional parallel trends (on observed characteristics). This should come at no surprise given our results in Section 3.2; there, the inference of DATE is valid even if the trends of potential outcomes are arbitrarily heterogeneous across units.

3.3.2. The assignment and outcome models. To characterize double robustness, we first define two non-nested models. The assignment model is characterized by the generalized propensity score, defined as

$$\boldsymbol{\pi}_i(\boldsymbol{w}) = \mathbb{P}[\boldsymbol{W}_i = \boldsymbol{w} \mid \boldsymbol{X}_i, \boldsymbol{U}_i], \quad \forall \boldsymbol{w}_i \in \{0,1\}^T. \tag{3.3.3}$$

Given an estimate $\hat{\boldsymbol{\pi}}_i$, we say that it estimates the assignment model well if $\hat{\boldsymbol{\pi}}_i$ is close to $\boldsymbol{\pi}_i$ in total variation distance. Specifically, we define the accuracy of $\hat{\boldsymbol{\pi}}_i$ as

$$\delta_{\pi i} \triangleq \sqrt{\mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i))^2]}. \tag{3.3.4}$$

Clearly, $\delta_{\pi i} = 0$ if $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ on the support of $\boldsymbol{W}_i$.

In the absence of unobserved confounders $\boldsymbol{U}_i$, it is typical to estimate $\boldsymbol{\pi}_i$ via parametric or nonparametric regression of the treatment on the observed confounders. The accuracy $\delta_{\pi i}$ is then governed by the complexity of the ground truth, as well as the complexity of the function class used for estimation. With unobserved $\boldsymbol{U}_i$, it is generally impossible to get an accurate estimate of $\boldsymbol{\pi}_i$. However, it can be constructed under additional structural assumptions. For instance, suppose that $U_{it} \equiv U_i$ is a time-invariant confounder and $(W_{i1}, \ldots, W_{iT})$ are independent with

$$\operatorname{logit}(\mathbb{P}(W_{it} = 1 \mid X_{it}, \boldsymbol{U}_i)) = X_{it}^\top \beta + \gamma(\boldsymbol{U}_i). \tag{3.3.5}$$

The term $\gamma(\boldsymbol{U}_i)$ is essentially a fixed effect and cannot be estimated consistently when $T = O(1)$ since there are only a bounded number of observations available for this parameter. Nonetheless, we can enrich $\boldsymbol{X}_i$ by including an extra covariate $\overline{W}_i = (1/T)\sum_{t=1}^T W_{it}$. It is easy to demonstrate that

$\boldsymbol{W}_i \perp\!\!\!\perp U_i \mid \boldsymbol{X}_i, \overline{W}_i$ and

$$\boldsymbol{\pi}_i(\boldsymbol{w}) \propto \frac{\exp\left\{\sum_{t=1}^T w_t X_{it}^\top \beta\right\}}{\sum_{\boldsymbol{j}\in\{0,1\}^T : \bar{\boldsymbol{j}}=\overline{\boldsymbol{w}}} \exp\left\{\sum_{t=1}^T j_t X_{it}^\top \beta\right\}} \cdot I\left\{\overline{\boldsymbol{w}} = \overline{W}_i\right\}.$$

The coefficient vector β can be consistently estimated via the conditional logistic regression [McFadden, 1973]. Arkhangelsky and Imbens [2019] discuss various other models under which the unobserved confounders do not hinder accurate estimation.

The outcome model considered in this paper is a TWFE model. Specifically, the outcome model assumes that

$$\mathbb{E}[Y_{it}(w) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] = \alpha(\boldsymbol{U}_i) + \lambda_t + m(X_{it}, U_{it}) + \tau^* w. \tag{3.3.6}$$

In particular, this implies a constant treatment effect. When $T = O(1)$, the unit fixed effect $\alpha(\boldsymbol{U}_i)$ cannot be estimated consistently without further assumptions on $\alpha(\cdot)$ and $\boldsymbol{U}_i$, because there are only T samples that carry information on $\alpha(\boldsymbol{U}_i)$. Thus we cannot hope to estimate $\mathbb{E}[Y_{it}(w) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]$ consistently even with infinite sample sizes.

Let m_{it} denote the doubly-centered version of $\{\mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] : i \in [n], t \in [T]\}$, i.e.

$$\begin{aligned} m_{it} \triangleq\ & \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] - \frac{1}{T}\sum_{t=1}^T \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i] \\ & + \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]. \end{aligned} \tag{3.3.7}$$

When the outcome model (3.3.6) is correct, it is easy to see that m_{it} is also the doubly-centered version of terms $\{m(X_{it}, U_{it}) : i \in [n], t \in [T]\}$. Given an estimate $\hat{\mu}_{it}$ of $\mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]$, instead of requiring $\hat{\mu}_{it} - \mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]$ to be small, which is generally impossible when $T = O(1)$, we only require $\hat{m}_{it} \approx m_{it}$, where

$$\hat{m}_{it} \triangleq \hat{\mu}_{it} - \frac{1}{n}\sum_{i=1}^n \hat{\mu}_{it} - \frac{1}{T}\sum_{t=1}^T \hat{\mu}_{it} + \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T \hat{\mu}_{it}. \tag{3.3.8}$$

For notational convenience, we denote by $\boldsymbol{m}_i$ the vector $(m_{i1}, \ldots, m_{iT})$ and $\hat{\boldsymbol{m}}_i$ the vector $(\hat{m}_{i1}, \ldots, \hat{m}_{iT})$. Specifically, we say that the outcome model is correctly specified and estimated well by $\hat{\mu}_{it}$ if $\delta_{yi} \approx 0$,

where

$$\delta_{yi} \triangleq \sqrt{\mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i\|_2^2]} + \|\boldsymbol{\tau}_i - \tau^* \mathbf{1}_T\|_2. \tag{3.3.9}$$

For instance, $\delta_{yi} = 0$ if

$$\hat{\mu}_{it} = \tilde{\alpha}_i + \tilde{\lambda}_t + m_{it}, \quad \text{and} \quad \tau_{it} = \tau^*,$$

where $\tilde{\alpha}_i$ and $\tilde{\lambda}_t$ can be data-dependent and arbitrarily different from the true unit and time fixed effects. Under the classical linear TWFE model, i.e.

$$Y_{it} = \mu + \alpha(\boldsymbol{U}_i) + \lambda_t + X_{it}^\top \beta + \epsilon_{it}$$

where $\{\epsilon_{it} : i \in [n], t \in [T]\}$ are i.i.d. exogenous errors, the unweighted TWFE regression yields a consistent estimator of β even when $T = O(1)$ [e.g., Arellano, 2003]. Then $\delta_{yi} = \sqrt{\sum_{t=1}^T \{X_{it}^\top (\hat{\beta} - \beta)\}^2} \approx 0$ for $\hat{\mu}_{it} = \hat{\alpha}_i + \hat{\lambda}_t + X_{it}^T \hat{\beta}$.

3.3.3. Consistency of RIPW estimators. Given an estimate $\hat{\mu}_{it}$ for $\mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i, \boldsymbol{U}_i]$ and $\hat{\boldsymbol{\pi}}_i$ for $\boldsymbol{\pi}_i$, we consider the following RIPW estimator

$$\hat{\tau}(\boldsymbol{\Pi}) \triangleq \underset{\tau, \mu, \sum_i \alpha_i = \sum_t \gamma_t = 0}{\arg\min} \sum_{i=1}^n \sum_{t=1}^T ((Y_{it}^{\text{obs}} - \hat{m}_{it}) - \mu - \alpha_i - \gamma_t - W_{it}\tau)^2 \frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)}. \tag{3.3.10}$$

This is more general than the following weighted TWFE regression estimator with covariates

$$\hat{\tau} \triangleq \underset{\tau, \mu, \beta, \sum_i \alpha_i = \sum_t \gamma_t = 0}{\arg\min} \sum_{i=1}^n \sum_{t=1}^T (Y_{it}^{\text{obs}} - \mu - \alpha_i - \gamma_t - W_{it}\tau - X_{it}^\top \beta)^2 \frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)},$$

which is a special case of (3.3.10) with $\hat{m}_{it} = X_{it}^\top \hat{\beta}$. The two-stage estimator (3.3.10) is more flexible since it does not require $\hat{m}_{it}$ to be estimated from the same weighted regression for DATE. For instance, when $\boldsymbol{U}_i$ does not appear in $m(\boldsymbol{X}_i, \boldsymbol{U}_i)$, we could obtain a more efficient estimate of $\mathbb{E}[Y_{it}(0) \mid \boldsymbol{X}_i]$, or via an advanced estimation technique to handle complicated functional forms. On the other hand, the two-stage formulation replaces the regression with covariates by a regression on the modified outcome $(Y_{it}^{\text{obs}} - \hat{m}_{it})$ without covariates, yielding a simplified structure which allows us to use the results from the previous section.

To investigate the consistency of $\hat{\tau}$, we need extra assumptions. We start with the simplified case where $\{\hat{\boldsymbol{m}}_i : i \in [n], t \in [T]\}$ an $\{\hat{\boldsymbol{\pi}}_i : i \in [n]\}$ are independent of the data and thus can be treated

as fixed. We consider the modified versions of Assumptions 3.2.1 - 3.2.3 (see Appendix B.1.1 for a general version).

Assumption 3.3.2. *There exists a universal constant $c > 0$ and a non-stochastic subset $\mathbb{S}^* \subset \{0,1\}^T$ with at least two elements and at least one element not in $\{\mathbf{0}_T, \mathbf{1}_T\}$, such that*

$$\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) > c, \boldsymbol{\pi}_i(\boldsymbol{w}) > c \quad \forall \boldsymbol{w} \in \mathbb{S}^*, i \in [n], \quad \textit{almost surely},$$

Assumption 3.3.3. $\mathcal{Z}_i = (\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{U}_i)$ *are i.i.d..*

Assumption 3.3.4. *There exists $M < \infty$ such that* $\max_{i,t,w} |Y_{it}(w) - \hat{m}_{it}| < M$.

Theorem 3.2.1 implies that the RIPW estimator with $\mathbf{\Pi}$ being a solution of the DATE equation, if any, is a consistent estimator of DATE without any outcome model when $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ is known. On the other hand, when the outcome model is correct, $\hat{\tau}$ should be intuitively consistent if all RIPWs are well-behaved, because $Y_{it}^{\text{obs}} - \hat{m}_{it}$ is a linear model with two-way fixed effects and a single predictor W_{it} and $\hat{\tau}$ is a general least squares estimator which is consistent under mild conditions on the weights [e.g., Wooldridge, 2010]. This shows a weak double robustness property that $\hat{\tau}(\mathbf{\Pi})$ is consistent if either the outcome model or the assignment model is exactly correct. The weak double robustness has been studied for other causal estimands for panel data under different assumptions [e.g., Arkhangelsky and Imbens, 2019, Sant'Anna and Zhao, 2020].

For cross-sectional data, the augmented IPW estimator enjoys a strong double robustness property, which states that the asymptotic bias is the product of estimation errors of the outcome and assignment models [e.g., Robins et al., 1994, Kang and Schafer, 2007]. Clearly, this implies the weak double robustness. It further implies the estimator has higher asymptotic precision than estimators based on merely the outcome or assignment modelling, when both models are estimated well. Next result provides a sufficient condition for strong double robustness of $\hat{\tau}(\mathbf{\Pi})$ when the estimated treatment and outcome models are independent of the data.

Theorem 3.3.1. *Assume that $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in [n]\}$ are independent of the data. Under Assumptions 3.3.1 - 3.3.4, $\hat{\tau}(\mathbf{\Pi})$ is a consistent estimator of $\tau^*(\xi)$ (conditional on the estimates) if*

$$\bar{\delta}_\pi \bar{\delta}_y = o(1), \quad \textit{where } \bar{\delta}_\pi = \sqrt{\frac{1}{n}\sum_{i=1}^n \delta_{\pi i}^2}, \quad \bar{\delta}_y = \sqrt{\frac{1}{n}\sum_{i=1}^n \delta_{yi}^2}.$$

3.3.4. Doubly robust inference. Similar to Theorem 3.2.2, we can derive an asymptotic linear expansion for $\mathcal{D}(\hat{\tau}(\mathbf{\Pi}) - \tau^*(\xi))$ when $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in [n]\}$ are independent of the data.

THEOREM 3.3.2. *Let $\Gamma_\theta, \Gamma_{ww}, \mathbf{\Gamma}_w$, and $\mathcal{D}$ be defined as in Theorem 3.2.2. Redefine $\Gamma_{wy}, \mathbf{\Gamma}_y$, and $\mathcal{V}_i$ by replacing $\boldsymbol{Y}_i^{\text{obs}}$ with $\tilde{\boldsymbol{Y}}_i^{\text{obs}} = \boldsymbol{Y}_i^{\text{obs}} - \hat{\boldsymbol{m}}_i$. Under Assumptions 3.3.1 - 3.3.4 and that $\bar{\delta}_\pi \bar{\delta}_y = o(1/\sqrt{n})$,*

$$\mathcal{D} \cdot \sqrt{n}(\hat{\tau}(\mathbf{\Pi}) - \tau^*(\xi)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1).$$

As with the design-based inference, we can estimate the asymptotic variance via (3.2.19) and construct the Wald-type confidence interval as (3.2.20).

THEOREM 3.3.3. *Under the same settings as in Theorem 3.3.2,*

$$\liminf_{n \to \infty} \mathbb{P}\left(\tau^*(\xi) \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha,$$

if, further, (3.2.21) *holds.*

In practice, it is uncommon to obtain estimates of $\hat{\boldsymbol{\pi}}_i$ and $\hat{\boldsymbol{m}}_i$ that are independent of the data, except in the design-based inference where $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ and $\hat{\boldsymbol{m}}_i = 0$, or when external data is available. Usually, both parameters need to be estimated from the data. The resulting dependence invalidates the assumptions of Theorem 3.3.2 and 3.3.3. Intuitively, $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ cannot depend on the data arbitrarily because the double-dipping may inflate the Type-I error.

To salvage the situation, we apply the cross fitting technique to restrict the dependency structure. Specifically, we split the data into K almost equal-sized folds with $\mathcal{I}_k$ denoting the index sets of the k-th fold and $|\mathcal{I}_k| \in \{\lfloor n/K \rfloor, \lceil n/K \rceil\}$. For each $i \in \mathcal{I}_k$, we estimate $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ using $\{\mathcal{Z}_i : i \notin \mathcal{I}_k\}$. Since $\{\mathcal{Z}_i : i \in [n]\}$ are independent under Assumption 3.3.3, it is obvious that

$$\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in \mathcal{I}_k\} \perp\!\!\!\perp \{\mathcal{Z}_i : i \in \mathcal{I}_k\}.$$

For valid inference, we need an additional assumption on the stability of the estimates.

ASSUMPTION 3.3.5. *There exist functions $\{\boldsymbol{\pi}_i' : i \in [n]\}$ which satisfy Assumption 3.3.2, and vectors $\{\boldsymbol{m}_i' : i \in [n]\}$ which satisfy Assumption 3.3.4, such that they only depend on $\{\boldsymbol{X}_i : i \in [n]\}$ and*

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2] + \mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2] \right\} = O(n^{-r}) \tag{3.3.11}$$

for some $r > 0$. Furthermore,

$$\boldsymbol{\pi}_i' = \boldsymbol{\pi}_i \text{ for all } i, \quad \text{or} \quad \boldsymbol{m}_i' = \boldsymbol{m}_i \text{ for all } i. \tag{3.3.12}$$

The condition (3.3.11) states that the estimates need to be asymptotically deterministic given the confounders $(\boldsymbol{X}_i, \boldsymbol{U}_i)$. This is a very mild assumption. For example, when $\hat{\boldsymbol{\pi}}_i$ is estimated from a parametric model $\{f(\boldsymbol{X}_i;\theta) : \theta \in \mathbb{R}^d\}$ as $f(\boldsymbol{X}_i;\hat{\theta})$, under standard regularity conditions, $\hat{\theta}$ converges to a limit θ_0 even if the model is misspecified. As a result, $\hat{\boldsymbol{\pi}}_i$ converges to $\boldsymbol{\pi}_i' = f(\boldsymbol{X}_i;\theta_0)$. Under certain smoothness assumption, the estimates converge in the standard parametric rate and thus (3.3.11) holds with $r = 1$. On the other hand, for design-based inference, (3.3.11) is always satisfied with $\boldsymbol{\pi}_i' = \boldsymbol{\pi}_i$ and $\boldsymbol{m}_i' = 0$. More generally, if $\bar{\delta}_\pi^2 + \bar{\delta}_y^2 = O(n^{-r})$, it is also satisfied with $\boldsymbol{\pi}_i' = \boldsymbol{\pi}_i$ and $\boldsymbol{m}_i' = \boldsymbol{m}_i$. A similar assumption was considered for cross-sectional data by Chernozhukov et al. [2020].

The condition (3.3.12) allows one of the treatment and outcome models to be inconsistently estimated. This covers the design-based inference where the outcome model does not need to be consistently estimated. It also covers the classical model-based inference in which case the assignment model can be arbitrarily misspecified.

Theorem 3.3.4. *Let $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in [n]\}$ be estimates obtained from K-fold cross-fitting where $K = O(1)$. Under Assumptions 3.3.1 - 3.3.5,*

(i) $\hat{\tau}(\boldsymbol{\Pi}) - \tau^*(\xi) = o_{\mathbb{P}}(1)$ if $\bar{\delta}_\pi \bar{\delta}_y = o(1)$;

(ii) Let $\hat{C}_{1-\alpha}$ be the same confidence interval as in Theorem 3.3.3. Then

$$\liminf_{n\to\infty} \mathbb{P}\left(\tau^*(\xi) \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha$$

if (a) $\bar{\delta}_\pi \bar{\delta}_y = o(1/\sqrt{n})$, (b) Assumption 3.3.5 holds with $r > 1/2$, and (c) (3.2.21) holds if $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ are replaced by $(\boldsymbol{\pi}_i', \boldsymbol{m}_i')$ in the definition of $\mathcal{V}_i$.

3.4. Solutions of the DATE equation

3.4.1. The case of two periods. When there are two periods, the DATE equation only involves four variables $\boldsymbol{\Pi}(0,0), \boldsymbol{\Pi}(0,1), \boldsymbol{\Pi}(1,0), \boldsymbol{\Pi}(1,1)$. Through some tedious algebra presented in Appendix

B.2.1, we can show that the DATE equation can be simplified into the following equation:

$$\{\mathbf{\Pi}(1,1)-\mathbf{\Pi}(0,0)\}\{\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1)\}=(\xi_1-\xi_2)\Big\{(\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))^2-(\mathbf{\Pi}(1,0)+\mathbf{\Pi}(0,1))\Big\}. \tag{3.4.1}$$

3.4.1.1. *Difference-in-difference designs.* In the setting of difference-in-difference (DiD), $(0,0)$ and $(0,1)$ are the only two possible treatment assignments. As a result, we should set the support of the reshaped distribution to be $\mathbb{S}^*=\{(0,0),(0,1)\}$. Then (3.4.1) reduces to

$$\mathbf{\Pi}(0,0)\mathbf{\Pi}(0,1)=(\xi_1-\xi_2)(\mathbf{\Pi}(0,1)^2-\mathbf{\Pi}(0,1))=(\xi_2-\xi_1)\mathbf{\Pi}(0,0)\mathbf{\Pi}(0,1).$$

It has a solution only when $\xi_2-\xi_1=1$, i.e. $(\xi_1,\xi_2)=(0,1)$ and hence $\tau^*(\xi)=\tau_2$, in which case any reshaped distribution $\mathbf{\Pi}$ with $\mathbf{\Pi}(0,0),\mathbf{\Pi}(0,1)>0$ is a solution. This is not surprising because for DiD, no unit is treated in the first period and thus τ_1 is unidentifiable. Nonetheless, τ_2 is an informative causal estimand in the literature of DiD. This implies that the RIPW estimator with any $\mathbf{\Pi}$ with $\mathbf{\Pi}(0,0),\mathbf{\Pi}(0,1)>0$ and $\mathbf{\Pi}(0,0)+\mathbf{\Pi}(0,1)=1$ yields a doubly robust DiD estimator.

3.4.1.2. *Cross-over designs.* For a two-period cross-over design, $(0,1)$ and $(1,0)$ are the only two possible treatment assignments. Since the support of $\mathbf{\Pi}$ must contain at least two elements, it has to be $\mathbb{S}^*=\{(1,0),(0,1)\}$. Then DATE equation reduces to

$$0=(\xi_1-\xi_2)\Big\{(\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))^2-(\mathbf{\Pi}(1,0)+\mathbf{\Pi}(0,1))\Big\}.$$

When $\xi_1\neq\xi_2$, it implies that

$$0=(\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))^2-(\mathbf{\Pi}(1,0)+\mathbf{\Pi}(0,1))=(\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))^2-1.$$

It never holds since $\mathbf{\Pi}(1,0),\mathbf{\Pi}(0,1)>0$. By contrast, when $\xi_1=\xi_2=1/2$, any $\mathbf{\Pi}$ with support $(1,0)$ and $(0,1)$ is a solution.

3.4.1.3. *Estimating equally-weighted DATE for general designs.* When $\xi_1=\xi_2=1/2$, the DATE equation reduces to

$$\{\mathbf{\Pi}(1,1)-\mathbf{\Pi}(0,0)\}\{\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1)\}=0\Longleftrightarrow\mathbf{\Pi}(1,1)=\mathbf{\Pi}(0,0)\text{ or }\mathbf{\Pi}(1,0)=\mathbf{\Pi}(0,1).$$

If $\mathbb{S}^* = \{(1,1),(0,0),(1,0),(0,1)\}$ in Assumption 3.3.2, that is, when all combinations of treatments are possible, the solutions are

$$(\mathbf{\Pi}(1,1),\mathbf{\Pi}(0,0),\mathbf{\Pi}(0,1),\mathbf{\Pi}(1,0)) = (a,a,b,1-2a-b), \quad a>0, 2a+b<1$$
$$\text{or } (\mathbf{\Pi}(1,1),\mathbf{\Pi}(0,0),\mathbf{\Pi}(0,1),\mathbf{\Pi}(1,0)) = (a,1-a-2b,b,b), \quad b>0, a+2b<1.$$

The uniform distribution on $\mathbb{S}^*$ is a solution, implying that the IPW weights deliver the average effect in this case. If $\mathbb{S}^* = \{(1,1),(0,0),(0,1)\}$ (staggered adoption), we cannot make $\mathbf{\Pi}(1,0)$ and $\mathbf{\Pi}(0,1)$ equal since the former must be zero while the latter must be positive. Therefore, the solutions can be characterized as

$$(\mathbf{\Pi}(1,1),\mathbf{\Pi}(0,0),\mathbf{\Pi}(0,1)) = (a,a,1-2a), \quad a\in(0,1/2). \tag{3.4.2}$$

Again, the uniform distribution on $\mathbb{S}^*$ is a solution. However, we will show in the next section that the uniform distribution is not a solution for staggered adoption designs with $T \geq 3$.

3.4.2. Staggered adoption with multiple periods. For staggered adoption designs, $\boldsymbol{\pi}_i$ is supported on

$$\mathcal{W}_T^{\text{sta}} \triangleq \{\boldsymbol{w} : \boldsymbol{w}_1 = \ldots = \boldsymbol{w}_i = 0, \boldsymbol{w}_{i+1} = \ldots = \boldsymbol{w}_T = 1 \text{ for some } i = 0,1,\ldots,T\}.$$

For notational convenience, we denote by $\boldsymbol{w}_{(j)}$ the vector in $\mathcal{W}_T^{\text{sta}}$ with j entries equal to 1 for $j = 0,1,\ldots,T$. Thus, the support $\mathbb{S}^*$ of $\mathbf{\Pi}$ must be a subset of $\mathcal{W}_T^{\text{sta}}$. For general weights, the DATE equation is a quadratic system with complicated structures. Nonetheless, when $\xi_1 = \ldots = \xi_T = 1/T$, the solution set is an union of segments on the T-dimensional simplex with closed-form expressions. We focus on the equally-weighted DATE in this section.

Theorem 3.4.1. *Let $\mathbb{S}^* = \{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(j_1)},\ldots,\boldsymbol{w}_{(j_r)},\boldsymbol{w}_{(T)}\}$ with $1 \leq j_1 < \ldots < j_r \leq T-1$. Then the set of solutions of the DATE equation with support $\mathbb{S}^*$ is characterized by the following linear system:*

$$\begin{cases} \mathbf{\Pi}(\boldsymbol{w}_{(T)}) = \frac{T-j_r}{T} - \mathbf{\Pi}(\boldsymbol{w}_{(j_r)}) + \frac{1}{T}\sum_{k=1}^{r} j_k \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}) \\ \mathbf{\Pi}(\boldsymbol{w}_{(j_{k+1})}) + \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}) = \frac{j_{k+1}-j_k}{T}, \quad k = 1,\ldots,r-1 \\ \mathbf{\Pi}(\boldsymbol{w}_{(0)}) = 1 - \mathbf{\Pi}(\boldsymbol{w}_{(T)}) - \sum_{k=1}^{r} \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}) \\ \mathbf{\Pi}(\boldsymbol{w}) > 0 \textit{ iff } \boldsymbol{w} \in \mathbb{S}^* \end{cases} \tag{3.4.3}$$

Furthermore, the solution set of (3.4.3) *is either an empty set or a* 1*-dimensional segment in the form of* $\{\lambda\mathbf{\Pi}^{(1)} + (1-\lambda)\mathbf{\Pi}^{(2)} : \lambda \in (0,1)\}$ *for some distributions* $\mathbf{\Pi}^{(1)}$ *and* $\mathbf{\Pi}^{(2)}$.

The proof of Theorem 3.4.1 is presented in Appendix B.2.2. In the following corollary, we show that the solution set with $\mathbb{S}^* = \mathcal{W}_T^{\text{sta}}$ is always non-empty with nice explicit expressions.

Corollary 3.4.1. *When* $\mathbb{S}^* = \mathcal{W}_T^{\text{sta}}$*, the solution set of* (3.4.3) *is* $\{\lambda\mathbf{\Pi}^{(1)} + (1-\lambda)\mathbf{\Pi}^{(2)} : \lambda \in (0,1)\}$ *where*

- *if* T *is odd,*

$$\mathbf{\Pi}^{(1)}(\boldsymbol{w}_{(T)}) = \frac{(T+1)^2}{4T^2},\ \mathbf{\Pi}^{(1)}(\boldsymbol{w}_{(0)}) = \frac{T^2-1}{4T^2},\ \mathbf{\Pi}^{(1)}(\boldsymbol{w}_j) = \frac{I(j \text{ is odd})}{T},\ j = 1,\dots,T-1,$$

$$\text{and}\ \ \mathbf{\Pi}^{(2)}(\boldsymbol{w}_{(j)}) = \mathbf{\Pi}^{(1)}(\boldsymbol{w}_{(T-j)}),\ j = 0,\dots,T;$$

- *if* T *is even,*

$$\mathbf{\Pi}^{(1)}(\boldsymbol{w}_{(T)}) = \mathbf{\Pi}^{(1)}(\boldsymbol{w}_{(0)}) = \frac{1}{4},\ \mathbf{\Pi}^{(1)}(\boldsymbol{w}_j) = \frac{I(j \text{ is odd})}{T},\ j = 1,\dots,T-1,$$

$$\text{and}\ \ \mathbf{\Pi}^{(2)}(\boldsymbol{w}_{(T)}) = \mathbf{\Pi}^{(2)}(\boldsymbol{w}_{(2)}) = \frac{T+2}{4T},\ \mathbf{\Pi}^{(2)}(\boldsymbol{w}_j) = \frac{I(j \text{ is even})}{T},\ j = 1,\dots,T-1.$$

In particular, when $T = 3$ and $\mathbb{S}^* = \mathcal{W}_T^{\text{sta}}$, the solution set is

$$\left\{(\mathbf{\Pi}(\boldsymbol{w}_{(0)}), \mathbf{\Pi}(\boldsymbol{w}_{(1)}), \mathbf{\Pi}(\boldsymbol{w}_{(2)}), \mathbf{\Pi}(\boldsymbol{w}_{(3)}) = \lambda\left(\frac{2}{9}, \frac{1}{3}, 0, \frac{4}{9}\right) + (1-\lambda)\left(\frac{4}{9}, 0, \frac{1}{3}, \frac{2}{9}\right) : \lambda \in (0,1)\right\}. \tag{3.4.4}$$

Clearly, the uniform distribution on $\mathbb{S}^*$ is excluded. Thus, although the RIPW estimator with a uniform reshaped distribution is inconsistent, the non-uniform distribution $(1/3, 1/6, 1/6, 1/3)$, namely the midpoint of the solution set, induces a consistent RIPW estimator. For general T, it is easy to see that the midpoint is

$$\mathbf{\Pi}(\boldsymbol{w}_{(T)}) = \mathbf{\Pi}(\boldsymbol{w}_{(0)}) = \frac{T+1}{4T},\ \mathbf{\Pi}(\boldsymbol{w}_{(j)}) = \frac{1}{2T},\ j = 1,\dots,T-1. \tag{3.4.5}$$

This distribution uniformly assigns probabilities on the subset $\{\boldsymbol{w}_{(1)},\dots,\boldsymbol{w}_{(T-1)}\}$ while puts a large mass on $\{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(T)}\}$. Intuitively, the asymmetry is driven by the special roles of $\boldsymbol{w}_{(0)}$ and $\boldsymbol{w}_{(T)}$: the former provides the only control group for period T while the latter provides the only treated group for period 1.

Corollary 3.4.1 offers a unified recipe for the reshaped distribution when the positivity Assumption 3.3.2 holds for all possible assignments. In some applications, certain assignment never or rarely occurs and we are forced to restrict the support of $\mathbf{\Pi}$ into a smaller subset $\mathbb{S}^*$. To start with, we provide a detailed account of the case $T = 3$. When $j_1 = 1, j_2 = 2$, (3.4.4) shows that $\mathbf{\Pi}(\boldsymbol{w}_{(0)}), \mathbf{\Pi}(\boldsymbol{w}_{(3)}) > 0$, and thus $\mathbb{S}^*$ must be $\mathcal{W}^3$ and cannot be $\{\boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}\}$, $\{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}\}$, or $\{\boldsymbol{w}_{(1)}, \boldsymbol{w}_{(2)}, \boldsymbol{w}_{(3)}\}$. When $j_1 = 1, r = 1$, via some tedious algebra, the solution set of (3.4.3) is

$$\left\{(\mathbf{\Pi}(\boldsymbol{w}_{(0)}), \mathbf{\Pi}(\boldsymbol{w}_{(1)}), \mathbf{\Pi}(\boldsymbol{w}_{(2)}), \mathbf{\Pi}(\boldsymbol{w}_{(3)}) = \lambda\,(0,1,0,0) + (1-\lambda)\left(\frac{1}{3}, 0, 0, \frac{2}{3}\right) : \lambda \in (0,1)\right\}. \tag{3.4.6}$$

Thus, $\{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(1)}, \boldsymbol{w}_{(3)}\}$ is the only support with $j_1 = 1, r = 1$ that induces a non-empty solution set of (3.4.3). Similarly, we can show that the only support with $j_2 = 1, r = 1$ that induces a non-empty solution set as

$$\left\{(\mathbf{\Pi}(\boldsymbol{w}_{(0)}), \mathbf{\Pi}(\boldsymbol{w}_{(1)}), \mathbf{\Pi}(\boldsymbol{w}_{(2)}), \mathbf{\Pi}(\boldsymbol{w}_{(3)}) = \lambda\,(0,0,1,0) + (1-\lambda)\left(\frac{2}{3}, 0, 0, \frac{1}{3}\right) : \lambda \in (0,1)\right\}. \tag{3.4.7}$$

In sum, $\mathcal{W}_T^{\text{sta}}, \mathcal{W}_T^{\text{sta}} \setminus \{\boldsymbol{w}_{(1)}\}, \mathcal{W}_T^{\text{sta}} \setminus \{\boldsymbol{w}_{(2)}\}$ are the only three supports with non-empty solution sets, characterized by (3.4.4), (3.4.6), and (3.4.7), respectively.

For $T = 3$, $\{j_1, \ldots, j_r\}$ can be any non-empty subset of $\{1, 2\}$. Via some tedious algebra, we can show that this continues to be true for $T = 4$. However, this no longer holds for $T \geq 5$. For instance, if $\{j_1, \ldots, j_r\} = \{1, 2, 4, 5\}$, the second equation of (3.4.3) implies that

$$\mathbf{\Pi}(\boldsymbol{w}_{(1)}) + \mathbf{\Pi}(\boldsymbol{w}_{(2)}) = \mathbf{\Pi}(\boldsymbol{w}_{(4)}) + \mathbf{\Pi}(\boldsymbol{w}_{(5)}) = \frac{1}{T}, \quad \mathbf{\Pi}(\boldsymbol{w}_{(2)}) + \mathbf{\Pi}(\boldsymbol{w}_{(4)}) = \frac{2}{T}.$$

Under the support constraint, the first two equations imply that $\mathbf{\Pi}(\boldsymbol{w}_{(2)}), \mathbf{\Pi}(\boldsymbol{w}_{(5)}) < 1/T$, contradicting with the third equation. Nonetheless, the contradiction can be resolved if any of these four elements is discarded. If this is the case in practice, we can discard the element that is believed to be the least likely assignment.

3.4.3. Other designs. In many applications, the treatment can be switched on and off at different periods for a single unit. In general, a design is characterized by a collection of possible assignments $\mathbb{S}_{\text{design}}$. If any subset $\mathbb{S}^* \subset \mathbb{S}_{\text{design}}$ yields a non-empty solution set of the DATE equation, we can derive a doubly robust estimator of the DATE. In this section, we consider several designs with more than two periods which are not staggered adoption designs.

First we consider transient designs with zero or one period being treated and with each period being treated with a non-zero chance, i.e.,

$$\mathcal{W}_{T,1}^{\mathrm{tra}} = \left\{ \boldsymbol{w} \in \{0,1\}^T : \sum_{t=1}^{T} \boldsymbol{w}_t \le 1 \right\}.$$

For notational convenience, we denote by $\tilde{\boldsymbol{w}}_{(0)}$ the never-treated assignment and $\tilde{\boldsymbol{w}}_{(j)}$ the assignment with only j-th period treated. The above design can be encountered, for example, when the treatment is a natural disaster. The following theorem characterizes all solutions of the DATE equation for any ξ.

Theorem 3.4.2. *When $\mathbb{S}^* = \mathcal{W}_{T,1}^{\mathrm{tra}}$, $\boldsymbol{\Pi}$ is a solution of the DATE equation iff there exists $b > 0$ such that*

$$\boldsymbol{\Pi}(\tilde{\boldsymbol{w}}_{(t)}) \left\{ 1 - \boldsymbol{\Pi}(\tilde{\boldsymbol{w}}_{(t)}) - \frac{\boldsymbol{\Pi}(\tilde{\boldsymbol{w}}_{(0)})}{T} \right\} = \xi_t b, \quad \forall t \in [T].$$

In particular, when $\xi_t = 1/T$ for every t, Theorem 3.4.2 implies that $\boldsymbol{\Pi} \sim \mathrm{Unif}(\mathcal{W}_{T,1}^{\mathrm{tra}})$ is a solution. In fact, for any given $\boldsymbol{\Pi}(\tilde{\boldsymbol{w}}_0) \in (0,1)$, $\boldsymbol{\Pi}$ is a solution if

$$\boldsymbol{\Pi}(\cdot \mid \boldsymbol{w} \neq \tilde{\boldsymbol{w}}_{(0)}) \sim \mathrm{Unif}(\{\tilde{\boldsymbol{w}}_{(1)}, \ldots, \tilde{\boldsymbol{w}}_{(T)}\}).$$

The above decomposition can be used to construct solutions for more general transient designs:

$$\mathcal{W}_{T,k}^{\mathrm{tra}} = \left\{ \boldsymbol{w} \in \{0,1\}^T : \sum_{t=1}^{T} \boldsymbol{w}_t \le k \right\}.$$

This design is common in marketing experiments where, for example, k is the maximal number of coupons given to a user and each user can receive coupons in any combination of up to k time periods.

Theorem 3.4.3. *When $\mathbb{S}^* = \mathcal{W}_{T,1}^{\mathrm{tra}}$, $\boldsymbol{\Pi}$ is a solution of the DATE equation with $\xi_t = 1/T$ $(t = 1, \ldots, T)$, if*

$$\boldsymbol{\Pi}\left(\cdot \mid \sum_{t=1}^{T} \boldsymbol{w}_t = k' \right) \sim \mathrm{Unif}(\mathcal{W}_{T,k'}^{\mathrm{tra}} \setminus \mathcal{W}_{T,k'-1}^{\mathrm{tra}}), \quad k' = 1, \ldots, k,$$

3.5. Numerical Studies

In this section, we investigate the properties of our estimator in simulations and show how to apply it to real datasets. The R programs to replicate all results in this section is available at `https://github.com/xiaomanluo/ripwPaper`.

3.5.1. Synthetic data. To highlight the central role of the reshaping function in eliminating the bias, we focus on design-based inference, where the propensity scores are known for every unit, with a large sample size to avoid finite sample bias. Put another way, in such settings, the bias of the unweighted or IPW estimators is purely driven by the wrong reshaping function, rather than other sources of variability. For simplicity, we consider the DATE with $\xi = \mathbf{1}_T/T$.

We consider a short panel with $T = 4$ and sample size $n = 10000$. We generate a single time-invariant covariate $X_{it} = X_i$ with $P(X_i = 1) = 0.7$ and $P(X_i = 2) = 0.3$ and a single time-invariant unobserved confounder $U_{it} = U_i$ with $U_i \sim \text{Unif}(\{1, \ldots, 10\})$. Within each experiment, the covariates and unobserved confounders are only generated once and then fixed to ensure a fixed design. For treatment assignments, we consider a staggered adoption design, i.e., $\boldsymbol{W}_i \in \mathcal{W}^{\text{sta}}$. We assume that $\boldsymbol{W}_i$ is less likely to be treated when $X_i = 1$. In particular,

$$\Big(\boldsymbol{\pi}_i(\boldsymbol{w}_{(0)}), \boldsymbol{\pi}_i(\boldsymbol{w}_{(1)}), \boldsymbol{\pi}_i(\boldsymbol{w}_{(2)}), \boldsymbol{\pi}_i(\boldsymbol{w}_{(3)}), \boldsymbol{\pi}_i(\boldsymbol{w}_{(4)})\Big) = \begin{cases} (0.8, 0.05, 0.05, 0.05, 0.05) & (X_i = 1) \\ (0.1, 0.1, 0.2, 0.3, 0.3) & (X_i = 2) \end{cases}.$$

The potential outcome $Y_{it}(0)$ and the treatment effect τ_{it} are generated as follows:

$$Y_{it}(0) = \mu + \alpha_i + \gamma_t + m_{it} + \epsilon_{it}, \quad m_{it} = \sigma_m X_i \beta_t, \quad \tau_{it} = \sigma_\tau a_i b_t,$$

where $\mu = 0$, $\beta_t = t - 1$, $\alpha_i = 0.5U_i$, $\gamma_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $b_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $\epsilon_{it} \overset{i.i.d.}{\sim} N(0,1)$. For a_i, we consider two settings: we either set $a_i = 1$ thus making τ_{it} unit-invariant; or $a_i \overset{i.i.d.}{\sim} \text{Unif}([0,1])$, in which case τ_{it} varies over units and periods. As with the covariates X_i, the time fixed effects γ_t and factors a_i, b_t are generated once and then fixed over runs. In contrast, ϵ_{it} will be resampled in every run as the stochastic errors. Note that both m_{it} and τ_{it} are generated from rank-one factor models.

The parameters σ_m and σ_τ measures two types of deviations from the TWFE model: σ_m measures the violation of parallel trend because we will not adjust for X_i in the design-based inference, and σ_τ measures the violation of constant treatment effects. We consider two settings: we either set $\sigma_m = 1, \sigma_\tau = 0$ — a model without parallel trends, but constant treatment effects;

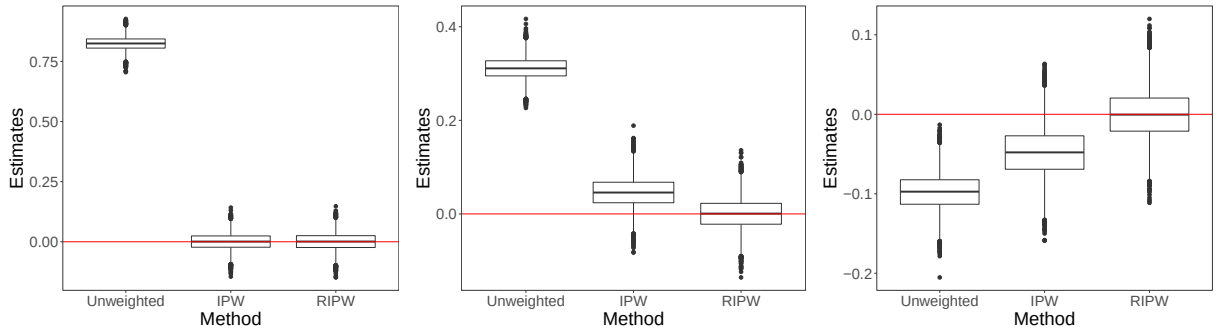


FIGURE 3.3. Boxplots of bias for the unweighted, IPW, and RIPW estimators
Boxplots of bias across 1000 replicates for the unweighted, IPW, and RIPW estimators under (left) violation of parallel trend ($\sigma_m = 1, \sigma_\tau = 0$), (middle) heterogeneous treatment effect with limited heterogeneity ($\sigma_m = 0, \sigma_\tau = 1, a_i = 1$), and (right) heterogeneous treatment effect with full heterogeneity ($\sigma_m = 0, \sigma_\tau = 1, a_i \sim \text{Unif}([0,1])$).

alternatively, we set $\sigma_m = 0, \sigma_\tau = 1$ — a TWFE model with heterogeneous effects, but parallel trends. In the first setting $\tau_{it} = 0$ regardless of the model for a_i, thus, we have 3 different scenarios in total.

We consider three estimators: the unweighted TWFE estimator, the IPW estimator, and the RIPW estimator with $\mathbf{\Pi}$ given by (3.4.5). For each of the three experiments, we resample W_{it}'s and ϵ_{it}'s, while keeping other quantities fixed, for 1000 times and collect the estimates and the confidence intervals. Figure 3.3 presents the boxplots of the bias $\hat{\tau}(\mathbf{\Pi}) - \tau^*(\xi)$. In all settings, the unweighted estimator is clearly biased, demonstrating that both the parallel trend and treatment effect homogeneity are indispensible for classical TWFE regression. In contrast, the IPW estimator is biased when the treatment effects are heterogeneous, but unbiased otherwise even if the parallel trend assumption is violated. This is by no means a coincidence; in this case, $\boldsymbol{\tau}_i = \tau^*(\xi)\mathbf{1}_T$ for all i and, by Theorem 3.2.1, the asymptotic bias $\Delta_\tau(\xi) = 0$ for RIPW estimators with any reshaped function including the IPW estimator. Finally, as implied by our theory, the RIPW estimator is unbiased in all settings. Moreover, the coverage of confidence intervals for the RIPW estimator is 95.1%, 95.0%, and 94.8% in these three settings, respectively, confirming the inferential validity stated in Theorem 3.2.3.

3.5.2. Reanalysis of Bachhuber et al. (2014) on medical cannabis law. In 1996, California voters first passed the law that legalized the medical usage of medical cannabis. By the end of 2017, 43 more states have passed similar laws. As more states pass the medical cannabis law, there has been debate on whether legal medical marijuana is associated with an increase or decrease in opioid overdose mortality. An influential paper by Bachhuber et al. [2014] analyzed the data

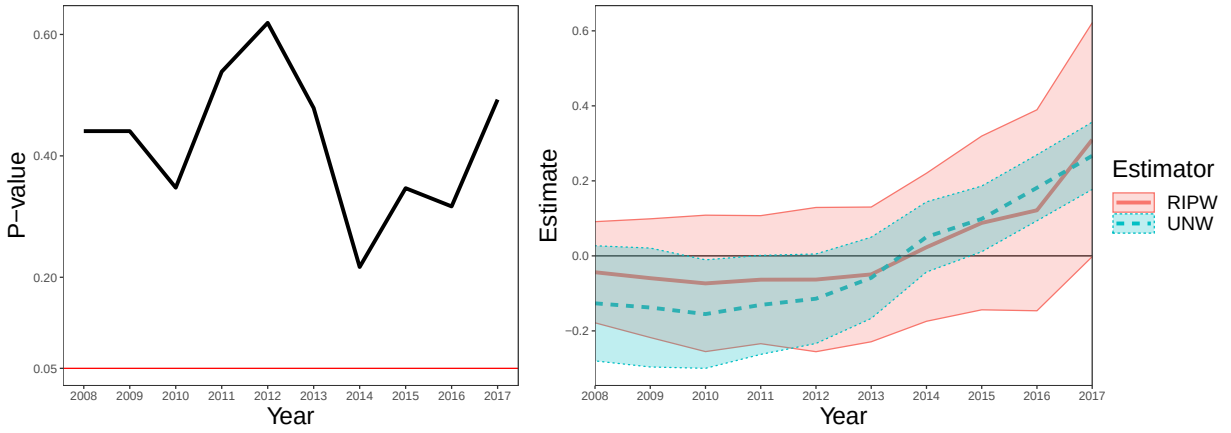
from 1999 to 2010 via the standard TWFE regression and found a significant negative effect on the state-level opioid overdose mortality rates. Later on, Shover et al. [2019] applied the same method on the data from 1999 to 2017 and found a significant positive effect instead, though they believe the association is spurious. Recently, Andrew Baker reanalyzed the data in his blog [1] using the modern DiD methods for staggered adoption, which is the design in this case since no state ever repeals the law, and raised concerns about the standard TWFE regression.

The treatment effect is highly heterogeneous in both states and time because of the complicated sociological and biological mechanisms through which the legal medical cannabis affects the opioid overdose mortality. On the other hand, the adoption time of the medical cannabis law involves a great amount of uncertainty, which is arguably easier to model than the mortality. For example, a duration model can be applied in this context. This suggests the potential benefit of our RIPW estimator which lends more robustness by leveraging the additional information from the adoption process, another important source of variation that might help with causal identification.

Following Bachhuber et al. [2014] and Shover et al. [2019], we use the logarithm of age-adjusted opioid overdose death rate per 100,000 population as the outcome, and include four time-varying covariates: annual state unemployment rate and presence of the following: prescription drug monitoring program, pain management clinic oversight laws, and law requiring or allowing pharmacists to request patient identification. As with Andrew Baker's blog, we remove North Dakota due to the high missing rate and impute the remaining missing values in the outcome, law adoption status, and unemployment using the matrix completion technique by Athey et al. [2018]. Since the previous contradicting finding occur at 2010 and 2017, we estimate the effect from 1999 to T_{end} for each $T_{\text{end}} \in \{2008, 2009, \ldots, 2017\}$. In particular, we choose the causal estimand as the equally-weighted DATE, which is close to the research question of Bachhuber et al. [2014] and Shover et al. [2019] in spirit.

For the RIPW estimator, we fit a standard TWFE regression to derive an estimate of the outcome model, i.e., $\hat{m}_{it} = X_{it}^{\top}\hat{\beta}$. This step guarantees that the resulting RIPW estimator is acceptable if the analyses of Bachhuber et al. [2014] and Shover et al. [2019] are because all estimate the same outcome model. On top of that, we fit a Cox proportional hazard model [Cox, 1972, Kalbfleisch and Prentice, 2011] with the same set of covariates to model the right-censored adoption time.

[1] https://andrewcbaker.netlify.app/2019/12/31/what-can-we-say-about-medical-marijuana-and-opioid-overdose-mortality/



(a) P-values of the Schoenfeld's test

(b) Effect estimates

FIGURE 3.4. Results for the data on medical cannabis law
(Left) Diagnostics for the Cox proportional hazard model on adoption times; (right) the unweighted TWFE regression estimates (UNW) and RIPW estimates. The x-axis represents T_{end}.

Specifically, letting T_i be the year in which the state i passes the medical cannabis law, a Cox proportional hazard model with time-varying covariates X_{it} assumes that

$$h_i(t \mid X_{it}) = h_0(t) \exp\{X_{it}^\top \beta\}$$

where $h_i(t \mid \cdot)$ denotes the hazard function for state i, and $h_0(t)$ denotea a nonparametric baseline hazard function. The estimates $\hat{h}_0$ and $\hat{\beta}$ yield an estimate $\hat{F}_i(t)$ of the survival function $\mathbb{P}(T_i \geq t)$ for state i, differencing which yields an estimate generalized propensity score

$$\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) = \begin{cases} \hat{F}_i(T_i) - \hat{F}_i(T_i + 1) & \text{(State } i \text{ passed the law before 2017)} \\ 1 - \hat{F}_i(2017) & \text{(otherwise)} \end{cases}.$$

The reshaped distribution is chosen as the midpoint solution (3.4.5). Finally, we apply the standard 10-fold cross-fitting to derive the estimates of the outcome and treatment models.

The proportional hazard assumption imposed by the Cox model is often controversial. Here, we apply the standard statistical tests based on Schoenfeld residuals [Schoenfeld, 1980] as a specification test for the Cox model. Figure 3.4a presents the p-values yielded by the Schoenfeld's test for each T_{end} without data splitting. Clearly, none of them show evidence against the proportional hazard assumption.

Figure 3.4b presents the RIPW estimates of equally-weighted DATE and the unweighted TWFE regression estimates for $T_{\text{end}} \in \{2008, 2009, \ldots, 2017\}$. It also displays the 95% pointwise confidence

intervals; here, the cluster-robust standard error is used for the unweighted estimator. The point estimates of the RIPW estimator and the unweighted estimator are similar when $T_{\text{end}} \geq 2013$, though the RIPW estimates are closer to zero otherwise. Moreover, the unweighted estimator shows significant negative effect when $T_{\text{end}} = 2010$ and significant positive effect when $T_{\text{end}} \geq 2015$. In contrast, the RIPW estimator does not show any significant effect due to the larger standard error to adjust for effect heterogeneity. This result corroborates the suspicion of Shover et al. [2019] on the invalidity of the unweighted TWFE regression estimates.

3.5.3. Analysis of OpenTable data in the early COVID-19 pandemic. On February 29th, 2020, Washington declared a state of emergency in response to the COVID-19 pandemic. A state of emergency is a situation in which a government is empowered to perform actions or impose policies that it would normally not be permitted to undertake [2]. It alerts citizens to change their behaviors and urges government agencies to implement emergency plans. As the pandemic has swept across the country, more states declared the state of emergency in response to the COVID-19 outbreak.

The state of emergency restricts various human activities. It would be valuable for governments and policymakers to get a sense of the short-term effect of this urgent action. Since mid-Feburary of 2020, OpenTable has been releasing daily data of year-over-year seated diners for a sample of restaurants on the OpenTable network through online reservations, phone reservations, and walk-ins.[3] This provides an opportunity to study how the state of emergency affects the restaurant industry in a short time frame. The data covers 36 states in the United States, which we will focus our analysis on.

Policy evaluation in the pandemic is extremely challenging due to the complex confounding and endogeneity issues [e.g., Chetty et al., 2020, Chinazzi et al., 2020, Goodman-Bacon and Marcus, 2020, Holtz et al., 2020, Kraemer et al., 2020, Abouk and Heydari, 2021]. Fortunately, compared to the policies in the later stage of the pandemic, the state of emergency was less confounded since it was basically the first policy that affected the vast majority of the public. On the other hand, the restaurant industry is responding to the policy swiftly because the restaurants are forced to limit and change operations, thereby eliminating some confounders that cannot take effect in a few days.

[2]Definition from Wikipedia: `https://en.wikipedia.org/wiki/State_of_emergency`.
[3]Source: `https://www.opentable.com/state-of-industry`.

Despite being more approachable, the problem remains challenging due to the effect heterogeneity and the difficulty to build a reliable model for the dine-in rates in a short time window. In contrast, the declaration time of the state of emergency is arguably less complex to model because it is mainly drive by the progress of the pandemic and the authority's attitude towards the pandemic.

We demonstrate our RIPW estimator on this data. The outcome variable is the daily state-level year-over-year percentage change in seated diners provided by OpenTable. [4] The treatment variable is the indicator whether the state of emergency has been declared. [5] We also include the state-level accumulated confirmed cases to measure the progress of the pandemic, [6] the vote share of Democrats based on the 2016 presidential election data to measure the political attitude towards COVID-19, [7] and the number of hospital beds per-capita as a proxy for the amount of regular medical resources. [8] For demonstration purpose, we restrict the analysis into Feburary 29th – March 13th, the first 14 days since the first declaration by Washington. As of March 13th, 34 out of 36 states have declared the state of emergency, and thus the declaration times are slightly right-censored.

Analogous to Section 3.5.2, we fit a Cox proportional hazard model on the declaration date to derive an estimate of the generalized propensity scores. Here, we include as the covariates the logarithms of the accumulated confirmed cases and the number of hospital beds per-capita, and the vote share. The p-value of the Schoenfeld's test is 0.34, suggesting no evidence against the specification. For the outcome model, we fit a standard TWFE regression with the same set of covariates, as detailed in Section 3.5.2. With these estimates, we compute the RIPW estimator for equally-weighted DATE with the reshaped distribution (3.4.5) and 10-fold cross-fitting. The RIPW estimate is -4.01% with the 95% confidence interval $[-8.63\%, 0.61\%]$ and the 90% confidence interval $[-7.89\%, -0.13\%]$. Thus, the effect is negative but only significant at the 10% level. As a comparison, the unweighted TWFE regression estimate is -1.1% with the 95% confidence interval is $[-4.28\%, 2.09\%]$ and 90% confidence interval is $[-3.77\%, 1.58\%]$.

[4]Source: https://www.opentable.com/state-of-industry.
[5]Source: https://www.businessinsider.com/california-washington-state-of-emergency-coronavirus-what-it-means-2020-3.
[6]Source: https://coronavirus.jhu.edu/.
[7]Source: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ.
[8]Source: https://github.com/rbracco/covidcompare.

3.6. Conclusion

We demonstrate both theoretically and empirically that the unit-specific reweighting of the OLS objective function improves the robustness of the resulting treatment effects estimator in applications with panel data. The proposed weights are constructed using the assignment process (either known or estimated) and thus appropriate in situations with substantial cross-sectional variation in the treatment paths. Practically, our results allow applied researchers to exploit domain knowledge about outcomes and assignments, thus resulting in a more balanced approach to identification and estimation.

Our focus on a very particular OLS problem – two-way fixed effects regression – is motivated by its popularity in applied work. We believe that our results can be extended to more general models, including those with interactive fixed effects, and models with dynamic treatment effects and state dependence. We view this as a part of the broad research agenda that connects different aspects of the causal inference problem – assignments and outcomes – to build more robust and transparent estimators.

APPENDIX A

Long Title of Appendix A

A.1. Mathematical Details of The Two-agent Model

A.1.1. Child utility maximization. The utility of child is

$$\begin{aligned} \max_{s} \quad & u_1^k(s, c_1^k) + \beta_k u_2^k(c_2), \\ s.t. \quad & c_1^k = \gamma(d) W_p(d), \\ & c_2 = g(h), \\ & h = f(d, s, c_1^k, h_0). \end{aligned} \tag{A.1.1}$$

Plugging constraints to utility function

$$L^k = u_1^k(s, \gamma(d) W_p(d)) + \beta_k u_2^k(g(f(d, s, c_1^k, h_0)))$$

Taking the derivative with respect to s and obtain the first order condition

$$\frac{\partial L^k}{\partial s} = \frac{\partial u_1^k}{\partial s} + \beta_k \frac{\partial u_2^k}{\partial c_2} \frac{\partial g}{\partial h} \frac{\partial f}{\partial s} = 0.$$

The marginal effect of studying time on current utility is $\mathrm{MU}_1^k = -\frac{\partial u_1^k}{\partial s}$, and its marginal effect on future utility is $\mathrm{MU}_2^k = \beta_k \frac{\partial u_2^k}{\partial c_2} \frac{\partial g}{\partial h} \frac{\partial f}{\partial s}$. Note that we assume $c_1^k = \gamma(d) W_p(d) = W_k(d)$ for simplicity here, so we don't assume a certain sign on $\frac{\partial W_k(d)}{\partial d}$. The goal is to study the effect of d on s^*, so further take the derivative of $\frac{\partial L^k}{\partial s}$ with respect to d,

$$\begin{aligned} \frac{\partial^2 L^k}{\partial s \partial d} =& \frac{\partial^2 u_1^k}{\partial s^2} \frac{\partial s}{\partial d} + \frac{\partial^2 u_1^k}{\partial s \partial c_1^k} \frac{\partial c_1^k}{\partial d} + \beta_k A \left(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial s} \frac{\partial s}{\partial d} + \frac{\partial f}{\partial c_1^k} \frac{\partial W_k(d)}{\partial d} \right) + \\ & \beta_k \frac{\partial u_2^k}{\partial c_2} \frac{\partial g}{\partial h} \frac{\partial^2 f}{\partial s^2} \frac{\partial s}{\partial d} + \beta_k \frac{\partial u_2^k}{\partial c_2} \frac{\partial g}{\partial h} \left(\frac{\partial^2 f}{\partial s \partial d} + \frac{\partial^2 f}{\partial s \partial c_1^k} \frac{\partial W_k(d)}{\partial d} \right) = 0, \end{aligned}$$

where

$$A = \frac{\partial^2 u_2^k}{\partial c_2^2}(\frac{\partial g}{\partial h})^2 \frac{\partial f}{\partial s} + \frac{\partial u_2^k}{\partial c_2}\frac{\partial f}{\partial s}\frac{\partial^2 g}{\partial h^2} < 0.$$

Therefore,

$$\frac{\partial s^*}{\partial d} = -\frac{\beta_k A\Big(\overbrace{\frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d}}^{\text{Investment effect}} + \overbrace{\frac{\partial f}{\partial d}}^{\text{Direct effect}}\Big) + \beta_k \frac{\partial u_2^k}{\partial c_2}\frac{\partial g}{\partial h}(\frac{\partial^2 f}{\partial s \partial d} + \frac{\partial^2 f}{\partial s \partial c_1^k}\frac{\partial W_k(d)}{\partial d}) + \frac{\partial^2 u_1^k}{\partial s \partial c_1^k}\frac{\partial W_k(d)}{\partial d}}{\frac{\partial^2 u_1^k}{\partial s^2} + \beta_k \frac{\partial u_2^k}{\partial c_2}\frac{\partial g}{\partial h}\frac{\partial^2 f}{\partial s^2} + \beta_k A \frac{\partial f}{\partial s}}.$$

If we further assume the separability of the child utility function and human capital production function, we will get rid of terms of $\frac{\partial^2 u_1^k}{\partial s \partial c_1^k}$, $\frac{\partial^2 f}{\partial s \partial d}$, and $\frac{\partial^2 f}{\partial s \partial c_1^k}$, then $\frac{\partial s}{\partial d}$ is simplified to

$$\frac{\partial s^*}{\partial d} = -\frac{\beta_k A\Big(\overbrace{\frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d}}^{\text{Investment effect}} + \overbrace{\frac{\partial f}{\partial d}}^{\text{Direct effect}}\Big)}{\frac{\partial^2 u_1^k}{\partial s^2} + \beta_k \frac{\partial u_2^k}{\partial c_2}\frac{\partial g}{\partial h}\frac{\partial^2 f}{\partial s^2} + \beta_k A \frac{\partial f}{\partial s}}.$$

The denominator of $\frac{\partial s^*}{\partial d}$ is negative, so the sign of $\frac{\partial s^*}{\partial d}$ depends on its numerator, and specifically depends on the relative size of $\frac{\partial f}{\partial d}$ and $\frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d}$. If assume that $\frac{\partial W_k(d)}{\partial d} \geq 0$ so that $\frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d} \geq 0$, and the negative direct effect of being left-behind is larger than the positive indirect effect through income, then $\frac{\partial s^*}{\partial d} \geq 0$, suggesting that the child will increase study time to compensate for worse performance due to the absence of parent, and vice versa.

Graphically, the original equilibrium of child study time should be at the intersection of the marginal utility of studying in period 1 and period 2, which is s^* in Figure 1.1. For the child, holding study time fixed, if there is a change in parent migration status d, then the child's human capital will change due to the direct effect of migration and its indirect effect through income/consumption, thereby affecting the income and consumption in the future. However, it will not affect consumption in the first period. This is equivalent to stating that holding s fixed, when d changes, h and c_2 will be affected. Recall that the marginal effect of study on current utility only depends on study time s, so even if d changes, the current marginal effect won't change because s is held fixed. The marginal utility of migration in period 2 depends on the discounted marginal effect of study time on future utility through consumption in that period ($\beta_k \frac{\partial u_2^k}{\partial c_2}\frac{\partial g}{\partial h}\frac{\partial f}{\partial s}$). When d

changes, the discount rate β_k won't change, and education production through the indirect effect of study time won' change either because s is held constant. However, education production will be directly affected by migration status and indirectly affected by migration through investment in the child, and thus the returns to education will change. In addition, the change in human capital leads to change in the consumption level, which will affect the marginal utility of consumption in the second period. Considering the changes in returns to education and marginal utility of consumption, the marginal utility of study time in period 2 will be affected.

If d increases, on the one hand, the direct effect of migration will lead to lower human capital. Since the return to education decreases as human capital increases, we would expect an increase in the return to education. As human capital worsens, future income and future consumption will drop, which leads to an increase in the marginal utility of consumption since it's decreasing in consumption levels. Therefore, considering the direct effect of migration, we expect the return to education and the marginal utility of consumption to increase as d increases, and thus the marginal utility of study time in period 2 increases as d increases. Graphically, the curve for marginal utility of study time in period 2 shifts up since the new marginal effect of studying on future utility becomes higher for every level of s. If d increases, on the other hand, the indirect effect of migration through investment in child will lead to higher human capital. Since the return to education decreases as human capital increases, we would expect to see a drop in the return to education. As the child human capital becomes higher, future income and future consumption will increase, leading to a decrease in the marginal utility of consumption since it's decreasing in consumption levels. Due to the indirect effect of migration through income, as d increases, we expect the return to education and the marginal utility of consumption to decrease, and thus the marginal effect of study time on future utility decreases. Graphically, the future marginal effect curve shifts down since the new marginal effect of study on future utility becomes lower for every level of s.

In summary, when migration status increases, although the current marginal effect curve of study time remains unchanged, the shift of the future marginal effect curve will depend on the relative sizes of the two forces from the direct and indirect effect of migration. If the two effects add up to be negative, then the curve will finally shift up and the new equilibrium study time will increase to $s^{*'}$, suggesting that if d increases, s^* is expected to increase, as shown in Figure 1.1. This

suggests that the child will have to study longer to compensate for the large detrimental effect of migration on their school performances. This is consistent to our findings in Appendix A.1.1.

A.1.2. Parental utility maximization. The utility of parent is

$$\begin{aligned} \max_{d} \quad & u_1^p(c_1^p) + \beta_p u_2^p(c_2), \\ s.t. \quad & c_1^p = \gamma_p W_p(d), \\ & c_2 = g(h), \\ & h = f(d, s, c_1^k, h_0). \end{aligned} \tag{A.1.2}$$

Since γ_p is a positive fixed number, I simply omit it from the first period consumption. Plugging constraints to the utility function:

$$L^p = u_1^p(W_p) + \beta_p u_2^p(g(h))$$

With a slight abuse of notation, we write $u_2(h)$ for $u_2(g(h))$. Then we know that

$$\begin{aligned} \frac{\partial u_2^p}{\partial h} &= \frac{\partial u_2^p}{\partial c_2}\frac{\partial g}{\partial h} \geq 0, \\ \frac{\partial^2 u_2^p}{\partial h^2} &= \frac{\partial^2 u_2}{\partial c_2^2}(\frac{\partial g}{\partial h})^2 + \frac{\partial^2 g}{\partial h^2}\frac{\partial u_2^p}{\partial c_2} \leq 0. \end{aligned}$$

Taking the derivative with respect to d and obtain the first order condition

$$\frac{\partial L^p}{\partial d} = \frac{\partial u_1^p}{\partial c_1^p}\frac{\partial c_1^p}{\partial d} + \beta_p \frac{\partial u_2^p}{\partial h}(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d}) = 0.$$

For the parent, when they migrate out, they are potentially benefiting from higher consumption due to higher income in the first period, but at the cost of their child's human capital and thus their future income and consumption. From the first-order condition, we know the marginal effect of parental migration on current utility is $\mathrm{MU}_1^p = \frac{\partial u_1^p}{\partial c_1^p}\frac{\partial c_1^p}{\partial d}$, and its marginal effect on future utility is $\mathrm{MU}_2^p = -\beta_p \frac{\partial u_2^p}{\partial h}(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d})$. To guarantee an interior solution, we need the marginal effect on future utility to be nonnegative, that is, $\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d} \leq 0$.

From the first-order condition, we can derive parent's optimal migration decision d^* as a function of child's study time s. Our goal is to study the effect of s on d^*, so further take the

derivative of $\frac{\partial L}{\partial d}$ with respect to s,

$$\begin{aligned}\frac{\partial^2 L^p}{\partial s \partial d} =& \frac{\partial u_1^p}{\partial c_1^p}\frac{\partial^2 c_1^p}{\partial d^2}\frac{\partial d}{\partial s} + \frac{\partial^2 u_1^p}{\partial {c_1^p}^2}(\frac{\partial c_1^p}{\partial d})^2\frac{\partial d}{\partial s} + \\ & \beta_p(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d})\frac{\partial^2 u_2^p}{\partial h^2}(\frac{\partial f}{\partial d}\frac{\partial d}{\partial s} + \frac{\partial f}{\partial s} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d}\frac{\partial d}{\partial s}) + \\ & \beta_p\frac{\partial u_2^p}{\partial h}[\frac{\partial W_k(d)}{\partial d}(\frac{\partial^2 f}{\partial c_1^k \partial d}\frac{\partial d}{\partial s} + \frac{\partial^2 f}{\partial c_1^k \partial s} + \frac{\partial^2 f}{\partial {c_1^k}^2}\frac{\partial W_k(d)}{\partial d}\frac{\partial d}{\partial s}) + \frac{\partial f}{\partial c_1^k}\frac{\partial^2 W_k(d)}{\partial d^2}\frac{\partial d}{\partial s}] = 0.\end{aligned}$$

Since we assume the separability of human capital production function, i.e., $\frac{\partial^2 f}{\partial s \partial d} = \frac{\partial^2 f}{\partial s \partial c_1^k} = \frac{\partial^2 f}{\partial c_1^k \partial d} = 0$, the second-order condition can be simplified, and thus

$$\frac{\partial d^*}{\partial s} = \frac{-\beta_p \frac{\partial^2 u_2^p}{\partial h^2}\frac{\partial f}{\partial s}(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d})}{\frac{\partial u_1^p}{\partial c_1^p}\frac{\partial^2 c_1^p}{\partial d^2} + \frac{\partial^2 u_1^p}{\partial {c_1^p}^2}(\frac{\partial c_1^p}{\partial d})^2 + \beta_p\frac{\partial^2 u_2^p}{\partial h^2}(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d})^2 + \beta_p\frac{\partial u_2^p}{\partial h}[\frac{\partial^2 f}{\partial d^2} + \frac{\partial^2 f}{\partial {c_1^k}^2}(\frac{\partial W_k(d)}{\partial d})^2 + \frac{\partial f}{\partial c_1^k}\frac{\partial^2 W_k(d)}{\partial d^2}]}.$$

Since $\frac{\partial u_1^p}{\partial c_1^p} \geq 0, \frac{\partial^2 u_1^p}{\partial {c_1^p}^2} \leq 0; \frac{\partial u_2^p}{\partial h} \geq 0, \frac{\partial^2 u_2^p}{\partial h^2} \leq 0; \frac{\partial c_1^p}{\partial d} \geq 0, \frac{\partial^2 c_1^p}{\partial d^2} \leq 0; \frac{\partial f}{\partial c_1^k} \geq 0, \frac{\partial^2 f}{\partial {c_1^k}^2} \leq 0; \frac{\partial f}{\partial d} \leq 0, \frac{\partial^2 f}{\partial d^2} \leq 0$, and $\beta_p > 0$, the denominator of $\frac{\partial d^*}{\partial s}$ is negative. The numerator is also negative since $\frac{\partial f}{\partial s} \geq 0$ and $\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d} \leq 0$. Thus, $\frac{\partial d^*}{\partial s} \geq 0$ as long as there is an interior solution. This suggests that if the child is willing to study for longer times, parent will be more "assured" and more likely to migrate out.

Graphically, the original equilibrium of parent migration decision should be at the intersection of the marginal utility of migration in the first period and the marginal utility in the second period, which is d_0^* in Figure 1.2. Holding parent migration status constant, if there is a change in child's study time, then the child's human capital will be affected, thereby affecting the income and consumption in the second period. However, it will not affect the consumption in the first period. This is equivalent to stating that when holding d fixed and changing s, c_1^p will remain unchanged but h and c_2 will be affected. Recall that the marginal utility from migration in the first period is the marginal effect of migration status on current utility through consumption in that period ($\frac{\partial u_1^p}{\partial c_1^p}\frac{\partial c_1^p}{\partial d}$), so even if s changes, the marginal utility in the first period won't change because d and c_1^p remain the same. The marginal utility in the second period is the discounted marginal effect of migration status on future utility through consumption in that period ($-\beta_p\frac{\partial u_2^p}{\partial h}(\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d})$). future consumption depends only on child human capital. When child study time s changes, education production

will be affected [1], the marginal utility in the second period from consumption/human capital will be affected, so the marginal utility of migration in the second period will be affected. If s increases, h will increase as the marginal effect of study time on human capital production is positive. Therefore, consumption in the second period increases as child human capital increases. Since the marginal utility is decreasing in consumption, we expect to see a decrease in marginal utility from future consumption, so the marginal utility of migration in Period 2 will decrease. Therefore, when the child increases study time, although parent's marginal utility in Period 1 will not change, the curve for marginal utility in Period 2 will shift down since it becomes lower for every level of d. This results in the new equilibrium of migration status to increase, as shown in Figure 1.2. That is, d^* is increasing in s. This is consistent to our findings in Appendix A.1.2.

A.1.3. A specific functional form. There might be some concern in the above decision making process since I am assuming simultaneous decisions. In this section, I will use specific functional forms to show that the joint decision process of parent and child will lead to one unique equilibrium. In that case, it makes no difference if we are assuming a simultaneous decision process or a sequential one. In addition, the specific functional forms I choose is also consistent with my empirical model.

For the child decision process, the utility maximization satisfying the previous assumptions could be depicted by:

$$\begin{aligned} \max_{s} \quad & \log[(1-s)T_0] + \log(c_1^k) + \beta_k \log(c_2), \\ s.t. \quad & c_1^k \leq a + w_1 \cdot d, \\ & c_2 \leq w_2 \cdot e, \\ & h \leq \gamma_T \cdot s \cdot T_0 + \gamma_W(a + w_1 d) + \gamma_D \cdot d, \end{aligned}$$

where T_0 is total weekly time available to the child, and d is a measure of parent migration status. For simplicity, I assume $\gamma(D)$ in Equation (A.1.1) to be constant.

[1] Education production through the direct effect of migration or the indirect effect through current consumption ($\frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d}$) won't change because c_1^k and d remain the same.

Plugging the constraints into the objective function, we have

$$L^k = \log[(1-s)T_0] + \log(a + w_1 \cdot d) + \beta_k \log[w_2(\gamma_T \cdot s \cdot T_0 + \gamma_W(a + w_1 d) + \gamma_D d)].$$

Taking its first-order derivative with respect to s, we have

$$\frac{\partial L^k}{\partial s} = -\frac{1}{1-s} + \frac{\gamma_T \cdot T_0 \cdot \beta_k}{\gamma_T \cdot s \cdot T_0 + \gamma_W(a + w_1 d) + \gamma_D d}$$

Setting the first-order condition to 0, we could solve for s^*, the optimal time decision of children:

$$s^* = \frac{\gamma_T \cdot T_0 \cdot \beta_k + a \cdot \gamma_W + (\gamma_D + w_1 \cdot \gamma_W) \cdot d}{\gamma_T T_0(\beta_k - 1)} \tag{A.1.3}$$

Since $\gamma_D + \gamma_W w_1 = \frac{\partial f}{\partial d} + \frac{\partial f}{\partial c_1^k}\frac{\partial W_k(d)}{\partial d} \leq 0$, and $\gamma_T T_0(\beta_k - 1) < 0$ due to the fact that discount factor $0 \leq \beta_k < 1$, we know that s^* is non-decreasing as d increases, which is consistent to our findings in Appendix A.1.1.

For the parent decision process, the utility maximization process is depicted by:

$$\begin{aligned} \max_d \quad & \log(c_1^p) + \beta_p \log(c_2), \\ s.t. \quad & c_1^p \leq a + w_1 \cdot d, \\ & c_2 \leq w_2 \cdot e, \\ & h \leq \gamma_T \cdot s \cdot T_0 + \gamma_W(a + w_1 d) + \gamma_D \cdot d. \end{aligned}$$

Plugging in the constraints to the objective function,

$$L^p = \log(a + w_1 \cdot d) + \beta_p \log[w_2(\gamma_T \cdot s \cdot T_0 + \gamma_W(a + w_1 d) + \gamma_D \cdot d)].$$

Taking the first-order derivative with respect to D, we have

$$\frac{\partial L^p}{\partial d} = \frac{w_1}{a + w_1 d} - \frac{\gamma_W w_1 + \gamma_D \beta_p}{\gamma_T \cdot s \cdot T_0 + \gamma_W(a + w_1 d) + \gamma_D \cdot d}.$$

Setting the first-order condition to 0, we have

$$d^* = \frac{-a\gamma_D \beta_p + w_1 \gamma_T T_0 s}{w_1 \gamma_D(\beta_p - 1)}. \tag{A.1.4}$$

Since $w_1\gamma_T T_0 \geq 0$, and $w_1\gamma_D(\beta_p - 1) > 0$ because $\gamma_D < 0$ and $0 \leq \beta_p < 1$, we know d^* is non-decreasing as s increases, which is consistent to our finding in Appendix A.1.2.

Next I will show there is a unique equilibrium (s^{**}, d^{**}). If we draw the reaction function of the parent and the child on one graph with d on the horizontal axis and s on the vertical axis, this is equivalent to show that the reactions curves have different slopes. The slope of child's reaction function is $\frac{\gamma_D + w_1\gamma_W}{\gamma_T T_0(\beta_k - 1)}$, and the slope of parent's reaction function is $\frac{w_1\gamma_D(\beta_p - 1)}{w_1\gamma_T T_0}$:

$$\begin{aligned}
&\frac{\gamma_D + w_1\gamma_W}{\gamma_T T_0(\beta_k - 1)} - \frac{w_1\gamma_D(\beta_p - 1)}{w_1\gamma_T T_0} \\
=&\frac{w_1(\gamma_D + w_1\gamma_W) - w_1\gamma_D(\beta_p - 1)(\beta_k - 1)}{w_1\gamma_T T_0(\beta_k - 1)} \\
=&\frac{w_1^2\gamma_W - w_1\gamma_D(\beta_p\beta_k - \beta_p - \beta_k)}{w_1\gamma_T T_0(\beta_k - 1)} \\
=&\frac{w_1[w_1\gamma_W - \gamma_D(\beta_p\beta_k - \beta_p - \beta_k)]}{w_1\gamma_T T_0(\beta_k - 1)}
\end{aligned}$$

We already know the denominator of the difference is negative since $\beta_k - 1 < 0$, so the value of the difference only depends on the numerator. As long as we have $w_1\gamma_W \neq \gamma_D(\beta_p\beta_k - \beta_p - \beta_k)$, the slopes will be different. Since $-1 < \beta_p\beta_k - \beta_p - \beta_k \leq 0$, if the negative direct effect γ_D is very large, then the numerator of the difference would be negative so the difference would be positive, suggesting that the slope of child's reaction function would be steeper. This also makes intuitive sense because if γ_D is very large, then based on the graph of marginal utility, to compensate for the negative direct effect, the child tends to increase study time by a lot, and the reaction is stronger than parent's. The graph for this case is depicted in Figure 1.3. The equilibrium study time s^{**} and equilibrium migration decision d^{**} is unique. Since the observed data are in equilibrium, we have s^{**} is given by (A.1.3) with $d = d^{**}$ and d^{**} is the solution of (A.1.3) and (A.1.4). This provides a concrete example for the abstract system (1.2.6).

A.2. Complementary empirical results

In this appendix, I will present supplementary results by imposing stronger yet potentially invalid assumptions like exogeneity of treatments/mediators or missing-at-random mediators. These results should be viewed as robustness checks or even sanity checks, which highlight the issues of failure to handle endogeneity and non-random missing values carefully.

A.2.1. Results without accounting for non-random missing values (no imputation). When the mediators are missing at random, there is no non-random missing issue and thus one can estimate the model on units without missing values. Table A.1 - A.3 present the results under the same setting as Table 1.3 - 1.7, except that Heckman model is not applied to impute the missing study time and investment in child.

We can observe that for both all sample and subgroup analysis, the direct effect shrinks slightly and the indirect effect through investment shrinks drastically, despite that the sample size only drops by 35%. This corroborates my speculation that simply removing these observations tend to underestimate the impact of migration, because those whose parents or guardians fail to report their study time or investment in them are likely suffering more from parental migration.

Table A.1. Effect of parental migration on child schooling outcomes (IV, all sample, not imputed)

			(1) Language	(2) Math
Direct Effect				
Parental Accompany			-0.458***	-0.411**
			(0.001)	(0.003)
Indirect Effect				
Study time			-0.008	-0.006
			(0.096)	(0.203)
Investment in children			-0.402**	-0.437***
			(0.003)	(0.001)
Sepecification Tests	*(1) Study time*	*(2) Investment*	*(3) Language*	*(4) Math*
Underidentification test				
(Anderson canon. corr. LM statistic)	34.769***	34.769***	18.155***	18.155***
	(0.000)	(0.000)	(0.000)	(0.000)
Overidentification test (Sargan statistic)	5.183*	3.629	1.561	1.487
	(0.023)	(0.057)	(0.458)	(0.475)
Endogeneity test	0.281	5.175*	19.310***	30.792***
	(0.596)	(0.023)	(0.000)	(0.000)
Obs.		1277		

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A.2. Effect of parental migration on child schooling outcomes (IV, subgroup by gender, not Imputed)

	Girl		Boy	
	(1) Language	(2) Math	(3) Language	(4) Math
Direct Effect				
Parental Accompany	-0.350*	-0.371	-0.340**	-0.236*
	(0.035)	(0.057)	(0.009)	(0.029)
Indirect Effect				
Study time	-0.000	-0.001	-0.026	-0.007
	(0.762)	(0.721)	(0.056)	(0.586)
Investment in children	-0.600*	-0.780*	-0.121*	-0.129***
	(0.043)	(0.030)	(0.011)	(0.001)
Sepecification Tests	*(1) Study time*	*(2) Investment*	*(3) Language*	*(4) Math*
Girl				
Underidentification test				
(Anderson canon. corr. LM statistic)	5.242	5.242	7.516	7.516
	(0.073)	(0.073)	(0.057)	(0.057)
Overidentification test (Sargan statistic)	2.518	5.040*	0.959	1.256
	(0.113)	(0.025)	(0.619)	(0.534)
Endogeneity test	0.021	3.314	3.684	10.138*
	(0.885)	(0.069)	(0.298)	(0.017)
Obs.		571		
Boy				
Underidentification test				
(Anderson canon. corr. LM statistic)	36.416***	36.416***	9.968*	9.968*
	(0.000)	(0.000)	(0.019)	(0.019)
Overidentification test (Sargan statistic)	3.559	0.531	1.081	0.781
	(0.059)	(0.466)	(0.582)	(0.677)
Endogeneity test	0.069	0.871	19.510***	21.147***
	(0.793)	(0.351)	(0.000)	(0.000)
Obs.		706		

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A.3. Effect of parental migration on child schooling outcomes (IV, subgroup by birth order, not imputed)

	First child		Subsequent children	
	(1) Language	(2) Math	(3) Language	(4) Math
Direct Effect				
Parental Accompany	-0.552	-0.647*	-0.469*	-0.282
	(0.075)	(0.040)	(0.014)	(0.091)
Indirect Effect				
Study time	-0.038	-0.046	0.001	-0.001
	(0.303)	(0.303)	(0.766)	(0.766)
Investment in children	-0.168	-0.260	-0.624*	-0.496*
	(0.227)	(0.114)	(0.025)	(0.028)
Sepecification Tests	*(1) Study time*	*(2) Investment*	*(3) Language*	*(4) Math*
First child				
Underidentification test				
(Anderson canon. corr. LM statistic)	14.970***	14.970***	2.693	2.693
	(0.001)	(0.001)	(0.260)	(0.260)
Overidentification test (Sargan statistic)	0.469	3.359	0.001	0.291
	(0.493)	(0.067)	(0.980)	(0.590)
Endogeneity test	0.636	0.119	6.198	14.192**
	(0.425)	(0.730)	(0.102)	(0.003)
Obs.		590		
Subsequent children				
Underidentification test				
(Anderson canon. corr. LM statistic)	7.995*	7.995*	4.635	4.635
	(0.018)	(0.018)	(0.099)	(0.099)
Overidentification test (Sargan statistic)	3.313	0.160	0.974	1.797
	(0.069)	(0.690)	(0.324)	(0.180)
Endogeneity test	1.288	9.416**	9.994*	6.080
	(0.256)	(0.002)	(0.019)	(0.108)
Obs.		535		

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A.2.2. Results without accounting for endogeneity (no IV). Recall the structural equation model (1.4.1) - (1.4.4) in Section 1.4. If D, T, W were all exogenous, $\epsilon_{Pi}, \epsilon_{Ti}, \epsilon_{Wi}, \epsilon_{Di}$ would be mutually independent, in which case no instrumental variable is needed for identification because the coefficients can be identified separately from each single equation. The results are presented in Table A.4 - 1.7.

Nevertheless, the endogeneity tests in Table 1.3 - 1.7 show strong evidence against exogeneity. Therefore, the results in Table A.4 - 1.7 should be taken as a robustness check. We can observe that both the direct effect and indirect effect through investment have the same signs as the results in Section 1.5, but they are underestimated drastically for both all samples and subgroups. This highlights the importance to handle the endogeneity in this problem.

Table A.4. Effect of parental migration on child schooling outcomes (no IV, all sample, imputed)

	Language Score	Math Score
Direct Effect		
Parental Accompany	-0.074**	-0.015
	(0.007)	(0.610)
Indirect Effect		
Study time	-0.002	-0.005**
	(0.159)	(0.008)
Investment in children	-0.004*	-0.005*
	(0.035)	(0.035)
Obs.	1971	

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.5. Effect of parental migration on child schooling outcomes (no IV, subgroup by gender, imputed)

	Girl		Boy	
	(1) Language	(2) Math	(3) Language	(4) Math
Direct Effect				
Parental Accompany	-0.126***	-0.077*	-0.018	0.050
	(0.001)	(0.045)	(0.650)	(0.253)
Indirect Effect				
Study time	-0.003	-0.003	-0.002	-0.010**
	(0.370)	(0.403)	(0.428)	(0.003)
Investment in children	-0.002	-0.002	-0.007*	-0.012**
	(0.387)	(0.387)	(0.026)	(0.009)
Obs.	887		1084	

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

A.2.3. Results without accounting for both. Finally I present the results without accounting for either endogeneity or non-random missing values in Table A.7 - A.9 as an additional sanity check. This analysis even fails to capture the significant negative effect through investment.

TABLE A.6. Effect of parental migration on child schooling outcomes (no IV, subgroup by birth order, imputed)

	First child		Subsequent children	
	(1) Language	(2) Math	(3) Language	(4) Math
Direct Effect				
Parental Accompany	-0.132***	-0.067	-0.042	0.037
	(0.000)	(0.094)	(0.340)	(0.437)
Indirect Effect				
Study time	0.001	-0.002	-0.007*	-0.013***
	(0.743)	(0.464)	(0.016)	(0.000)
Investment in children	-0.001	-0.007*	-0.005	-0.005
	(0.713)	(0.026)	(0.217)	(0.217)
Obs.	891		860	

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A.7. Effect of parental migration on child schooling outcomes (no IV, all sample, not imputed)

	Language Score	Math Score
Direct Effect		
Parental Accompany	-0.074*	-0.013
	(0.016)	(0.688)
Indirect Effect		
Study time	-0.001	-0.002
	(0.392)	(0.392)
Investment in children	-0.002	-0.004
	(0.301)	(0.301)
Obs.	1277	

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A.8. Effect of parental migration on child schooling outcomes (no IV, subgroup by gender, not imputed)

	Girl		Boy	
	(1) Language	(2) Math	(3) Language	(4) Math
Direct Effect				
Parental Accompany	-0.110*	-0.052	-0.045	0.021
	(0.013)	(0.254)	(0.281)	(0.659)
Indirect Effect				
Study time	-0.001	-0.001	-0.000	-0.003
	(0.501)	(0.501)	(0.700)	(0.391)
Investment in children	-0.001	-0.001	-0.005	-0.009
	(0.806)	(0.806)	(0.169)	(0.169)
Obs.	571		706	

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.9. Effect of parental migration on child schooling outcomes (no IV, subgroup by birth order, not imputed)

	First child		Subsequent children	
	(1) Language	(2) Math	(3) Language	(4) Math
Direct Effect				
Parental Accompany	-0.084*	-0.024	-0.107*	-0.018
	(0.036)	(0.601)	(0.035)	(0.730)
Indirect Effect				
Study time	0.001	-0.001	0.005	0.008
	(0.751)	(0.716)	(0.174)	(0.174)
Investment in children	-0.000	-0.007	0.010	0.011
	(0.899)	(0.101)	(0.200)	(0.200)
Obs.	590		535	

p-values in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APPENDIX B

Long Title of Appendix B

B.1. Statistical Properties of RIPW Estimators

B.1.1. Notation, assumptions, and preliminaries. Throughout this section we consider a generalized version of the extended causal framework introduced in Section 3.3.1. Since the inferential target DATE is defined conditional on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$, we treat them as fixed quantities. As discussed in Section 3.3.1, we can suppress $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$ and treat the quantities $\mathcal{Z}_i = (\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{W}_i)$ as non-identically distributed units.

In Section 3.3.1, we consider the special case where $(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{W}_i, \boldsymbol{X}_i, \boldsymbol{U}_i)$ are i.i.d. (Assumption 3.3.3), in which case $\mathcal{Z}_i$ are independent. Here we consider the more general case where $\mathcal{Z}_i$ can be dependent (conditioning on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$). The (conditional) maximal correlation ρ_{ij} between $\mathcal{Z}_i = (\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{W}_i)$ and $\mathcal{Z}_j = (\boldsymbol{Y}_j(1), \boldsymbol{Y}_j(0), \boldsymbol{W}_j)$ is defined as

$$\rho_{ij} = \sup_{f,g} \mathrm{Corr}(f(\mathcal{Z}_i), g(\mathcal{Z}_j) \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_n).$$

For design-based inference, $(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{X}_i, \boldsymbol{U}_i)$ are all fixed and thus the above definition coincides with (3.2.8).

To simplify the notation, we use the symbol $\mathbb{E}$ and $\mathbb{P}$ to denote the expectation and probability conditioning on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$. Occasionally, we use $\mathbb{E}_{\text{full}}$ and $\mathbb{P}_{\text{full}}$ to denote the unconditional expectation and probability. We emphasize that $\mathbb{E}_{\text{full}}$ and $\mathbb{P}_{\text{full}}$ will only be used to make connections with the simplified results in Section 3.3, but will not be used anywhere in the proofs. With this notation,

$$\tau_{it} = \mathbb{E}[Y_{it}(1) - Y_{it}(0)], \quad \boldsymbol{\pi}_i(\boldsymbol{w}) = \mathbb{P}[\boldsymbol{W}_i = \boldsymbol{w}],$$

and

$$m_{it} = \mathbb{E}[Y_{it}(0)] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_{it}(0)] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{it}(0)] + \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathbb{E}[Y_{it}(0)].$$

The DATE estimand with weight ξ is defined as

$$\tau^*(\xi) = \sum_{t=1}^{T} \xi_t \left(\frac{1}{n} \sum_{i=1}^{n} \tau_{it} \right).$$

With a reshaped distribution $\mathbf{\Pi}$ on $\{0,1\}^T$, the RIPW estimator is defined as

$$\hat{\tau}(\mathbf{\Pi}) \triangleq \underset{\tau,\mu,\sum_i \alpha_i = \sum_t \gamma_t = 0}{\arg\min} \sum_{i=1}^{n} \sum_{t=1}^{T} ((Y_{it}^{\text{obs}} - \hat{m}_{it}) - \mu - \alpha_i - \gamma_t - W_{it}\tau)^2 \frac{\mathbf{\Pi}(\boldsymbol{W}_i)}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)}.$$

We will suppress ξ from $\tau^*(\xi)$ and $\mathbf{\Pi}$ from $\hat{\tau}(\mathbf{\Pi})$ throughout the section.

It is easy to see that $\hat{\tau}$ is invariant if $Y_{it}(w)$ is replaced by $Y_{it}(w) - \mu' - \alpha_i' - \gamma_t'$ for any constants $\mu', \{\alpha_i' : i \in [n]\}, \{\gamma_t' : t \in [T]\}$, and in particular,

$$\mu' = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E}[Y_{it}(0)], \quad \alpha_i' = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[Y_{it}(0)] - \mu', \quad \gamma_t' = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_{it}(0)] - \mu'.$$

Therefore, we can assume without loss of generality that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_{it}(0)] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[Y_{it}(0)] = 0, \quad t = 1, \ldots, T,\ i = 1, \ldots, n.$$

In this case,

$$\boldsymbol{m}_i = \mathbb{E}[\boldsymbol{Y}_i(0)], \quad \tilde{\boldsymbol{Y}}_i(0) = \boldsymbol{Y}_i(0) - \mathbb{E}[\boldsymbol{Y}_i(0)] - (\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i). \tag{B.1.1}$$

To be self-contained, we summarize all notation here. First, for notational convenience, we use the bold letter with a subscript i to denote the vector across T time periods. For instance, $\boldsymbol{Y}_i^{\text{obs}} = (Y_{i1}^{\text{obs}}, \ldots, Y_{iT}^{\text{obs}})$. It also includes $\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0), \boldsymbol{m}_i, \hat{\boldsymbol{m}}_i, \boldsymbol{\tau}_i$. Let $J = I_T - \mathbf{1}_T \mathbf{1}_T^\top / T$,

$$\Theta_i = \mathbf{\Pi}(\boldsymbol{W}_i)/\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i), \quad \tilde{\boldsymbol{Y}}_i^{\text{obs}} = \boldsymbol{Y}_i^{\text{obs}} - \hat{\boldsymbol{m}}_i, \quad \tilde{\boldsymbol{Y}}_i(w) = \boldsymbol{Y}_i(w) - \hat{\boldsymbol{m}}_i,\ w \in \{0,1\},$$

and

$$\Gamma_\theta \triangleq \frac{1}{n} \sum_{i=1}^{n} \Theta_i, \quad \Gamma_{ww} \triangleq \frac{1}{n} \sum_{i=1}^{n} \Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i, \quad \Gamma_{wy} \triangleq \frac{1}{n} \sum_{i=1}^{n} \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}},$$

$$\mathbf{\Gamma}_w \triangleq \frac{1}{n} \sum_{i=1}^{n} \Theta_i J \boldsymbol{W}_i, \quad \mathbf{\Gamma}_y \triangleq \frac{1}{n} \sum_{i=1}^{n} \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}.$$

Furthermore, we define

$$\mathcal{V}_i = \Theta_i \Bigg\{ \Big(\mathbb{E}[\Gamma_{wy}] - \tau^* \mathbb{E}[\Gamma_{ww}]\Big) - \Big(\mathbb{E}[\boldsymbol{\Gamma}_y] - \tau^* \mathbb{E}[\boldsymbol{\Gamma}_w]\Big)^\top J \boldsymbol{W}_i$$

$$+ \mathbb{E}[\Gamma_\theta] \boldsymbol{W}_i^\top J \Big(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^* \boldsymbol{W}_i\Big) - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top J \Big(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^* \boldsymbol{W}_i\Big) \Bigg\}.$$

This coincides with the definition in Theorem 3.2.2 when $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ and $\hat{\boldsymbol{m}}_i = 0$.

Next, we define the treatment and outcome models. the treatment model is perfectly estimated for unit i if

$$\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) = \boldsymbol{\pi}_i(\boldsymbol{w}), \quad \forall \boldsymbol{w} \in \{0,1\}^T,$$

and the precision is defined as

$$\Delta_{\pi i} = \sqrt{\mathbb{E}\left[\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i) \mid \hat{\boldsymbol{\pi}}_i\right]^2}.$$

The outcome model is perfectly estimated for unit i if

$$\hat{m}_{it} = m_{it}, \quad \tau_{it} = \tau^*,$$

and the precision is defined as

$$\Delta_{yi} = \sqrt{\mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i\|_2^2 \mid \hat{\boldsymbol{m}}_i]} + \|\boldsymbol{\tau}_i - \tau^* \mathbf{1}_T\|_2.$$

We also define the average precision $\bar{\Delta}_\pi, \bar{\Delta}_y$, and $\bar{\Delta}_{\pi y}$ as

$$\bar{\Delta}_\pi = \sqrt{\frac{1}{n}\sum_{i=1}^n \Delta_{\pi i}^2}, \quad \bar{\Delta}_y = \sqrt{\frac{1}{n}\sum_{i=1}^n \Delta_{yi}^2}.$$

The above precision measures are essentially the conditional versions of $\delta_{\pi i}, \delta_{yi}, \bar{\delta}_y$, and $\bar{\delta}_\pi$ in Section 3.3. Specifically,

$$\delta_{\pi i}^2 = \mathbb{E}_{\text{full}}[\Delta_{\pi i}^2], \quad \delta_{yi}^2 = \mathbb{E}_{\text{full}}[\Delta_{yi}^2], \quad \bar{\delta}_\pi^2 = \mathbb{E}_{\text{full}}[\bar{\Delta}_\pi^2], \quad \bar{\delta}_y^2 = \mathbb{E}_{\text{full}}[\bar{\Delta}_y^2].$$

As a result, by Markov's inequality,

$$\bar{\delta}_\pi \bar{\delta}_y = o(1) \Longrightarrow \mathbb{P}\Big(\bar{\Delta}_\pi \bar{\Delta}_y \geq \epsilon_n\Big) \geq 1 - \epsilon_n,$$

for some deterministic sequence $\epsilon_n \to 0$. Therefore, if we can prove the double robustness only assuming $\bar{\Delta}_\pi \bar{\Delta}_y = o(1)$ conditional on $(\boldsymbol{X}_i, \boldsymbol{U}_i, \hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)_{i=1}^n$, we can prove it assuming that $\bar{\delta}_\pi \bar{\delta}_y = o(1)$ as in Section 3.3.

Finally, we state the core assumptions. We start by restating the latent mean ignorability assumption based on the simplified notation.

Assumption B.1.1. (Latent mean ignorability)

$$\mathbb{E}[(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0)) \mid \boldsymbol{W}_i] = \mathbb{E}[(\boldsymbol{Y}_i(1), \boldsymbol{Y}_i(0))] \tag{B.1.2}$$

Next, we restate the overlap condition (Assumption 3.3.2) below with the constant c replaced by c_π to be more informative in the proofs.

Assumption B.1.2. *There exists a universal constant $c > 0$ and a non-stochastic subset $\mathbb{S}^* \subset \{0,1\}^T$ with at least two elements and at least one element not in $\{\boldsymbol{0}_T, \boldsymbol{1}_T\}$, such that*

$$\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) > c_\pi, \boldsymbol{\pi}_i(\boldsymbol{w}) > c_\pi, \quad \forall \boldsymbol{w} \in \mathbb{S}^*, i \in [n], \quad \textit{almost surely.} \tag{B.1.3}$$

Finally, we state the following assumption that unify and substantially generalize Assumptions 3.2.2-3.2.3 for design-based inference and Assumptions 3.3.3-3.3.4 for doubly-robust inference.

Assumption B.1.3. *There exists $q \in (0,1]$,*

$$\frac{1}{n^2} \sum_{i=1}^n \rho_i \left\{ \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + 1 \right\} = O(n^{-q}),$$

and

$$\frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 \right\} = O(1).$$

We close this section by a basic property of the maximal correlation.

Lemma B.1.1. *Let f_i be any deterministic function on the domain of $\mathcal{Z}_i$. Then*

$$\mathrm{Var}\left[\sum_{i=1}^n f_i(\mathcal{Z}_i) \right] \le \frac{1}{2} \sum_{i=1}^n \mathrm{Var}[f_i(\mathcal{Z}_i)] \rho_i.$$

Proof. By definition of ρ_{ij},

$$\mathrm{Cov}\left(f_i(\mathcal{Z}_i), f_j(\mathcal{Z}_j) \right) \le \rho_{ij} \sqrt{\mathrm{Var}[f_i(\mathcal{Z}_i)] \mathrm{Var}[f_j(\mathcal{Z}_j)]} \le \frac{\rho_{ij}}{2} \left\{ \mathrm{Var}[f_i(\mathcal{Z}_i)] + \mathrm{Var}[f_j(\mathcal{Z}_j)] \right\}.$$

Thus,

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^{n} f_i(\mathcal{Z}_i)\right] &= \sum_{i,j=1}^{n} \mathrm{Cov}(f_i(\mathcal{Z}_i), f_j(\mathcal{Z}_j)) \\ &\le \sum_{i,j=1}^{n} \frac{\rho_{ij}}{2}\left\{\mathrm{Var}[f_i(\mathcal{Z}_i)] + \mathrm{Var}[f_j(\mathcal{Z}_j)]\right\} = \sum_{i=1}^{n} \mathrm{Var}[f_i(\mathcal{Z}_i)]\rho_i.\end{aligned}$$

□

B.1.2. A non-stochastic formula of RIPW estimators.

THEOREM B.1.1. *With the same notation as Theorem 3.2.2, $\hat{\tau} = \mathcal{N}/\mathcal{D}$, where*

$$\mathcal{N} = \Gamma_{wy}\Gamma_\theta - \mathbf{\Gamma}_w^\top \mathbf{\Gamma}_y, \quad \mathcal{D} = \Gamma_{ww}\Gamma_\theta - \mathbf{\Gamma}_w^\top \mathbf{\Gamma}_w. \tag{B.1.4}$$

PROOF. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_t)^\top$ be any vector with $\boldsymbol{\gamma}^\top \mathbf{1}_T = 0$. First we derive the optimum $\hat{\mu}(\boldsymbol{\gamma}, \tau)$, $\hat{\alpha}_i(\boldsymbol{\gamma}, \tau)$ given any values of $\boldsymbol{\gamma}$ and τ. Recall that

$$(\hat{\mu}(\boldsymbol{\gamma}, \tau), \hat{\alpha}_i(\boldsymbol{\gamma}, \tau)) = \underset{\sum_i \alpha_i = 0}{\arg\min} \sum_{i=1}^{n} \left(\sum_{t=1}^{T} (\tilde{Y}_{it}^{\mathrm{obs}} - \mu - \alpha_i - \gamma_t - W_{it}\tau)^2 \right) \Theta_i.$$

Since the weight Θ_i only depends on i, it is easy to see that

$$\hat{\mu}(\boldsymbol{\gamma}, \tau) + \hat{\alpha}_i(\boldsymbol{\gamma}, \tau) = \frac{1}{T}\sum_{t=1}^{T} (\tilde{Y}_{it}^{\mathrm{obs}} - \gamma_t - W_{it}\tau), \quad \hat{\mu}(\boldsymbol{\gamma}, \tau) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} (\tilde{Y}_{it}^{\mathrm{obs}} - \gamma_t - W_{it}\tau).$$

As a result,

$$\begin{aligned}&\sum_{t=1}^{T} (\tilde{Y}_{it}^{\mathrm{obs}} - \hat{\mu}(\boldsymbol{\gamma}, \mu) - \hat{\alpha}_i(\boldsymbol{\gamma}, \mu) - \gamma_t - W_{it}\tau)^2 \\ =&\left\| \left(\tilde{Y}_i^{\mathrm{obs}} - \boldsymbol{\gamma} - \boldsymbol{W}_i\tau\right) - \frac{\mathbf{1}_T \mathbf{1}_T^\top}{T}\left(\tilde{Y}_i^{\mathrm{obs}} - \boldsymbol{\gamma} - \boldsymbol{W}_i\tau\right) \right\|_2^2 \\ =&\left\| J\left(\tilde{Y}_i^{\mathrm{obs}} - \boldsymbol{\gamma} - \boldsymbol{W}_i\tau\right) \right\|_2^2.\end{aligned}$$

This yields a profile loss function for $\boldsymbol{\gamma}$ and τ:

$$(\hat{\boldsymbol{\gamma}}, \hat{\tau}) = \underset{\boldsymbol{\gamma}^\top \mathbf{1}_T = 0}{\arg\min} \sum_{i=1}^{n} \left\| J\left(\tilde{Y}_i^{\mathrm{obs}} - \boldsymbol{\gamma} - \boldsymbol{W}_i\tau\right) \right\|_2^2 \Theta_i = \underset{\boldsymbol{\gamma}^\top \mathbf{1}_T = 0}{\arg\min} \sum_{i=1}^{n} \left\| J\left(\tilde{Y}_i^{\mathrm{obs}} - \boldsymbol{W}_i\tau\right) - \boldsymbol{\gamma} \right\|_2^2 \Theta_i,$$

where the last equality uses the fact that $J\boldsymbol{\gamma} = \boldsymbol{\gamma}$. Given τ, the optimizer $\hat{\boldsymbol{\gamma}}(\tau)$ is simply the weighted average of $\{J(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau)\}_{i=1}^n$ in absence of the constraint $\boldsymbol{\gamma}^\top \mathbf{1}_T = 0$, i.e.

$$\hat{\boldsymbol{\gamma}}(\tau) = \frac{\sum_{i=1}^n \Theta_i J(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau)}{\sum_{i=1}^n \Theta_i} = \frac{\boldsymbol{\Gamma}_y}{\Gamma_\theta} - \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\tau.$$

Noting that $\hat{\boldsymbol{\gamma}}(\tau)^\top \mathbf{1}_T = 0$ since $J\mathbf{1}_T = 0$, $\hat{\boldsymbol{\gamma}}(\tau)$ is also the minimizer of the constrained problem, i.e.

$$\hat{\boldsymbol{\gamma}}(\tau) = \operatorname*{arg\,min}_{\boldsymbol{\gamma}^\top \mathbf{1}_T = 0} \sum_{i=1}^n \left\| J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau\right) - \boldsymbol{\gamma} \right\|_2^2 \Theta_i.$$

Plugging in $\hat{\boldsymbol{\gamma}}(\tau)$ yields a profile loss function for τ

$$\hat{\tau} = \arg\min \sum_{i=1}^n \left\| J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau\right) - \hat{\boldsymbol{\gamma}}(\tau) \right\|_2^2 \Theta_i \triangleq L(\tau).$$

A direct calculation shows that

$$\begin{aligned}
\frac{L'(\tau)}{2n} =& \frac{1}{n}\sum_{i=1}^n \Theta_i \left(-J\boldsymbol{W}_i + \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right)^\top \left(J\left(\tilde{Y}_i^{\text{obs}} - \boldsymbol{W}_i\tau\right) - \frac{\boldsymbol{\Gamma}_y}{\Gamma_\theta} + \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\tau\right) \\
=& \frac{1}{n}\left\{\sum_{i=1}^n \Theta_i \left(J\boldsymbol{W}_i - \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right)^\top \left(J\boldsymbol{W}_i - \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right)\right\}\tau - \frac{1}{n}\left\{\sum_{i=1}^n \Theta_i \left(J\boldsymbol{W}_i - \frac{\boldsymbol{\Gamma}_w}{\Gamma_\theta}\right)^\top \left(J\tilde{\boldsymbol{Y}}_i - \frac{\boldsymbol{\Gamma}_y}{\Gamma_\theta}\right)\right\} \\
=& \left\{\Gamma_{ww} - \frac{\boldsymbol{\Gamma}_w^\top \boldsymbol{\Gamma}_w}{\Gamma_\theta}\right\}\tau - \left\{\Gamma_{wy} - \frac{\boldsymbol{\Gamma}_w^\top \boldsymbol{\Gamma}_y}{\Gamma_\theta}\right\}
\end{aligned}$$

Since $L(\tau)$ is a convex quadratic function of τ, the first-order condition is sufficient and necessary to determine the optimality. The proof is then completed by solving $L'(\hat{\tau}) = 0$. □

B.1.3. Statistical properties of RIPW estimators with deterministic $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$.

B.1.3.1. *Asymptotic linear expansion of RIPW estimators.* As a warm-up, we assume $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)_{i=1}^n$ are deterministic. This, for example, includes the pure design-based inference where $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ and $\hat{\boldsymbol{m}}_i = 0$. In this case, the measures of accuracy can be simplified as

$$\Delta_{\pi i} = \sqrt{\mathbb{E}\left[\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i)\right]^2}, \quad \Delta_{yi} = \sqrt{\mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i\|_2^2]} + \|\boldsymbol{\tau}_i - \tau^* \mathbf{1}_T\|_2. \tag{B.1.5}$$

As a result, $(\Delta_{\pi i}, \Delta_{yi})$ are deterministic (conditional on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$).

We start by a lemma showing that $\Gamma_\theta, \Gamma_{wy}, \Gamma_{ww}, \boldsymbol{\Gamma}_w, \boldsymbol{\Gamma}_y$ concentrate around their means. For notational convenience, we let $\mathrm{Var}(Z)$ denote $\mathbb{E}\|Z - \mathbb{E}[Z]\|_2^2$ for a random vector Z.

Lemma B.1.2. *Under Assumptions B.1.2 and B.1.3,*

$$|\mathbb{E}[\Gamma_\theta]| + |\mathbb{E}[\Gamma_{wy}]| + |\mathbb{E}[\Gamma_{ww}]| + \|\mathbb{E}[\mathbf{\Gamma}_w]\|_2 + \|\mathbb{E}[\mathbf{\Gamma}_y]\|_2 = O(1),$$

and

$$\mathrm{Var}(\Gamma_\theta) + \mathrm{Var}(\Gamma_{wy}) + \mathrm{Var}(\Gamma_{ww}) + \mathrm{Var}(\mathbf{\Gamma}_w) + \mathrm{Var}(\mathbf{\Gamma}_y) = O(n^{-q}).$$

As a consequence,

$$\left|\Gamma_\theta - \mathbb{E}[\Gamma_\theta]\right| + \left|\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}]\right| + \left|\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}]\right| + \left\|\mathbf{\Gamma}_w - \mathbb{E}[\mathbf{\Gamma}_w]\right\|_2 + \left\|\mathbf{\Gamma}_y - \mathbb{E}[\mathbf{\Gamma}_y]\right\|_2 = O_{\mathbb{P}}\left(n^{-q/2}\right).$$

Proof. By Assumption B.1.2, $\Theta_i \le 1/c_\pi$ almost surely. Moreover, $\|\boldsymbol{W}_i\|_2 \le \sqrt{T}$ since $W_{it} \in \{0,1\}$. Thus,

$$\|\mathbf{\Gamma}_w\|_2 \le \frac{\sqrt{T}}{c_\pi}, \quad |\Gamma_{ww}| \le \frac{T}{c_\pi}, \quad |\Gamma_\theta| \le \frac{1}{c_\pi} \Longrightarrow \mathbb{E}\|\mathbf{\Gamma}_w\|_2 + \mathbb{E}|\Gamma_{ww}| + \mathbb{E}|\Gamma_\theta| = O(1).$$

Next, we derive bounds for $(\mathbb{E}[\Gamma_{wy}])^2$ and $\|\mathbb{E}[\mathbf{\Gamma}_y]\|_2^2$ separately. For $(\mathbb{E}[\Gamma_{wy}])^2$,

$$\begin{aligned}
(\mathbb{E}[\Gamma_{wy}])^2 &\le \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}]\right)^2 \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}]^2 \\
&\le \frac{1}{nc_\pi^2}\sum_{i=1}^n \mathbb{E}[\boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}]^2 \le \frac{T}{nc_\pi^2}\sum_{i=1}^n \mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 \\
&\le \frac{T}{nc_\pi^2}\sum_{i=1}^n \left\{\mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2\right\} \\
&= O(1),
\end{aligned}$$

where the last step follows from the Assumption B.1.3. For $\|\mathbb{E}[\mathbf{\Gamma}_y]\|_2^2$,

$$\begin{aligned}
\|\mathbb{E}[\mathbf{\Gamma}_y]\|_2^2 &\le \frac{1}{n}\sum_{i=1}^n (\mathbb{E}[\Theta_i J \tilde{Y}_i^{\mathrm{obs}}])^2 \le \frac{1}{nc_\pi^2}\sum_{i=1}^n \|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 \\
&\le \frac{1}{nc_\pi^2}\sum_{i=1}^n \left\{\mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2\right\} \\
&= O(1),
\end{aligned}$$

where the last step follows from the Assumption B.1.3. Putting the pieces together, the bound on the sum of expectations is proved.

Next, we turn to the bound on the variances. By Lemma B.1.1,

$$\mathrm{Var}(\Gamma_\theta) \le \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\Theta_i)\rho_i \le \frac{1}{n^2 c_\pi^2}\sum_{i=1}^n \rho_i.$$

The Assumption B.1.2 implies that

$$\frac{1}{n^2}\sum_{i=1}^n \rho_i = O(n^{-q}).$$

Therefore, $\mathrm{Var}(\Gamma_\theta) = O(n^{-q})$. For Γ_{ww},

$$\begin{aligned}\mathrm{Var}(\Gamma_{ww}) &\le \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i)\rho_i \le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}(\Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i)^2\rho_i \\ &\le \frac{T}{n^2 c_\pi^2}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{W}_i\|_2^2\rho_i \le \frac{T}{n^2 c_\pi^2}\sum_{i=1}^n \rho_i = O(n^{-q}),\end{aligned}$$

where the last equality uses the fact that $\|\boldsymbol{W}_i\|_2 \le \sqrt{T}$. For Γ_{wy},

$$\begin{aligned}\mathrm{Var}(\Gamma_{wy}) &\le \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}})\rho_i \le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}(\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}})^2\rho_i \\ &\overset{(i)}{\le} \frac{1}{n^2 c_\pi^2}\sum_{i=1}^n \mathbb{E}\Big[\|\boldsymbol{W}_i\|_2^2 \cdot \|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2\Big]\rho_i \\ &\overset{(ii)}{\le} \frac{T}{n^2 c_\pi^2}\sum_{i=1}^n \Big(\mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2\Big)\rho_i \\ &\overset{(iii)}{=} O(n^{-q}),\end{aligned}$$

where (i) follows from the Cauchy-Schwarz inequality and that $\|J\|_{\mathrm{op}} = 1$, (ii) is obtained from the fact that $\|\boldsymbol{W}_i\|_2^2 \le T$ and $\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}} \in \{\tilde{\boldsymbol{Y}}_i(1), \tilde{\boldsymbol{Y}}_i(0)\}$, and (iii) follows from the Assumption B.1.3.

For $\boldsymbol{\Gamma}_w$, recall that $\mathrm{Var}(\boldsymbol{\Gamma}_w)$ is the sum of the variance of each coordinate of $\boldsymbol{\Gamma}_w$. By Lemma B.1.1,

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\Gamma}_w) &\le \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\Theta_i J \boldsymbol{W}_i)\rho_i \le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\|\Theta_i J \boldsymbol{W}_i\|_2^2\rho_i \\ &\le \frac{1}{n^2 c_\pi^2}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{W}_i\|_2^2\rho_i \le \frac{T}{n^2 c_\pi^2}\sum_{i=1}^n \rho_i = O(n^{-q}).\end{aligned}$$

For $\mathbf{\Gamma}_y$, analogues to inequalities (i) - (iii) for Γ_{wy}, we obtain that

$$\begin{aligned}\mathrm{Var}(\mathbf{\Gamma}_y) &\le \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\Theta_i J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}})\rho_i \le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\|\Theta_i J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2\rho_i \\ &\le \frac{1}{n^2 c_\pi^2}\sum_{i=1}^n \left(\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2 + \|\tilde{\boldsymbol{Y}}_i(0)\|_2^2\right)\rho_i = O(n^{-q}),\end{aligned}$$

where the last step follows from the Assumption B.1.3.

Finally, by Markov's inequality,

$$\begin{aligned}&\left|\Gamma_\theta - \mathbb{E}[\Gamma_\theta]\right| + \left|\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}]\right| + \left|\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}]\right| + \left\|\mathbf{\Gamma}_w - \mathbb{E}[\mathbf{\Gamma}_w]\right\|_2 + \left\|\mathbf{\Gamma}_y - \mathbb{E}[\mathbf{\Gamma}_y]\right\|_2 \\ =&O_{\mathbb{P}}\left(\sqrt{\mathrm{Var}(\Gamma_\theta) + \mathrm{Var}(\Gamma_{wy}) + \mathrm{Var}(\Gamma_{ww}) + \mathrm{Var}(\mathbf{\Gamma}_w) + \mathrm{Var}(\mathbf{\Gamma}_y)}\right) = O_{\mathbb{P}}(n^{-q/2}).\end{aligned}$$

□

The following lemma shows that the denominator of $\hat{\tau}$ is bounded away from 0.

Lemma B.1.3. *Under Assumptions B.1.3, regardless of the dependence between $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ and the data,*

$$\mathcal{D} \ge c_{\mathcal{D}}^2\left(\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{W}_i = \boldsymbol{w}_1)\right)\left(\frac{1}{n}\sum_{i=1}^n I(\boldsymbol{W}_i = \boldsymbol{w}_2)\right),$$

for some constant $c_{\mathcal{D}}$ that only depends on $\mathbf{\Pi}$. As a result, $\mathcal{D} \ge 0$ almost surely. If Assumption B.1.2 also holds, [1]

$$\mathbb{E}[\mathcal{D}] \ge c_{\mathcal{D}}^2 c_\pi^2 - \frac{1}{n^2}\sum_{i=1}^n \rho_i, \quad \mathcal{D} \ge c_{\mathcal{D}}^2 c_\pi^2 - o_{\mathbb{P}}(1).$$

Proof. By definition,

$$\begin{aligned}\mathcal{D} &= \left(\frac{1}{n}\sum_{i=1}^n \Theta_i \tilde{\boldsymbol{W}}_i^\top J\tilde{\boldsymbol{W}}_i\right)\left(\frac{1}{n}\sum_{i=1}^n \Theta_i\right) - \left\|\frac{1}{n}\sum_{i=1}^n \Theta_i J\tilde{\boldsymbol{W}}_i\right\|_2^2 \\ &= \frac{1}{n^2}\sum_{i,j=1}^n \Theta_i\Theta_j\left(\tilde{\boldsymbol{W}}_i^\top J\tilde{\boldsymbol{W}}_i + \tilde{\boldsymbol{W}}_j^\top J\tilde{\boldsymbol{W}}_j - 2\tilde{\boldsymbol{W}}_i^\top J\tilde{\boldsymbol{W}}_j\right) \\ &= \frac{1}{n^2}\sum_{i,j=1}^n \Theta_i\Theta_j\|J(\boldsymbol{W}_i - \boldsymbol{W}_j)\|_2^2.\end{aligned}$$

[1] A more rigorous version of the second statement is $\max\{c_{\mathcal{D}}^2 c_\pi^2 - \mathcal{D}, 0\} = o_{\mathbb{P}}(1)$

Let $\boldsymbol{w}_1, \boldsymbol{w}_2$ be two distinct elements from $\mathbb{S}^*$ with $\boldsymbol{w}_1 \notin \{\mathbf{0}_T, \mathbf{1}_T\}$ and

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\pi}_i(\boldsymbol{w}_k) > c_\pi, \quad k \in \{1,2\}. \tag{B.1.6}$$

This is enabled by Assumption B.1.2. Note that $J(\boldsymbol{w}_1 - \boldsymbol{w}_2) = 0$ iff $\boldsymbol{w}_1 - \boldsymbol{w}_2 = a\mathbf{1}_T$ for some $a \in \mathbb{R}$, which is impossible since $\boldsymbol{w}_1 \notin \{\mathbf{0}_T, \mathbf{1}_T\}$ and all entries of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are binary. In addition, since $\boldsymbol{\Pi}$ has support $\mathbb{S}^*$, $\boldsymbol{\Pi}(\boldsymbol{w}_1), \boldsymbol{\Pi}(\boldsymbol{w}_2) > 0$. Let

$$c_{\mathcal{D}} = \min\{\boldsymbol{\Pi}(\boldsymbol{w}_1), \boldsymbol{\Pi}(\boldsymbol{w}_2)\}\|J(\boldsymbol{w}_1 - \boldsymbol{w}_2)\|_2 > 0.$$

Then

$$\begin{aligned}
\mathcal{D} &\geq \frac{c_{\mathcal{D}}^2}{n^2}\sum_{i,j=1}^{n} \frac{1}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)\hat{\boldsymbol{\pi}}_j(\boldsymbol{W}_j)} I(\boldsymbol{W}_i = \boldsymbol{w}_1, \boldsymbol{W}_j = \boldsymbol{w}_2) \\
&\geq \frac{c_{\mathcal{D}}^2}{n^2}\sum_{i,j=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1, \boldsymbol{W}_j = \boldsymbol{w}_2) \\
&= c_{\mathcal{D}}^2 \left(\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1)\right)\left(\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2)\right),
\end{aligned}$$

where the second inequality follows from the fact that $\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) \leq 1$. By (B.1.6),

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_k)\right] = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\pi}_i(\boldsymbol{w}_k) > c_\pi, \quad k \in \{1,2\}.$$

Furthermore, by Lemma B.1.1,

$$\begin{aligned}
&\left|\mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1), \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2)\right]\right| \\
&= \frac{1}{n^2}\left|\sum_{i,j=1}^{n} \mathrm{Cov}(I(\boldsymbol{W}_i = \boldsymbol{w}_1), I(\boldsymbol{W}_j = \boldsymbol{w}_2))\right| \\
&\leq \frac{1}{n^2}\sum_{i,j=1}^{n} \left|\mathrm{Cov}(I(\boldsymbol{W}_i = \boldsymbol{w}_1), I(\boldsymbol{W}_j = \boldsymbol{w}_2))\right| \\
&\leq \frac{1}{n^2}\sum_{i,j=1}^{n} \rho_{ij}\sqrt{\mathrm{Var}(I(\boldsymbol{W}_i = \boldsymbol{w}_1))\mathrm{Var}(I(\boldsymbol{W}_j = \boldsymbol{w}_2))} \\
&\leq \frac{1}{n^2}\sum_{i,j=1}^{n} \rho_{ij} = \frac{1}{n^2}\sum_{i=1}^{n} \rho_i.
\end{aligned}$$

Putting pieces together, we obtain that

$$\begin{aligned}
\mathbb{E}[\mathcal{D}] &\geq c_{\mathcal{D}}^2 \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1)\right)\left(\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2)\right)\right] \\
&= c_{\mathcal{D}}^2 \Bigg\{ \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1)\right] \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2)\right] \\
&\qquad + \mathrm{Cov}\left[\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_1), \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_2)\right] \Bigg\} \\
&\geq c_{\mathcal{D}}^2 c_\pi^2 - \frac{1}{n^2}\sum_{i=1}^{n} \rho_i.
\end{aligned}$$

On the other hand, by Lemma B.1.1, for $k \in \{1, 2\}$,

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_k)\right) \leq \frac{1}{n^2}\sum_{i=1}^{n} \rho_i = O(n^{-q}) = o(1).$$

By Markov's inequality, for $k \in \{1, 2\}$,

$$\frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{W}_i = \boldsymbol{w}_k) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(\boldsymbol{W}_i = \boldsymbol{w}_1) - o_{\mathbb{P}}(1) \geq c_\pi - o_{\mathbb{P}}(1).$$

Therefore,

$$\mathcal{D} \geq c_{\mathcal{D}}^2 (c_\pi - o_{\mathbb{P}}(1))(c_\pi - o_{\mathbb{P}}(1)) \geq c_{\mathcal{D}}^2 c_\pi^2 - o_{\mathbb{P}}(1).$$

□

Based on Lemma B.1.2 and B.1.3, we can derive an asymptotic linear expansion for the RIPW estimator.

Theorem B.1.2. *Under Assumptions B.1.2 and B.1.3,*

$$\mathcal{D}(\hat{\tau} - \tau^*) = N_* + \frac{1}{n}\sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + O_{\mathbb{P}}(n^{-q}),$$

where

$$N_* = \frac{1}{2n}\sum_{i=1}^{n} \mathbb{E}[\mathcal{V}_i] = \mathbb{E}[\Gamma_{wy}]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\boldsymbol{\Gamma}_y] - \tau^* \Big(\mathbb{E}[\Gamma_{ww}]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\boldsymbol{\Gamma}_w] \Big).$$

Furthermore,

$$\hat{\tau} - \tau^* = O_{\mathbb{P}}(|N_*|) + O_{\mathbb{P}}(n^{-q}).$$

Proof. Note that

$$\mathcal{D}(\hat{\tau}-\tau^*)=\mathcal{N}-\tau^*\mathcal{D}.$$

By Lemma B.1.2,

$$\begin{aligned}
&\left|(\Gamma_{wy}-\mathbb{E}[\Gamma_{wy}])(\Gamma_\theta-\mathbb{E}[\Gamma_\theta])\right|+\left|(\mathbf{\Gamma}_w-\mathbb{E}[\mathbf{\Gamma}_w])^\top(\mathbf{\Gamma}_y-\mathbb{E}[\mathbf{\Gamma}_y])\right|\\
\leq&\frac{1}{2}\left\{(\Gamma_{wy}-\mathbb{E}[\Gamma_{wy}])^2+(\Gamma_\theta-\mathbb{E}[\Gamma_\theta])^2+\|\mathbf{\Gamma}_w-\mathbb{E}[\mathbf{\Gamma}_w]\|_2^2+\|\mathbf{\Gamma}_y-\mathbb{E}[\mathbf{\Gamma}_y]\|_2^2\right\}\\
=&O_{\mathbb{P}}\left(\mathrm{Var}(\Gamma_{wy})+\mathrm{Var}(\Gamma_\theta)+\mathrm{Var}(\mathbf{\Gamma}_w)+\mathrm{Var}(\mathbf{\Gamma}_y)\right)=O_{\mathbb{P}}(n^{-q}).
\end{aligned}$$

Let

$$\mathcal{V}_{i1}=\Theta_i\left\{\mathbb{E}[\Gamma_{wy}]-\mathbb{E}[\mathbf{\Gamma}_y]^\top J\boldsymbol{W}_i+\mathbb{E}[\Gamma_\theta]\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}-\mathbb{E}[\mathbf{\Gamma}_w]^\top J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right\}.$$

Then,

$$\mathcal{N}=\mathbb{E}[\Gamma_{wy}]\mathbb{E}[\Gamma_\theta]-\mathbb{E}[\mathbf{\Gamma}_w]^\top\mathbb{E}[\mathbf{\Gamma}_y]+\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_{i1}-\mathbb{E}[\mathcal{V}_{i1}])+O_{\mathbb{P}}(n^{-q}),$$

Similarly,

$$\mathcal{D}=\mathbb{E}[\Gamma_{ww}]\mathbb{E}[\Gamma_\theta]-\mathbb{E}[\mathbf{\Gamma}_w]^\top\mathbb{E}[\mathbf{\Gamma}_w]+\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_{i2}-\mathbb{E}[\mathcal{V}_{i2}])+O_{\mathbb{P}}(n^{-q}),$$

where

$$\mathcal{V}_{i2}=\Theta_i\left\{\mathbb{E}[\Gamma_{ww}]-\mathbb{E}[\mathbf{\Gamma}_w]^\top J\boldsymbol{W}_i+\mathbb{E}[\Gamma_\theta]\boldsymbol{W}_i^\top J\boldsymbol{W}_i-\mathbb{E}[\mathbf{\Gamma}_w]^\top J\boldsymbol{W}_i\right\}.$$

Since $\mathcal{V}_i=\mathcal{V}_{i1}-\tau^*\mathcal{V}_{i2}$,

$$\mathcal{D}(\hat{\tau}-\tau^*)=\mathcal{N}-\tau^*\mathcal{D}=N_*+\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i-\mathbb{E}[\mathcal{V}_i])+O_{\mathbb{P}}(n^{-q}).$$

This proves the first statement.

Next, we prove the second statement on $\hat{\tau}-\tau^*$. By Lemma B.1.3, $1/\mathcal{D}=O_{\mathbb{P}}(1)$. It is left to show that

$$\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i-\mathbb{E}[\mathcal{V}_i])=o_{\mathbb{P}}(1).$$

Applying the inequality that $\mathrm{Var}(Z_1+Z_2)=2\mathrm{Var}(Z_1)+2\mathrm{Var}(Z_2)-\mathrm{Var}(Z_1-Z_2)\leq 2(\mathrm{Var}(Z_1)+\mathrm{Var}(Z_2))$, we obtain that

$$\frac{1}{4}\mathrm{Var}(\mathcal{V}_{i1})$$

$$
\begin{aligned}
&\leq \mathrm{Var}\left(\Theta_i \mathbb{E}[\Gamma_{wy}]\right) + \mathrm{Var}\left(\Theta_i \mathbb{E}[\Gamma_\theta] \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right) + \mathrm{Var}\left(\Theta_i \mathbb{E}[\boldsymbol{\Gamma}_w]^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right) + \mathrm{Var}\left(\Theta_i \mathbb{E}[\boldsymbol{\Gamma}_y]^\top J \boldsymbol{W}_i\right) \\
&\leq \mathbb{E}\left(\Theta_i \mathbb{E}[\Gamma_{wy}]\right)^2 + \mathbb{E}\left(\Theta_i \mathbb{E}[\Gamma_\theta] \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right)^2 + \mathbb{E}\left(\Theta_i \mathbb{E}[\boldsymbol{\Gamma}_w]^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right)^2 + \mathbb{E}\left(\Theta_i \mathbb{E}[\boldsymbol{\Gamma}_y]^\top J \boldsymbol{W}_i\right)^2 \\
&\overset{(i)}{\leq} \frac{1}{c_\pi^2}\left\{(\mathbb{E}[\Gamma_{wy}])^2 + (\mathbb{E}[\Gamma_\theta])^2 \mathbb{E}(\boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}})^2 + \mathbb{E}\left(\mathbb{E}[\boldsymbol{\Gamma}_w]^\top J \tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\right)^2 + \mathbb{E}\left(\mathbb{E}[\boldsymbol{\Gamma}_y]^\top J \boldsymbol{W}_i\right)^2\right\} \\
&\overset{(ii)}{\leq} \frac{1}{c_\pi^2}\left\{(\mathbb{E}[\Gamma_{wy}])^2 + (\mathbb{E}[\Gamma_\theta])^2 \mathbb{E}\|\boldsymbol{W}_i\|_2^2 \mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 + \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2^2 \mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 + \|\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2^2 \mathbb{E}\|\boldsymbol{W}_i\|_2^2\right\} \\
&\overset{(iii)}{\leq} \frac{1}{c_\pi^2}\left\{(\mathbb{E}[\Gamma_{wy}])^2 + T(\mathbb{E}[\Gamma_\theta])^2 \mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 + \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2^2 \mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2 + T\|\mathbb{E}[\boldsymbol{\Gamma}_y]\|_2^2\right\},
\end{aligned}
$$

where (i) follows from the Assumption B.1.2 that $\Theta_i \leq 1/c_\pi$ almost surely, (ii) follows from the Cauchy-Schwarz inequality and the fact that $\|J\|_{\mathrm{op}} = 1$, and (iii) follows from the fact that $\|\boldsymbol{W}_i\|_2^2 \leq T$. By Lemma B.1.2, we obtain that for all $i \in [n]$,

$$
\mathrm{Var}(\mathcal{V}_{i1}) \leq C_1\left(1 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}\|_2^2\right) \leq C_1\left(1 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2\right), \tag{B.1.7}
$$

for some constant C_1 that only depends on c_π and T. Similarly, we have that $\mathrm{Var}(\mathcal{V}_{i2}) \leq C_2$ for some constant C_2 that only depends on c_π and T. By Assumption B.1.3,

$$
\tau^* = \frac{1}{nT}\sum_{i=1}^n \sum_{t=1}^T \left(\mathbb{E}[\tilde{Y}_{it}(1)] - \mathbb{E}[\tilde{Y}_{it}(0)]\right) = O(1).
$$

Therefore,

$$
\mathrm{Var}(\mathcal{V}_i) \leq 2\mathrm{Var}(\mathcal{V}_{i1}) + 2(\tau^*)^2 \mathrm{Var}(\mathcal{V}_{i2}) \leq C\left(1 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^2 + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^2\right).
$$

for some constant C that only depends on c_π and T. Since $\mathcal{V}_i$ is a function of $\mathcal{Z}_i$, by Lemma B.1.1 and Assumption B.1.3,

$$
\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{V}_i\right) \leq \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\mathcal{V}_i)\rho_i = o(1).
$$

By Chebyshev's inequality,

$$
\frac{1}{n}\sum_{i=1}^n (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) = o_{\mathbb{P}}(1).
$$

The proof is then completed. □

B.1.3.2. *DATE equation and consistency.* Theorem B.1.2 shows that the asymptotic limit of $\mathcal{D}(\hat{\tau} - \tau^*)$ is N_*. For consistency, it remains to prove that $N_* = o(1)$. We start by proving that the asymptotic bias is zero when either the treatment or the outcome model is perfectly estimated.

Lemma B.1.4. *Under Assumptions B.1.1, B.1.2, and B.1.3, $N_* = 0$, if either (1) $\Delta_{yi} = 0$ for all $i \in [n]$, or (2) $\Delta_{\pi i} = 0$ for all $i \in [n]$, and $\mathbf{\Pi}$ satisfies the DATE equation* (3.2.13).

Proof. Without loss of generality, we assume that $\tau^* = 0$; otherwise, we replace $Y_{it}(1)$ by $Y_{it}(1) - \tau^*$ and the resulting $\hat{\tau}$ becomes $\hat{\tau} - \tau^*$. Then

$$N_* = \mathbb{E}[\Gamma_{wy}]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\mathbf{\Gamma}_w]^\top \mathbb{E}[\mathbf{\Gamma}_y].$$

It remains to prove that $N_* = 0$. Note that

$$\tilde{\boldsymbol{Y}}_i^{\text{obs}} = \tilde{\boldsymbol{Y}}_i(0) + \text{diag}(\boldsymbol{W}_i)\boldsymbol{\tau}_i.$$

Since $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ are deterministic, by Assumption B.1.1,

$$\begin{aligned}\mathbb{E}[\Gamma_{wy}] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J(\tilde{\boldsymbol{Y}}_i(0) + \text{diag}(\boldsymbol{W}_i)\boldsymbol{\tau}_i)] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i J \boldsymbol{W}_i]^\top \mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \,\text{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.\end{aligned}$$

Similarly,

$$\mathbb{E}[\mathbf{\Gamma}_y] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i] J \mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i \,\text{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.$$

As a result,

$$\begin{aligned}N_* = &\frac{1}{n}\sum_{i=1}^n \{\mathbb{E}[\Theta_i J \boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Theta_i]\mathbb{E}[\mathbf{\Gamma}_w]\}^\top \mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] \\ &+ \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \,\text{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\mathbf{\Gamma}_w]^\top \mathbb{E}[\Theta_i \,\text{diag}(\boldsymbol{W}_i)]\right\}\boldsymbol{\tau}_i. \qquad \text{(B.1.8)}\end{aligned}$$

If $\Delta_{yi} = 0$, $\hat{\boldsymbol{m}}_i = \boldsymbol{m}_i$ and $\boldsymbol{\tau}_i = 0$ (because $\tau^* = 0$). By (B.1.1), $\mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] = 0$. It is then obvious from (B.1.8) that $N_* = 0$.

If $\Delta_{\pi i} = 0$, $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ and thus for any function $f(\cdot)$,

$$\mathbb{E}[\Theta_i f(\boldsymbol{W}_i)] = \sum_{\boldsymbol{w}\in\{0,1\}^T} \frac{\mathbf{\Pi}(\boldsymbol{w})}{\boldsymbol{\pi}_i(\boldsymbol{w})} f(\boldsymbol{w})\boldsymbol{\pi}_i(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[f(\boldsymbol{W})].$$

As a result,

$$\mathbb{E}[\Theta_i J \boldsymbol{W}_i] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}] = \mathbb{E}[\boldsymbol{\Gamma}_w], \quad \mathbb{E}[\Theta_i] = 1 = \mathbb{E}[\Gamma_\theta],$$

and

$$\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W} J \operatorname{diag}(\boldsymbol{W})], \quad \mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\operatorname{diag}(\boldsymbol{W})].$$

Then

$$\mathbb{E}[\Theta_i J \boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Theta_i]\mathbb{E}[\boldsymbol{\Gamma}_w] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}] = 0,$$

and by DATE equation,

$$\begin{aligned}
&\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \operatorname{diag}(\boldsymbol{W})] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}]\xi^\top.
\end{aligned}$$

By (B.1.8),

$$\begin{aligned}
N_* &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}]\xi^\top \boldsymbol{\tau}_i \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}]\left(\frac{1}{n}\sum_{i=1}^n \xi^\top \boldsymbol{\tau}_i\right) \\
&= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}]\tau^* = 0.
\end{aligned}$$

□

Next, we prove a general bound for the asymptotic bias N_* as a function of $(\Delta_{yi}, \Delta_{\pi i})_{i=1}^n$.

Theorem B.1.3. *Let* $\boldsymbol{\Pi}$ *be an solution of the DATE equation* (3.2.13)*. Under Assumptions B.1.1, B.1.2, and B.1.3,*

$$|N_*| = O\left(\bar{\Delta}_\pi \bar{\Delta}_y\right).$$

Proof. As in the proof of Lemma B.1.4, we assume that $\tau^* = 0$. Let

$$\Theta_i^* = \frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)}{\boldsymbol{\pi}_i(\boldsymbol{W}_i)}, \quad \tilde{\boldsymbol{Y}}_i^{*\text{obs}} = \boldsymbol{Y}_i^{\text{obs}} - \boldsymbol{m}_i.$$

Further, let Γ_θ^* and $\boldsymbol{\Gamma}_w^*$ be the counterpart of Γ_θ and $\boldsymbol{\Gamma}_w$ with $(\Theta_i, \tilde{\boldsymbol{Y}}_i^{\text{obs}})$ replaced by $(\Theta_i^*, \tilde{\boldsymbol{Y}}_i^{*\text{obs}})$. For any function $f : \{0,1\}^T \mapsto \mathbb{R}$ such that $\mathbb{E}[f^2(\boldsymbol{W}_i)] \le C_1$ for some constant $C_1 > 0$, by Cauchy-Schwarz

inequality,

$$\mathbb{E}[\Theta_i f(\boldsymbol{W}_i) - \Theta_i^* f(\boldsymbol{W}_i)] = \mathbb{E}[(\Theta_i - \Theta_i^*) f(\boldsymbol{W}_i)] \le \sqrt{C_1}\sqrt{\mathbb{E}\left(\Theta_i - \Theta_i^*\right)^2}$$

$$= \sqrt{C_1}\sqrt{\mathbb{E}\left[\frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)^2}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)^2 \boldsymbol{\pi}_i(\boldsymbol{W}_i)^2}\left(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i)\right)^2\right]} \le \frac{\sqrt{C_1}}{c_\pi^2}\Delta_{\pi i}. \tag{B.1.9}$$

Thus, there exists a constant C_2 that only depends on c_π and T such that

$$|\mathbb{E}[\Theta_i] - \mathbb{E}[\Theta_i^*]| + \|\mathbb{E}[\Theta_i J \boldsymbol{W}_i] - \mathbb{E}[\Theta_i^* J \boldsymbol{W}_i]\|_2 + \|\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] - \mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\|_2$$
$$+ \|\mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)] - \mathbb{E}[\Theta_i^* \operatorname{diag}(\boldsymbol{W}_i)]\|_{\mathrm{op}} \le C_2 \Delta_{\pi i}.$$

By triangle inequality and Cauchy-Schwarz inequality, we also have

$$|\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Gamma_\theta^*]| + \|\mathbb{E}[\boldsymbol{\Gamma}_w] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]\| \le \frac{C_2}{n}\sum_{i=1}^n \Delta_{\pi i} \le C_2 \bar{\Delta}_\pi.$$

On the other hand, by Lemma B.1.2, there exists a constant C_3 that only depends on c_π and T,

$$|\mathbb{E}[\Gamma_\theta]| + \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2 \le C_3.$$

Without loss of generality, we assume that

$$C_3 \ge 1 + \|\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}]\|_2 = \mathbb{E}[\Theta_i^*] + \|\mathbb{E}[\Theta_i^* J \boldsymbol{W}_i]\|_2.$$

Putting pieces together,

$$\left|\mathbb{E}[\Theta_i J \boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Theta_i]\mathbb{E}[\boldsymbol{\Gamma}_w] - \left(\mathbb{E}[\Theta_i^* J \boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\Theta_i^*]\mathbb{E}[\boldsymbol{\Gamma}_w^*]\right)\right|$$
$$\le \left|\mathbb{E}[\Theta_i J \boldsymbol{W}_i] - \mathbb{E}[\Theta_i^* J \boldsymbol{W}_i]\right| \cdot \mathbb{E}[\Gamma_\theta] + \left|\mathbb{E}[\Theta_i] - \mathbb{E}[\Theta_i^*]\right| \cdot \|\mathbb{E}[\boldsymbol{\Gamma}_w]\|_2$$
$$+ \left|\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\Gamma_\theta^*]\right| \cdot \|\mathbb{E}[\Theta_i^* J \boldsymbol{W}_i]\|_2 + \left\|\mathbb{E}[\boldsymbol{\Gamma}_w] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]\right\| \cdot \mathbb{E}[\Theta_i^*]$$
$$\le 2C_3 C_2(\Delta_{\pi i} + \bar{\Delta}_\pi).$$

Similarly,

$$\Big|\mathbb{E}[\Theta_i \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta] - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top \mathbb{E}[\Theta_i \operatorname{diag}(\boldsymbol{W}_i)]$$
$$- \left(\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* \operatorname{diag}(\boldsymbol{W}_i)]\right)\Big|$$

$$\leq 2C_3C_2(\Delta_{\pi i} + \bar{\Delta}_\pi).$$

Let

$$N'_* = \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}[\Theta_i^* J \boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\Theta_i^*]\mathbb{E}[\boldsymbol{\Gamma}_w^*]\right\}^\top \mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)] + \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* \operatorname{diag}(\boldsymbol{W}_i)]\right\}\boldsymbol{\tau}_i.$$

Using the same arguments as in the proof of Lemma B.1.4,

$$\mathbb{E}[\Theta_i^* J \boldsymbol{W}_i]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\Theta_i^*]\mathbb{E}[\boldsymbol{\Gamma}_w^*] = 0,$$

and

$$\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}]\xi^\top.$$

Then

$$N'_* = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}]\xi^\top \boldsymbol{\tau}_i = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}]\tau^* = 0.$$

This entails that

$$|N_*| = |N_* - N'_*| \leq \frac{2C_3C_2}{n}\sum_{i=1}^n (\Delta_{\pi i} + \bar{\Delta}_\pi)(\|\mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)]\|_2 + \|\boldsymbol{\tau}_i\|_2).$$

By (B.1.1),

$$\|\mathbb{E}[\tilde{\boldsymbol{Y}}_i(0)]\|_2 = \mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i\|_2].$$

Since $(1/n)\sum_{i=1}^n \Delta_{yi} \leq \sqrt{(1/n)\sum_{i=1}^n \Delta_{yi}^2}$,

$$|N_*| \leq \frac{2C_3C_2}{n}\sum_{i=1}^n (\Delta_{\pi i} + \bar{\Delta}_\pi)\Delta_{yi} = 4C_3C_2\bar{\Delta}_\pi\bar{\Delta}_y.$$

The proof is then completed. □

B.1.3.3. *Doubly robust inference.* Theorem B.1.2 and Theorem B.1.3 imply the following properties of RIPW estimators.

Theorem B.1.4. *Let* $\mathbf{\Pi}$ *be an solution of the DATE equation* (3.2.13)*. Under Assumptions B.1.1, B.1.2, and B.1.3,*

$$\hat{\tau} - \tau^* = o_{\mathbb{P}}(1), \quad \text{if } \bar{\Delta}_\pi \bar{\Delta}_y = o(1).$$

If, further, $q > 1/2$ *in Assumption B.1.3 and* $\bar{\Delta}_\pi \bar{\Delta}_y = o(1/\sqrt{n})$*,*

$$\mathcal{D} \cdot \sqrt{n}(\hat{\tau} - \tau^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1).$$

Recalling (B.1.5) that $(\Delta_{\pi i}, \Delta_{yi})$ are deterministic, $\bar{\Delta}_\pi \bar{\Delta}_y = \mathbb{E}[\bar{\Delta}_\pi \bar{\Delta}_y]$. Since Assumptions B.1.2 and B.1.3 generalize Assumptions 3.2.1-3.2.3, Theorem B.1.4 implies Theorem 3.2.1 and 3.2.2. For doubly robust inference, Theorem 3.2.2 assumes that the unconditional expectation of $\bar{\Delta}_\pi \bar{\Delta}_y$ is $o(1)$ or $o(1/\sqrt{n})$. By Markov's inequality, it implies that $\mathbb{E}[\bar{\Delta}_\pi \bar{\Delta}_y] = o(1)$ or $o(1/\sqrt{n})$ with high probability (conditional on $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i = 1, \ldots, n\}$). Thus, Theorem 3.3.1 is also implied by Theorem B.1.4 since Assumptions B.1.2 - B.1.3 generalize Assumptions 3.3.2 - 3.3.4.

Throughout the rest of the subsection, we focus on the special case where $\{\mathcal{Z}_i : i \in [n]\}$ are independent. In this case, Assumption B.1.3 holds with $q = 1 > 1/2$ and thus the asymptotically linear expansion in Theorem B.1.4 holds. To obtain the asymptotic normality and a consistent variance estimator, we modify Assumption B.1.3 as follows.

Assumption B.1.4. $\{\mathcal{Z}_i : i = 1, \ldots, n\}$ *are independent (but not necessarily identically distributed), and there exists* $\omega > 0$ *such that*

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(1)\|_2^{2+\omega} + \mathbb{E}\|\tilde{\boldsymbol{Y}}_i(0)\|_2^{2+\omega} \right\} = O(1).$$

To derive the asymptotic normality of the RIPW estimator, we need the following assumption that prevents the variance from being too small.

Assumption B.1.5. *There exists* $\nu_0 > 0$ *such that*

$$\sigma^2 \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}(\mathcal{V}_i) \geq \nu_0.$$

The following lemma shows the asymptotic normality of the term $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])$.

Lemma B.1.5. *Then under Assumptions B.1.2, B.1.4, and B.1.5,*

$$d_K\left(\mathcal{L}\left(\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])\right), N(0,1)\right) \to 0,$$

where $\mathcal{L}(\cdot)$ denotes the probability law, d_K denotes the Kolmogorov-Smirnov distance (i.e., the ℓ_∞-norm of the difference of CDFs)

Proof. Since $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ are deterministic, by Assumption B.1.4, $\{\mathcal{V}_i : i \in [n]\}$ are independent. Recalling the definition of $\mathcal{V}_i$, it is easy to see that Assumption B.1.4 implies

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\mathcal{V}_i|^{2+\omega} = O(1). \tag{B.1.10}$$

By Assumption B.1.4,

$$\sum_{i=1}^{n}\mathbb{E}\left|\frac{\mathcal{V}_i}{\sqrt{n}\sigma}\right|^{2+\omega} = O\left(n^{-\omega/2}\right) = o(1).$$

The proof is completed by the Berry-Esseen inequality (Proposition B.1.1) with $g(x) = x^\omega$. □

Let $\hat{\mathcal{V}}_i$ denote the plug-in estimate of $\mathcal{V}_i$, i.e.,

$$\hat{\mathcal{V}}_i = \Theta_i\left\{\left(\Gamma_{wy} - \hat{\tau}\Gamma_{ww}\right) - \left(\boldsymbol{\Gamma}_y - \hat{\tau}\boldsymbol{\Gamma}_w\right)^\top J\boldsymbol{W}_i + \Gamma_\theta \boldsymbol{W}_i^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \hat{\tau}\boldsymbol{W}_i\right) - \boldsymbol{\Gamma}_w^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \hat{\tau}\boldsymbol{W}_i\right)\right\}. \tag{B.1.11}$$

We first prove that $\hat{\mathcal{V}}_i$ is an accurate approximation of $\mathcal{V}_i$ on average, even without the independence assumption.

Lemma B.1.6. *Let $\boldsymbol{\Pi}$ be a solution of the DATE equation. Under Assumptions B.1.1, B.1.2, and B.1.3,*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 = o_{\mathbb{P}}(1), \quad \text{if } \bar{\Delta}_\pi\bar{\Delta}_y = o(1).$$

Proof. Let

$$\hat{\mathcal{V}}_i' = \Theta_i\left\{\left(\Gamma_{wy} - \tau^*\Gamma_{ww}\right) - \left(\boldsymbol{\Gamma}_y - \tau^*\boldsymbol{\Gamma}_w\right)^\top J\boldsymbol{W}_i + \Gamma_\theta \boldsymbol{W}_i^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right) - \boldsymbol{\Gamma}_w^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right)\right\}.$$

Then

$$\hat{\mathcal{V}}_i - \hat{\mathcal{V}}_i' = (\hat{\tau} - \tau^*)\Theta_i\left\{-\Gamma_{ww} + \boldsymbol{\Gamma}_w^\top J\boldsymbol{W}_i - \Gamma_\theta \boldsymbol{W}_i^\top J\boldsymbol{W}_i + \boldsymbol{\Gamma}_w^\top J\boldsymbol{W}_i\right\}.$$

Under Assumption B.1.2, there exists a constant C that only depends on c_π and T such that

$$|\hat{\mathcal{V}}_i - \hat{\mathcal{V}}_i'| \le C|\hat{\tau} - \tau^*|.$$

By Theorem B.1.4,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \hat{\mathcal{V}}_i')^2 = O((\hat{\tau} - \tau^*)^2) = o_{\mathbb{P}}(1) \tag{B.1.12}$$

Next,

$$\begin{aligned}\hat{\mathcal{V}}_i' - \mathcal{V}_i = \Theta_i\Bigg\{&\Big((\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}]) - \tau^*(\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}])\Big) - \Big((\mathbf{\Gamma}_y - \mathbb{E}[\mathbf{\Gamma}_y]) - \tau^*(\mathbf{\Gamma}_w - \mathbb{E}[\mathbf{\Gamma}_w])\Big)^\top J\boldsymbol{W}_i \\ &+ (\Gamma_\theta - \mathbb{E}[\Gamma_\theta])\boldsymbol{W}_i^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right) - (\mathbf{\Gamma}_w - \mathbb{E}[\mathbf{\Gamma}_w])^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right)\Bigg\}.\end{aligned}$$

By Jensen's inequality and Assumption B.1.2,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \mathcal{V}_i)^2 \\ &\le \frac{5}{nc_\pi^2}\sum_{i=1}^{n}\Bigg\{(\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}])^2 + (\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}])^2\cdot\tau^{*2} + \|(\mathbf{\Gamma}_y - \mathbb{E}[\mathbf{\Gamma}_y])\|_2^2\cdot\|J\boldsymbol{W}_i\|_2^2 \\ &\qquad + \|(\mathbf{\Gamma}_w - \mathbb{E}[\mathbf{\Gamma}_w])\|_2^2\cdot\|J(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - 2\tau^*\boldsymbol{W}_i)\|_2^2 + (\Gamma_\theta - \mathbb{E}[\Gamma_\theta])^2\left(\boldsymbol{W}_i^\top J\left(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\right)\right)^2\Bigg\} \\ &= \frac{5}{c_\pi^2}\Bigg\{(\Gamma_{wy} - \mathbb{E}[\Gamma_{wy}])^2 + (\Gamma_{ww} - \mathbb{E}[\Gamma_{ww}])^2\cdot\tau^{*2} + \|(\mathbf{\Gamma}_y - \mathbb{E}[\mathbf{\Gamma}_y])\|_2^2\cdot T \\ &\qquad + \|(\mathbf{\Gamma}_w - \mathbb{E}[\mathbf{\Gamma}_w])\|_2^2\cdot\frac{1}{n}\sum_{i=1}^{n}\|(\tilde{\boldsymbol{Y}}_i^{\text{obs}} - 2\tau^*\boldsymbol{W}_i)\|_2^2 \\ &\qquad + \|(\Gamma_\theta - \mathbb{E}[\Gamma_\theta])\|_2^2\cdot\frac{T}{n}\sum_{i=1}^{n}\|\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \tau^*\boldsymbol{W}_i\|_2^2\Bigg\}.\end{aligned}$$

By Lemma B.1.2,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \mathcal{V}_i)^2\right] = o(1).$$

By Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \mathcal{V}_i)^2 = o_{\mathbb{P}}(1). \tag{B.1.13}$$

Putting (B.1.12) and (B.1.13) together, we obtain that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 \le \frac{2}{n}\sum_{i=1}^{n}\{(\hat{\mathcal{V}}_i - \hat{\mathcal{V}}_i')^2 + (\hat{\mathcal{V}}_i' - \mathcal{V}_i)^2\} = o_{\mathbb{P}}(1).$$

□

As in Section 3.2, we estimate the (conservative) variance of the term $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])$ as

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mathcal{V}}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2\right\}. \tag{B.1.14}$$

This yields a Wald-type confidence interval for DATE,

$$\hat{C}_{1-\alpha} = [\hat{\tau} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}, \hat{\tau} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}], \tag{B.1.15}$$

where z_η is the η-th quantile of the standard normal distribution.

Theorem B.1.5. *Assume that $\bar{\Delta}_\pi\bar{\Delta}_y = o(1/\sqrt{n})$. Under Assumptions B.1.1, B.1.2, B.1.4, and B.1.5,*

$$\liminf_{n\to\infty}\mathbb{P}\left(\tau^* \in \hat{C}_{1-\alpha}\right) \ge 1-\alpha.$$

Proof. By Theorem B.1.2, Theorem B.1.3, Lemma B.1.5, and Assumption B.1.5,

$$\frac{\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*)}{\sigma} = \frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1) \xrightarrow{d} N(0,1) \text{ in Kolmogorov-Smirnov distance,}$$

As a result,

$$\left|\mathbb{P}\left(\left|\frac{\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*)}{\sigma}\right| \le z_{1-\alpha/2}\cdot\frac{\hat{\sigma}}{\sigma}\right) - \left\{2\Phi\left(z_{1-\alpha/2}\cdot\frac{\hat{\sigma}}{\sigma}\right) - 1\right\}\right| = o(1), \tag{B.1.16}$$

where Φ is the cumulative distribution function of the standard normal distribution. Let

$$\sigma_+^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\mathcal{V}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i]\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i^2] - \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i]\right)^2. \tag{B.1.17}$$

Clearly, σ_+^2 is deterministic and

$$\sigma_+^2 = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\mathcal{V}_i] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{V}_i]\right)^2 \ge \sigma^2.$$

It remains to show that

$$\left|\frac{n-1}{n}\hat{\sigma}^2 - \sigma_+^2\right| = o_{\mathbb{P}}(1). \tag{B.1.18}$$

In fact, by Assumption B.1.5, (B.1.18) implies that

$$\sqrt{\frac{n-1}{n}}\frac{\hat{\sigma}}{\sigma} \xrightarrow{p} \frac{\sigma_+}{\sigma} \geq 1 \Longrightarrow \frac{\hat{\sigma}}{\sigma} \xrightarrow{p} \frac{\sigma_+}{\sigma} \geq 1.$$

By continuous mapping theorem,

$$2\Phi\left(z_{1-\alpha/2}\cdot\frac{\hat{\sigma}}{\sigma}\right) - 1 \xrightarrow{p} 2\Phi\left(z_{1-\alpha/2}\cdot\frac{\sigma_+}{\sigma}\right) - 1 \geq 1-\alpha,$$

which completes the proof.

Now we prove (B.1.18). By Proposition B.1.2 and Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i^2 - \mathbb{E}[\mathcal{V}_i^2])\right|^{1+\omega/2} &\leq \frac{2}{n^{1+\omega/2}}\sum_{i=1}^n \mathbb{E}|\mathcal{V}_i^2 - \mathbb{E}[\mathcal{V}_i^2]|^{1+\omega/2} \\ \leq \frac{2^{1+\omega/2}}{n^{1+\omega/2}}\sum_{i=1}^n\left(\mathbb{E}[|\mathcal{V}_i|^{2+\omega}] + \mathbb{E}[\mathcal{V}_i^2]^{1+\omega/2}\right) &\leq \frac{2^{2+\omega/2}}{n^{1+\omega/2}}\sum_{i=1}^n \mathbb{E}\left[|\mathcal{V}_i|^{2+\omega}\right].\end{aligned}$$

By (B.1.10),

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i^2 - \mathbb{E}[\mathcal{V}_i^2])\right|^{1+\omega/2} = o(1).$$

By Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i^2 - \mathbb{E}[\mathcal{V}_i^2]) = o_{\mathbb{P}}(1). \tag{B.1.19}$$

Similarly, we have that

$$\frac{1}{n}\sum_{i=1}^n(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) = o_{\mathbb{P}}(1). \tag{B.1.20}$$

In addition, (B.1.10) and Hölder's inequality imply that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i^2] = O(1), \quad \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathcal{V}_i] = O(1).$$

As a result,

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i^2 = O_{\mathbb{P}}(1), \quad \frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i = O_{\mathbb{P}}(1). \tag{B.1.21}$$

By Lemma B.1.6, (B.1.21), and Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 - \frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i^2\right| &\le \frac{2}{n}\sum_{i=1}^{n}\mathcal{V}_i|\hat{\mathcal{V}}_i - \mathcal{V}_i| + \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 \\ &\le 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2} + \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 = o_{\mathbb{P}}(1). \end{aligned} \tag{B.1.22}$$

Similarly,

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i\right)^2\right| = o_{\mathbb{P}}(1). \tag{B.1.23}$$

By (B.1.19), (B.1.20), and (B.1.21),

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{V}_i\right)^2 - \sigma_+^2\right| = o_{\mathbb{P}}(1). \tag{B.1.24}$$

Putting (B.1.22) - (B.1.24) together, we complete the proof of (B.1.18). □

B.1.4. Doubly robust inference with deterministic $(\hat{\pi}_i, \hat{m}_i)$ and dependent assignments. Recall Theorem B.1.4 that

$$\mathcal{D}\cdot\sqrt{n}(\hat{\tau} - \tau^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1).$$

This is true even when $\mathcal{Z}_i$'s are dependent as long as Assumption B.1.3 holds. If $\mathcal{V}_i$'s are observable, a valid confidence interval for τ^* can be derived if the distribution of $(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])$ can be approximated. Specifically, assume that

$$\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\sqrt{(1/n)\mathrm{Var}[\sum_{i=1}^{n}\mathcal{V}_i]}} \xrightarrow{d} N(0,1), \tag{B.1.25}$$

and there exists a conservative oracle variance estimator $\hat{\sigma}^{*2}$ based on $(\mathcal{V}_1, \ldots, \mathcal{V}_n)$ in the sense that

$$\frac{(1/n)\mathrm{Var}[\sum_{i=1}^{n}\mathcal{V}_i]}{\hat{\sigma}^{*2}} \le 1 + o_{\mathbb{P}}(1). \tag{B.1.26}$$

Then, $[\hat{\tau} - z_{1-\alpha/2}\hat{\sigma}^*/\sqrt{n}\mathcal{D}, \hat{\tau} + z_{1-\alpha/2}\hat{\sigma}^*/\sqrt{n}\mathcal{D}]$ is an asymptotically valid confidence interval for τ^*. Of course, this interval cannot be computed in practice because $\mathcal{V}_i$ is unobserved due to the unknown quantities including $\mathbb{E}[\Gamma_\theta], \mathbb{E}[\mathbf{\Gamma}_w], \mathbb{E}[\mathbf{\Gamma}_y], \mathbb{E}[\Gamma_{ww}]$, $\mathbb{E}[\Gamma_{wy}]$, and τ^*. A natural variance estimator can be obtained by replacing $\boldsymbol{\mathcal{V}} \triangleq (\mathcal{V}_1, \ldots, \mathcal{V}_n)$ with $\hat{\boldsymbol{\mathcal{V}}} \triangleq (\hat{\mathcal{V}}_1, \ldots, \hat{\mathcal{V}}_n)$ in $\hat{\sigma}^{*2}$. The following theorem makes this intuition rigorous for generic quadratic oracle variance estimators.

Theorem B.1.6. *Suppose there exists an oracle variance estimator $\hat{\sigma}^{*2}$ such that*

(i) $\hat{\sigma}^{*2} = \boldsymbol{\mathcal{V}}^\top \boldsymbol{A}_n \boldsymbol{\mathcal{V}}/n$ *for some positive semidefinite (and potentially random) matrix* $\boldsymbol{A}_n$ *with* $\|\boldsymbol{A}_n\|_{\text{op}} = O_{\mathbb{P}}(1)$*;*

(ii) $\hat{\sigma}^{*2}$ *is conservative in the sense that, for every* η *in a neighborhood of* α*,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^n (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}^*}\right| \geq z_{1-\eta/2}\right) \leq \eta;$$

(iii) $1/\hat{\sigma}^{*2} = O_{\mathbb{P}}(1)$.

Let $\hat{\sigma}^2 = \hat{\boldsymbol{\mathcal{V}}}^\top \boldsymbol{A}_n \hat{\boldsymbol{\mathcal{V}}}/n$ *and*

$$\hat{C}_{1-\alpha} = [\hat{\tau} - z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}, \hat{\tau} + z_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\mathcal{D}].$$

Under Assumptions B.1.1, B.1.2, and B.1.3 with $q > 1/2$*, if* $\mathbf{\Pi}$ *be an solution of the DATE equation* (3.2.13) *and* $\bar{\Delta}_\pi \bar{\Delta}_y = o(1/\sqrt{n})$*,*

$$\liminf_{n\to\infty} \mathbb{P}\left(\tau^* \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha.$$

Proof. By Lemma B.1.6,

$$\frac{1}{n}\|\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}}\|_2^2 = \frac{1}{n}\sum_{i=1}^n (\hat{\mathcal{V}}_i - \mathcal{V}_i)^2 = o_{\mathbb{P}}(1).$$

Since $\boldsymbol{A}_n$ is positive semidefinite, for any $\epsilon \in (0,1)$,

$$(1-\epsilon)\hat{\sigma}^{*2} - \left(\frac{1}{\epsilon} - 1\right)\frac{1}{n}(\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}})^\top \boldsymbol{A}_n (\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}}) \leq \hat{\sigma}^2 \leq (1+\epsilon)\hat{\sigma}^{*2} + \left(\frac{1}{\epsilon} + 1\right)\frac{1}{n}(\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}})^\top \boldsymbol{A}_n (\hat{\boldsymbol{\mathcal{V}}} - \boldsymbol{\mathcal{V}})$$

Thus, for any $\epsilon \in (0,1)$,

$$\mathbb{P}\left(\hat{\sigma}^2 \notin [(1-\epsilon)\hat{\sigma}^{*2}, (1+\epsilon)\hat{\sigma}^{*2}]\right) = o(1).$$

By condition (iii), the above result implies that

$$\left|\frac{\hat{\sigma}}{\hat{\sigma}^*} - 1\right| = o_{\mathbb{P}}(1). \tag{B.1.27}$$

By Theorem B.1.4,

$$\mathcal{D} \cdot \sqrt{n}(\hat{\tau} - \tau^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i]) + o_{\mathbb{P}}(1).$$

It remains to show that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}}\right| \geq z_{1-\alpha/2}\right) \leq \alpha.$$

Let $\eta(\epsilon)$ be the quantity such that $z_{1-\eta(\epsilon)/2} = z_{1-\alpha/2} \cdot (1-\epsilon)$. For any sufficiently small ϵ such that $\eta(\epsilon)$ lies in the neighborhood of α in condition (ii),

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}}\right| \geq z_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}^*}\right| \geq z_{1-\alpha/2} \cdot \frac{\hat{\sigma}}{\hat{\sigma}^*}\right) \\
&\leq \mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}^*}\right| \geq z_{1-\eta(\epsilon)/2}\right) + \mathbb{P}\left(\frac{\hat{\sigma}}{\hat{\sigma}^*} \leq 1-\epsilon\right).
\end{aligned}$$

By (B.1.27), when n tends to infinity,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{(1/\sqrt{n})\sum_{i=1}^{n}(\mathcal{V}_i - \mathbb{E}[\mathcal{V}_i])}{\hat{\sigma}}\right| \geq z_{1-\alpha/2}\right) \leq \eta(\epsilon).$$

The proof is completed by letting $\epsilon \to 0$ and noting that $\lim_{\epsilon\to 0} \eta(\epsilon) = \alpha$. □

When $\boldsymbol{W}_i$'s are independent,

$$\hat{\sigma}^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathcal{V}_i - \bar{\mathcal{V}})^2.$$

Thus, $\boldsymbol{A}_n = (n/(n-1))(I_n - \mathbf{1}_n\mathbf{1}_n^T/n)$. Clearly, the condition (i) is satisfied because $\|\boldsymbol{A}_n\|_{\text{op}} = n/(n-1)$. Under the assumptions in Theorem B.1.5, the condition (ii) is satisfied. Moreover, we have shown that $\hat{\sigma}^{*2}$ converges to $\sigma_+^2 \geq \sigma^2 > 0$, and thus the condition (iii) is satisfied. Therefore, Theorem B.1.5 can be implied by Theorem B.1.6.

When $\mathcal{V}_i$'s are observed, the variance estimators are quadratic under nearly all types of dependent assignment mechanisms. This includes completely randomized experiments [Hoeffding, 1951, Li and Ding, 2017], blocked and matched experiments [Pashley and Miratrix, 2021], two-stage

randomized experiments [Ohlsson, 1989], and so on. Below, we prove the results for completely randomized experiments with fixed potential outcomes to illustrate how to apply Theorem B.1.6. The notation is chosen to mimic Theorem 5 and Proposition 3 in Li and Ding [2017].

THEOREM B.1.7. *Assume that* $(Y_{it}(1), Y_{it}(0))$ *are fixed, and* $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i$ *as in Section 3.2 (while* $\hat{m}_{it}$ *is allowed to be non-zero). Consider a completely randomized experiments where the treatment assignments are sampled without replacement from* Q *possible assignments* $\{\boldsymbol{w}_{[1]}, \ldots, \boldsymbol{w}_{[Q]}\}$ *with* n_q *units assigned* $\boldsymbol{w}_{[q]}$*. Let* $\boldsymbol{\Pi}$ *be a solution of the DATE equation* (3.2.13) *with support* $\{\boldsymbol{w}_{[1]}, \ldots, \boldsymbol{w}_{[Q]}\}$*, and* $\mathcal{V}_i(q)$ *be the "potential outcome" for* $\mathcal{V}_i$ *where* $(Y_{it}^{\text{obs}}, W_{it})$ *is replaced by* $(Y_{it}(\boldsymbol{w}_{[q],t}), \boldsymbol{w}_{[q],t})$*, i.e.,*

$$\mathcal{V}_i(q) = \frac{\boldsymbol{\Pi}(\boldsymbol{w}_{[q]})}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}_{[q]})} \Bigg\{ \Big(\mathbb{E}[\Gamma_{wy}] - \tau^* \mathbb{E}[\Gamma_{ww}] \Big) - \Big(\mathbb{E}[\boldsymbol{\Gamma}_y] - \tau^* \mathbb{E}[\boldsymbol{\Gamma}_w] \Big)^\top J \boldsymbol{w}_{[q]} + \mathbb{E}[\Gamma_\theta] \boldsymbol{w}_{[q]}^\top J \Big(\tilde{\boldsymbol{Y}}_i(q) - \tau^* \boldsymbol{w}_{[q]} \Big) - \mathbb{E}[\boldsymbol{\Gamma}_w]^\top J \Big(\tilde{\boldsymbol{Y}}_i(q) - \tau^* \boldsymbol{w}_{[q]} \Big) \Bigg\},$$

and $\tilde{\boldsymbol{Y}}_i(q) = \Big(Y_{i1}(\boldsymbol{w}_{[q],1}) - \hat{m}_{i1}, Y_{i2}(\boldsymbol{w}_{[q],2}) - \hat{m}_{i2}, \ldots, Y_{iT}(\boldsymbol{w}_{[q],T}) - \hat{m}_{iT} \Big)$*. Further, for any* $q, r = 1, \ldots, Q$*, let*

$$S_q^2 = \frac{1}{n-1} \sum_{i=1}^n (\mathcal{V}_i(q) - \bar{\mathcal{V}}(q))^2, \quad S_{qr} = \frac{1}{n-1} \sum_{i=1}^n (\mathcal{V}_i(q) - \bar{\mathcal{V}}(q))(\mathcal{V}_i(r) - \bar{\mathcal{V}}(r)),$$

where $\bar{\mathcal{V}}(q) = (1/n) \sum_{i=1}^n \mathcal{V}_i(q)$*. Define the variance estimate* $\hat{\sigma}^2$ *as*

$$\hat{\sigma}^2 = \sum_{q=1}^Q \frac{n_q}{n} s_q^2, \quad \text{where } s_q^2 = \frac{1}{n_q - 1} \sum_{i: \boldsymbol{w}_i = \boldsymbol{w}_{[q]}} (\hat{\mathcal{V}}_i - \hat{\bar{\mathcal{V}}}(q))^2, \ \hat{\bar{\mathcal{V}}}(q) = \frac{1}{n_q} \sum_{i: \boldsymbol{w}_i = \boldsymbol{w}_{[q]}} \hat{\mathcal{V}}_i.$$

Further, define the confidence interval as

$$\hat{C}_{1-\alpha} = [\hat{\tau} - z_{1-\alpha/2} \hat{\sigma} / \sqrt{n} \mathcal{D}, \hat{\tau} + z_{1-\alpha/2} \hat{\sigma} / \sqrt{n} \mathcal{D}].$$

Assume that

(a) $Q = O(1)$ *and* $n_q / n \to \pi_q$ *for some constant* $\pi_q > 0$*;*

(b) for any $q, r = 1, \ldots, Q$*,* S_q^2 *and* S_{qr} *have limiting values* S_q^{*2}, S_{qr}^**;*

(c) there exists a constant $c_\tau > 0$ *such that* $\sum_{q=1}^Q \pi_q S_q^{*2} > c_\tau$*;*

(d) there exists a constant $M < \infty$ *such that* $\max_{i,q} \{ \| \tilde{\boldsymbol{Y}}_i(q) - \hat{\boldsymbol{m}}_i \|_2 \} < M$*.*

Then,

$$\liminf_{n \to \infty} \mathbb{P}\Big(\tau^* \in \hat{C}_{1-\alpha} \Big) \geq 1 - \alpha.$$

Proof. By definition, for any $i \neq j \in [n]$ and $q \neq r \in [Q]$,

$$\mathbb{P}(\boldsymbol{W}_i = \boldsymbol{w}_{[q]}) = \frac{n_q}{n}, \quad \mathbb{P}(\boldsymbol{W}_i = \boldsymbol{W}_j = \boldsymbol{w}_{[q]}) = \frac{n_q(n_q-1)}{n(n-1)}, \quad \mathbb{P}(\boldsymbol{W}_i = \boldsymbol{w}_{[q]}, \boldsymbol{W}_j = \boldsymbol{w}_{[r]}) = \frac{n_q n_r}{n(n-1)}.$$

For any functions f and g on $[0,1]^T$,

$$\mathbb{E}[f(\boldsymbol{W}_i)] = \sum_{q=1}^{Q} \frac{n_q}{n} f(\boldsymbol{w}_{[q]}), \quad \mathbb{E}[g(\boldsymbol{W}_j)] = \sum_{q=1}^{Q} \frac{n_q}{n} g(\boldsymbol{w}_{[q]}),$$

$$\mathbb{E}[f^2(\boldsymbol{W}_i)] = \sum_{q=1}^{Q} \frac{n_q}{n} f^2(\boldsymbol{w}_{[q]}), \quad \mathbb{E}[g^2(\boldsymbol{W}_j)] = \sum_{q=1}^{Q} \frac{n_q}{n} g^2(\boldsymbol{w}_{[q]}),$$

and

$$\mathbb{E}[f(\boldsymbol{W}_i)g(\boldsymbol{W}_i)] = \sum_{q=1}^{Q} \frac{n_q}{n} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}),$$

$$\mathbb{E}[f(\boldsymbol{W}_i)g(\boldsymbol{W}_j)] = \sum_{q=1}^{Q} \frac{n_q(n_q-1)}{n(n-1)} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}) + \sum_{q\neq r} \frac{n_q n_r}{n(n-1)} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[r]}).$$

As a result, for any $i \neq j$

$$\begin{aligned}
&\mathrm{Cov}(f(\boldsymbol{W}_i), g(\boldsymbol{W}_j)) = \mathbb{E}[f(\boldsymbol{W}_i)g(\boldsymbol{W}_j)] - \mathbb{E}[f(\boldsymbol{W}_i)]\mathbb{E}[g(\boldsymbol{W}_j)] \\
&= \sum_{q=1}^{Q} \left(\frac{n_q(n_q-1)}{n(n-1)} - \frac{n_q^2}{n^2} \right) f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}) + \sum_{q\neq r} \left(\frac{n_q n_r}{n(n-1)} - \frac{n_q n_r}{n^2} \right) f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[r]}) \\
&= \sum_{q=1}^{Q} -\frac{n_q(n-n_q)}{n^2(n-1)} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}) + \sum_{q\neq r} \frac{n_q n_r}{n^2(n-1)} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[r]}) \\
&= -\frac{1}{n-1} \sum_{q=1}^{Q} \frac{n_q}{n} f(\boldsymbol{w}_{[q]})g(\boldsymbol{w}_{[q]}) + \frac{1}{n-1} \left(\sum_{q=1}^{Q} \frac{n_q}{n} f(\boldsymbol{w}_{[q]}) \right) \left(\sum_{q=1}^{Q} \frac{n_q}{n} g(\boldsymbol{w}_{[r]}) \right) \\
&= -\frac{1}{n-1} \left(\mathbb{E}[f(\boldsymbol{W}_i)g(\boldsymbol{W}_i)] - \mathbb{E}[f(\boldsymbol{W}_i)]\mathbb{E}[g(\boldsymbol{W}_i)] \right) \\
&= -\frac{1}{n-1} \mathrm{Cov}(f(\boldsymbol{W}_i), g(\boldsymbol{W}_i))
\end{aligned}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}
&|\mathrm{Cov}(f(\boldsymbol{W}_i), g(\boldsymbol{W}_j))| \leq \frac{1}{n-1} |\mathrm{Cov}(f(\boldsymbol{W}_i), g(\boldsymbol{W}_i))| \\
&\leq \frac{1}{n-1} \sqrt{\mathrm{Var}[f(\boldsymbol{W}_i)]\mathrm{Var}[g(\boldsymbol{W}_i)]} = \frac{1}{n-1} \sqrt{\mathrm{Var}[f(\boldsymbol{W}_i)]\mathrm{Var}[g(\boldsymbol{W}_j)]}.
\end{aligned}$$

This implies that

$$\rho_{ij} \le \frac{1}{n-1} \Longrightarrow \boldsymbol{\rho}_i \le 2.$$

It is then clear that Assumption B.1.3 holds under the condition (d). Further, since $\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}_{[q]}) = \boldsymbol{\pi}_i(\boldsymbol{w}_{[q]}) = n_q/n$, the condition (a) implies Assumption B.1.2 and that $\bar{\Delta}_\pi \bar{\Delta}_y = 0$. On the other hand, Assumption B.1.1 holds because $\boldsymbol{W}_i$ is completely randomized. Therefore, it remains to check the condition (i) - (iii) in Theorem B.1.6 with

$$\hat{\sigma}^{*2} = \sum_{q=1}^{Q} \frac{n_q}{n} s_q^{*2}, \quad \text{where } s_q^{*2} = \frac{1}{n_q - 1} \sum_{i:\boldsymbol{w}_i = \boldsymbol{w}_{[q]}} (\mathcal{V}_i - \bar{\mathcal{V}}(q))^2,\ \bar{\mathcal{V}}(q) = \frac{1}{n_q} \sum_{i:\boldsymbol{w}_i = \boldsymbol{w}_{[q]}} \mathcal{V}_i.$$

In this case, $\boldsymbol{A}_n$ is a block-diagonal matrix with

$$\boldsymbol{A}_{n,\mathcal{I}_q,\mathcal{I}_q} = \frac{n_q}{n_q - 1} \left(I_{n_q} - \frac{\mathbf{1}_{n_q} \mathbf{1}_{n_q}^\top}{n_q} \right),$$

where $\mathcal{I}_q = \{i : \boldsymbol{W}_i = \boldsymbol{w}_{[q]}\}$. As a result,

$$\|\boldsymbol{A}_n\|_{\text{op}} = \max_q \frac{n_q}{n_q - 1} = O(1).$$

Thus, the condition (i) holds. The condition (ii) is implied by Proposition 3 in Li and Ding [2017] and the condition (iii) is implied by the condition (c). The theorem is then imply by Theorem B.1.6. □

B.1.5. Statistical properties of RIPW estimators with cross-fitted $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$. In this section, we consider the K-fold cross-fitting where K is treated as a constant. Let $\{\mathcal{I}_k : k = 1, \ldots, K\}$ denote the index sets of each fold, each with size $m \in \{\lfloor n/K \rfloor, \lceil n/K \rceil\}$. For convenience, we assume that $m = n/K$ is an integer. All proofs in this subsection can be easily extended to the general case.

For each $i \in \mathcal{I}_k$, $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ are estimated using $\{\mathcal{Z}_i : i \notin \mathcal{I}_k\}$. When $\{\mathcal{Z}_i : i \in [n]\}$ are independent, it is obvious that

$$\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i \in \mathcal{I}_k\} \perp\!\!\!\perp \{\mathcal{Z}_i : i \in \mathcal{I}_k\}.$$

We use a superscript (k) to denote the corresponding quantity in fold k, i.e.,

$$\Gamma_\theta^{(k)} \triangleq \frac{1}{m} \sum_{i \in \mathcal{I}_k} \Theta_i, \quad \Gamma_{ww}^{(k)} \triangleq \frac{1}{m} \sum_{i \in \mathcal{I}_k} \Theta_i \boldsymbol{W}_i^\top J \boldsymbol{W}_i, \quad \Gamma_{wy}^{(k)} \triangleq \frac{1}{m} \sum_{i \in \mathcal{I}_k} \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}},$$

$$\mathbf{\Gamma}_w^{(k)} \triangleq \frac{1}{m}\sum_{i\in\mathcal{I}_k}\Theta_i J\boldsymbol{W}_i, \quad \mathbf{\Gamma}_y^{(k)} \triangleq \frac{1}{m}\sum_{i\in\mathcal{I}_k}\Theta_i J\tilde{\boldsymbol{Y}}_i^{\text{obs}}.$$

To prove the asymptotic properties of RIPW estimators with cross-fitting, we state an analogy of Assumption 3.3.5 below.

Assumption B.1.6. *There exist deterministic functions* $\{\boldsymbol{\pi}_i' : i \in [n]\}$ *which satisfy* (B.1.3), *and vectors* $\{\boldsymbol{m}_i' : i \in [n]\}$ *such that*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2] + \mathbb{E}[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2]\right\} = O(n^{-r})$$

for some $r > 0$. *Furthermore,*

$$\text{either } \boldsymbol{\pi}_i' = \boldsymbol{\pi}_i \text{ for all } i, \quad \text{or} \quad \boldsymbol{m}_i' = \boldsymbol{m}_i \text{ for all } i.$$

Remark B.1.1. *Without loss of generality, we can assume that*

$$\mathbb{E}[\bar{\Delta}_\pi^2] = \Omega(n^{-r}), \quad \mathbb{E}[\bar{\Delta}_y^2] = \Omega(n^{-r}), \tag{B.1.28}$$

where $a_n = \Omega(b_n)$ *iff* $b_n = O(a_n)$. *Otherwise, we can replace* $\boldsymbol{\pi}_i'$ *by* $\boldsymbol{\pi}_i$ *or* $\boldsymbol{m}_i'$ *by* $\boldsymbol{m}_i$ *to increase* r.

Theorem B.1.8. *Let* $\{(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) : i = 1, \ldots, n\}$ *be estimates obtained from* K*-fold cross-fitting. Under Assumption B.1.1, B.1.2, B.1.4, and B.1.6,*

(i) $\hat{\tau} - \tau^* = o_{\mathbb{P}}(1)$ *if* $\sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_y^2]} = o(1)$;

(ii) $\liminf_{n\to\infty} \mathbb{P}\left(\tau^* \in \hat{C}_{1-\alpha}\right) \geq 1 - \alpha$ *if (1)* $\sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_y^2]} = o(1/\sqrt{n})$, *(2) Assumption B.1.5 holds when* $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i) = (\boldsymbol{\pi}_i', \boldsymbol{m}_i')$, *and (3) Assumption B.1.6 holds with* $r > 1/2$.

Proof. As in the proof of Theorem B.1.3, we assume $\tau^* = 0$ without loss of generality. Let $\left(\Gamma_{wy}', \Gamma_\theta', \mathbf{\Gamma}_w', \mathbf{\Gamma}_y'\right)$ and $(\Theta_i', \tilde{\boldsymbol{Y}}_i'^{\text{obs}})$ be the counterpart of $(\Gamma_{wy}, \Gamma_\theta, \mathbf{\Gamma}_w, \mathbf{\Gamma}_y)$ and $(\Theta_i, \tilde{\boldsymbol{Y}}_i^{\text{obs}})$ with $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ replaced by $(\boldsymbol{\pi}_i', \boldsymbol{m}_i')$. We first claim that

$$\Gamma_{wy}\Gamma_\theta - \mathbf{\Gamma}_w^\top \mathbf{\Gamma}_y - \left\{\Gamma_{wy}'\Gamma_\theta' - \mathbf{\Gamma}_w'^\top \mathbf{\Gamma}_y'\right\} = O_{\mathbb{P}}\left(n^{-\min\{r,(r'+1)/2\}} + \sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_y^2]}\right), \tag{B.1.29}$$

where $r' = r\omega/(2+\omega)$. The proof of (B.1.29) is relegated to the end. Here we prove the rest of the theorem under (B.1.29).

Note that $\Gamma'_{wy}\Gamma'_\theta - \mathbf{\Gamma}'^\top_w \mathbf{\Gamma}'_y$ is the numerator of $\hat{\tau}$ when $\{(\boldsymbol{\pi}'_i, \boldsymbol{m}'_i) : i = 1, \ldots, n\}$ are used as the estimates. Let

$$\Delta'_{\pi i} = \sqrt{\mathbb{E}[(\boldsymbol{\pi}'_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i(\boldsymbol{W}_i))^2]}, \quad \Delta'_{yi} = \sqrt{\mathbb{E}[\|\boldsymbol{m}'_i - \boldsymbol{m}_i\|_2^2]} + \|\boldsymbol{\tau}_i\|_2, \tag{B.1.30}$$

and

$$\bar{\Delta}'_\pi = \sqrt{\frac{1}{n}\sum_{i=1}^n \Delta'^2_{\pi i}}, \quad \bar{\Delta}'_y = \sqrt{\frac{1}{n}\sum_{i=1}^n \Delta'^2_{yi}}. \tag{B.1.31}$$

By Assumption B.1.6 and (B.1.28) in Remark (B.1.1),

$$\begin{aligned}
\bar{\Delta}'^2_\pi &= \frac{1}{n}\sum_{i=1}^n \Delta'^2_{\pi i} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Delta'^2_{\pi i}] \\
&\leq \frac{2}{n}\sum_{i=1}^n \left\{\mathbb{E}[\Delta^2_{\pi i}] + \mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}'_i(\boldsymbol{W}_i))^2]\right\} \\
&= O\left(\mathbb{E}[\bar{\Delta}^2_\pi] + n^{-r}\right) = O\left(\mathbb{E}[\bar{\Delta}^2_\pi]\right). \qquad \text{(B.1.32)}
\end{aligned}$$

Similarly,

$$\bar{\Delta}'^2_y = \frac{1}{n}\sum_{i=1}^n \Delta'^2_{yi} = O\left(\mathbb{E}[\bar{\Delta}^2_y]\right). \tag{B.1.33}$$

As a result,

$$\bar{\Delta}'_\pi \bar{\Delta}'_y = O\left(\sqrt{\mathbb{E}[\bar{\Delta}^2_\pi]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}^2_y]}\right).$$

Note that Assumption B.1.4 implies Assumption B.1.3 with $q = 1$. By Theorem B.1.2 and Theorem B.1.3,

$$\Gamma'_{wy}\Gamma'_\theta - \mathbf{\Gamma}'^\top_w \mathbf{\Gamma}'_y = \frac{1}{n}\sum_{i=1}^n (\mathcal{V}'_i - \mathbb{E}[\mathcal{V}'_i]) + O_{\mathbb{P}}\left(\sqrt{\mathbb{E}[\bar{\Delta}^2_\pi]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}^2_y]}\right) + o_{\mathbb{P}}(1/\sqrt{n}) \tag{B.1.34}$$

$$= O_{\mathbb{P}}\left(\sqrt{\mathbb{E}[\bar{\Delta}^2_\pi]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}^2_y]}\right) + o_{\mathbb{P}}(1), \tag{B.1.35}$$

where

$$\mathcal{V}'_i = \Theta'_i \left\{\mathbb{E}[\Gamma'_{wy}] - \mathbb{E}[\mathbf{\Gamma}'_y]^\top J \boldsymbol{W}_i + \mathbb{E}[\Gamma'_\theta] \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}} - \mathbb{E}[\mathbf{\Gamma}'_w]^\top J \tilde{\boldsymbol{Y}}_i'^{\text{obs}}\right\}.$$

On the other hand, by (B.1.29),

$$\mathcal{D}(\hat{\tau}-\tau^*)=\Gamma_{wy}\Gamma_\theta-\mathbf{\Gamma}_w^\top\mathbf{\Gamma}_y=\Gamma'_{wy}\Gamma'_\theta-\mathbf{\Gamma}_w^{'\top}\mathbf{\Gamma}'_y+O_{\mathbb{P}}\left(n^{-\min\{r,(r'+1)/2\}}+\sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}_y^2]}\right). \tag{B.1.36}$$

When $\sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}_y^2]}=o(1)$, (B.1.35) and (B.1.36) imply that

$$\mathcal{D}(\hat{\tau}-\tau^*)=o_{\mathbb{P}}(1).$$

The consistency then follows from Lemma B.1.3.

When $\sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}_y^2]}=o(1/\sqrt{n})$ and $r>1/2$, (B.1.35) and (B.1.36) imply that

$$\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*)=\frac{1}{\sqrt{n}}\sum_{i=1}^n(\mathcal{V}'_i-\mathbb{E}[\mathcal{V}'_i])+o_{\mathbb{P}}(1).$$

Let $\hat{\mathcal{V}}'_i$ denote the plug-in estimate of $\mathcal{V}'_i$ assuming that $(\boldsymbol{\pi}'_i,\boldsymbol{m}'_i)$ is known, i.e.,

$$\hat{\mathcal{V}}'_i=\Theta'_i\left\{\Gamma'_{wy}-\mathbf{\Gamma}_y^{'\top}J\boldsymbol{W}_i+\Gamma'_\theta\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}-\mathbf{\Gamma}_w^{'\top}J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\right\}. \tag{B.1.37}$$

By Lemma B.1.5, under Assumption B.1.5 (with $(\hat{\boldsymbol{\pi}}_i,\hat{\boldsymbol{m}}_i)=(\boldsymbol{\pi}'_i,\boldsymbol{m}'_i)$),

$$\frac{\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^*)}{\sigma'}\xrightarrow{d}N(0,1)\text{ in Kolmogorov-Smirnov distance,}$$

where

$$\sigma^{'2}=\frac{1}{n}\sum_{i=1}^n\text{Var}(\mathcal{V}'_i)\geq\nu_0.$$

Similar to (B.1.17), define

$$\sigma_+^{'2}=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left(\mathcal{V}'_i-\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\mathcal{V}'_i]\right)^2=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\mathcal{V}_i^{'2}]-\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\mathcal{V}'_i]\right)^2.$$

Obviously, $\sigma_+^{'2}\geq\sigma^{'2}$. Furthermore, define an oracle variance estimate $\hat{\sigma}^{'2}$ as

$$\hat{\sigma}^{'2}=\frac{1}{n-1}\sum_{i=1}^n\left(\hat{\mathcal{V}}'_i-\frac{1}{n}\sum_{i=1}^n\hat{\mathcal{V}}'_i\right)^2=\frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^n\hat{\mathcal{V}}_i^{'2}-\left(\frac{1}{n}\sum_{i=1}^n\hat{\mathcal{V}}'_i\right)^2\right\}.$$

Recalling (B.1.14) that

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mathcal{V}}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2 = \frac{n}{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2\right\}.$$

Similar to (B.1.18) in Theorem B.1.5, it remains to prove that

$$|\hat{\sigma}^2 - \sigma_+^{'2}| = o_{\mathbb{P}}(1).$$

Using the same arguments as in Theorem B.1.5, we can prove that

$$|\hat{\sigma}^{'2} - \sigma_+^{'2}| = o_{\mathbb{P}}(1).$$

Therefore, the proof will be completed if

$$|\hat{\sigma}^2 - \hat{\sigma}^{'2}| = o_{\mathbb{P}}(1). \tag{B.1.38}$$

We present the proof of (B.1.38) in the end.

Proof of (B.1.29)**:** Let $\left(\Gamma_{wy}^{'(k)}, \Gamma_{\theta}^{'(k)}, \mathbf{\Gamma}_{w}^{'(k)}, \mathbf{\Gamma}_{y}^{'(k)}\right)$ be the counterpart of $(\Gamma_{wy}^{(k)}, \Gamma_{\theta}^{(k)}, \mathbf{\Gamma}_{w}^{(k)}, \mathbf{\Gamma}_{y}^{(k)})$ with $(\hat{\boldsymbol{\pi}}_i, \hat{\boldsymbol{m}}_i)$ replaced by $(\tilde{\boldsymbol{\pi}}_i, \tilde{\boldsymbol{m}}_i)$. Since the proof is lengthy, we decompose it into seven steps.

Step 1 By triangle inequality and Cauchy-Schwarz inequality,

$$\begin{aligned}
&|\Gamma_{wy} - \Gamma'_{wy}| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}|\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta'_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{'\text{obs}}| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\Theta_i \boldsymbol{W}_i^\top J(\hat{\boldsymbol{m}}_i - \boldsymbol{m}'_i)\right| + \frac{1}{n}\sum_{i=1}^{n}\left|(\Theta_i - \Theta'_i)\boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{'\text{obs}}\right| \\
&\leq \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\|\Theta_i \boldsymbol{W}_i^\top J\|_2^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}'_i\|_2^2\right)} \\
&\quad + \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\|(\Theta_i - \Theta'_i)\boldsymbol{W}_i^\top J\|_2^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2^2\right)}.
\end{aligned}$$

By Assumption B.1.5 and Hölder's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2^2] \le \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2^{2+\omega}]\right)^{2/(2+\omega)} = O(1).$$

By Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2^2 = O_{\mathbb{P}}(1). \tag{B.1.39}$$

By Assumption B.1.2 and the boundedness of $\|\boldsymbol{W}_i J\|_2$,

$$\frac{1}{n}\sum_{i=1}^{n}\|\Theta_i \boldsymbol{W}_i^\top J\|_2^2 = O(1),$$

and, further, by Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\|(\Theta_i - \Theta_i')\boldsymbol{W}_i^\top J\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2]\right).$$

Putting pieces together and using Assumption B.1.6, we arrive at

$$|\Gamma_{wy} - \Gamma_{wy}'| = O_{\mathbb{P}}(n^{-r/2}).$$

Similarly, we can prove that

$$|\Gamma_{wy} - \Gamma_{wy}'| + |\Gamma_\theta - \Gamma_\theta'| + \|\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}_w'\|_2 + \|\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}_y'\|_2 = O_{\mathbb{P}}(n^{-r/2}). \tag{B.1.40}$$

As a consequence,

$$\left|(\Gamma_{wy} - \Gamma_{wy}')(\Gamma_\theta - \Gamma_\theta') - (\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}_w')^\top(\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}_y')\right| = O_{\mathbb{P}}(n^{-r}). \tag{B.1.41}$$

Step 2 Note that Assumption B.1.4 implies Assumption B.1.3 with $q = 1$. By Lemma B.1.2,

$$\left|\Gamma_\theta' - \mathbb{E}[\Gamma_\theta']\right| + \left|\Gamma_{wy}' - \mathbb{E}[\Gamma_{wy}']\right| + \left\|\boldsymbol{\Gamma}_w' - \mathbb{E}[\boldsymbol{\Gamma}_w']\right\|_2 + \left\|\boldsymbol{\Gamma}_y' - \mathbb{E}[\boldsymbol{\Gamma}_y']\right\|_2 = O_{\mathbb{P}}\left(n^{-1/2}\right).$$

By (B.1.40), we have

$$\begin{aligned}&\Big|(\Gamma_{wy} - \Gamma_{wy}')(\Gamma_\theta' - \mathbb{E}[\Gamma_\theta']) + (\Gamma_{wy}' - \mathbb{E}[\Gamma_{wy}'])(\Gamma_\theta - \Gamma_\theta') \\ &- (\boldsymbol{\Gamma}_w - \boldsymbol{\Gamma}_w')^\top(\boldsymbol{\Gamma}_y' - \mathbb{E}[\boldsymbol{\Gamma}_y']) - (\boldsymbol{\Gamma}_w' - \mathbb{E}[\boldsymbol{\Gamma}_w'])^\top(\boldsymbol{\Gamma}_y - \boldsymbol{\Gamma}_y')\Big| = O_{\mathbb{P}}(n^{-(r+1)/2}).\end{aligned} \tag{B.1.42}$$

Step 3 Note that

$$\Gamma_{wy} - \Gamma'_{wy} = \frac{1}{K}\sum_{k=1}^{K}\left(\Gamma_{wy}^{(k)} - \Gamma_{wy}^{'(k)}\right).$$

For each k,

$$\Gamma_{wy}^{(k)} - \Gamma_{wy}^{'(k)} = \frac{1}{m}\sum_{i\in\mathcal{I}_k}(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta_i'\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}).$$

Under Assumption B.1.4, the summands are independent conditional on $\mathcal{D}_{-[k]} \triangleq \{\mathcal{Z}_i : i \notin \mathcal{I}_k\}$. Let $\mathbb{E}^{(k)}$ and $\text{Var}^{(k)}$ denote the expectation and variance conditional on $\mathcal{D}_{-k}$ (and $\{(\boldsymbol{X}_i, \boldsymbol{U}_i) : i \in [n]\}$). By Chebyshev's inequality,

$$\begin{aligned}
&\left(\Gamma_{wy}^{(k)} - \Gamma_{wy}^{'(k)} - \mathbb{E}^{(k)}[\Gamma_{wy}^{(k)} - \Gamma_{wy}^{'(k)}]\right)^2 \\
&= O_{\mathbb{P}}\left(\frac{1}{m^2}\sum_{i\in\mathcal{I}_k}\text{Var}^{(k)}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta_i'\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\right)\right) \\
&\overset{(i)}{=} O_{\mathbb{P}}\left(\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}^{(k)}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta_i'\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\right)^2\right) \\
&\overset{(ii)}{=} O_{\mathbb{P}}\left(\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta_i'\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\right)^2\right),
\end{aligned} \tag{B.1.43}$$

where (i) follows from $K = O(1)$ and (ii) applies Markov's inequality. By Jensen's inequality and Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\left(\Theta_i \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{\text{obs}} - \Theta_i'\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\right)^2 \\
&\le 2\left\{\mathbb{E}\left(\Theta_i \boldsymbol{W}_i^\top J(\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i')\right)^2 + \mathbb{E}\left((\Theta_i - \Theta_i')\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\right)^2\right\} \\
&\le 2\left\{\mathbb{E}\left[\|\Theta_i \boldsymbol{W}_i^\top J\|_2^2(\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i')^2\right] + \mathbb{E}\left[(\Theta_i - \Theta_i')^2(\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}})^2\right]\right\} \\
&\le C\left\{\mathbb{E}\left[(\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i')^2\right] + \mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2(\tilde{\boldsymbol{Y}}_i^{'\text{obs}})^2\right]\right\},
\end{aligned} \tag{B.1.44}$$

where C is a constant that only depends on c_π and T. The second term can be bounded by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2(\tilde{\boldsymbol{Y}}_i^{'\text{obs}})^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2(\tilde{\boldsymbol{Y}}_i^{'\text{obs}})^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)-\boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2(1+2/\omega)}\right)^{\omega/(2+\omega)}\left(\frac{1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{Y}}_i^{'\mathrm{obs}})^{2+\omega}\right)^{2/(2+\omega)}\right]\\
&\overset{(ii)}{\leq} \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)-\boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2(1+2/\omega)}\right]\right)^{\omega/(2+\omega)}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\tilde{\boldsymbol{Y}}_i^{'\mathrm{obs}})^{2+\omega}\right]\right)^{2/(2+\omega)}\\
&\overset{(iii)}{\leq} \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)-\boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2}\right)^{\omega/(2+\omega)}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\tilde{\boldsymbol{Y}}_i^{'\mathrm{obs}})^{2+\omega}\right]\right)^{2/(2+\omega)},
\end{aligned}
$$

where (i) applies the Hölder's inequality for sums, (ii) applies the Hölder's inequality that $\mathbb{E}[XY] \leq \mathbb{E}[X^{(2+\omega)/\omega}]^{\omega/(2+\omega)}\mathbb{E}[Y^{(2+\omega)/2}]^{2/(2+\omega)}$, and (iii) uses the fact that $|\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)-\boldsymbol{\pi}_i'(\boldsymbol{W}_i)| \leq 1$. By Assumptions B.1.5 and B.1.6,

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)-\boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2(\tilde{\boldsymbol{Y}}_i^{'\mathrm{obs}})^2\right] = O\left(n^{-r\omega/(2+\omega)}\right) = O\left(n^{-r'}\right) \tag{B.1.45}
$$

(B.1.44) and (B.1.45) together imply that

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\Theta_i\boldsymbol{W}_i^{\top}J\tilde{\boldsymbol{Y}}_i^{\mathrm{obs}}-\Theta_i'\boldsymbol{W}_i^{\top}J\tilde{\boldsymbol{Y}}_i^{'\mathrm{obs}}\right)^2 = O\left(n^{-r'}\right). \tag{B.1.46}
$$

By (B.1.43), for each k,

$$
\Gamma_{wy}^{(k)}-\Gamma_{wy}^{'(k)}-\mathbb{E}^{(k)}[\Gamma_{wy}^{(k)}-\Gamma_{wy}^{'(k)}] = O_{\mathbb{P}}\left(n^{-(r'+1)/2}\right).
$$

Since $K = O(1)$, it implies that

$$
\left|\Gamma_{wy}-\Gamma_{wy}^{'}-\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma_{wy}^{(k)}-\Gamma_{wy}^{'(k)}]\right| = O_{\mathbb{P}}\left(n^{-(r'+1)/2}\right).
$$

Similarly, we have

$$
\begin{aligned}
&\left|\Gamma_{\theta}-\Gamma_{\theta}^{'}-\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma_{\theta}^{(k)}-\Gamma_{\theta}^{'(k)}]\right|+\left\|\boldsymbol{\Gamma}_w-\boldsymbol{\Gamma}_w^{'}-\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_w^{(k)}-\boldsymbol{\Gamma}_w^{'(k)}]\right\|_2\\
&+\left\|\boldsymbol{\Gamma}_y-\boldsymbol{\Gamma}_y^{'}-\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_y^{(k)}-\boldsymbol{\Gamma}_y^{'(k)}]\right\|_2 = O_{\mathbb{P}}\left(n^{-(r'+1)/2}\right).
\end{aligned}
$$

By Lemma B.1.2,

$$
|\mathbb{E}[\Gamma_{\theta}']|+|\mathbb{E}[\Gamma_{wy}']|+\|\mathbb{E}[\boldsymbol{\Gamma}_w']\|_2+\|\mathbb{E}[\boldsymbol{\Gamma}_y']\|_2 = O(1).
$$

Therefore,

$$\begin{aligned}
&\left|\left(\Gamma_{wy}-\Gamma'_{wy}-\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma_{wy}^{(k)}-\Gamma_{wy}^{'(k)}]\right)\mathbb{E}[\Gamma'_{\theta}]\right.\\
&+\mathbb{E}[\Gamma'_{wy}]\left(\Gamma_{\theta}-\Gamma'_{\theta}-\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\Gamma_{\theta}^{(k)}-\Gamma_{\theta}^{'(k)}]\right)\\
&-\left(\boldsymbol{\Gamma}_{w}-\boldsymbol{\Gamma}'_{w}-\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_{w}^{(k)}-\boldsymbol{\Gamma}_{w}^{'(k)}]\right)^{\top}\mathbb{E}[\boldsymbol{\Gamma}'_{y}]\\
&\left.-\mathbb{E}[\boldsymbol{\Gamma}'_{w}]^{\top}\left(\boldsymbol{\Gamma}_{y}-\boldsymbol{\Gamma}'_{y}-\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_{y}^{(k)}-\boldsymbol{\Gamma}_{y}^{'(k)}]\right)\right|\\
&=O_{\mathbb{P}}(n^{-(r'+1)/2}).
\end{aligned}\tag{B.1.47}$$

Step 4 Note that $\mathbb{E}[\Gamma'_{wy}]\mathbb{E}[\Gamma'_{\theta}]-\mathbb{E}[\boldsymbol{\Gamma}'_{w}]^{\top}\mathbb{E}[\boldsymbol{\Gamma}'_{y}]$ is the limit of $\mathcal{D}\cdot\sqrt{n}(\hat{\tau}-\tau^{*})$ when $\{(\boldsymbol{\pi}'_{i},\boldsymbol{m}'_{i}):i=1,\ldots,n\}$ are plugged in as the estimates. Under Assumption B.1.6, either $\boldsymbol{\pi}'_{i}=\boldsymbol{\pi}_{i}$ for all $i\in[n]$ or $\boldsymbol{m}'_{i}=\boldsymbol{m}_{i}$ for all $i\in[n]$. Then, by Lemma B.1.4,

$$\mathbb{E}[\Gamma'_{wy}]\mathbb{E}[\Gamma'_{\theta}]-\mathbb{E}[\boldsymbol{\Gamma}'_{w}]^{\top}\mathbb{E}[\boldsymbol{\Gamma}'_{y}]=0.\tag{B.1.48}$$

Step 5 We shall prove that

$$\frac{1}{K}\sum_{k=1}^{K}\left|\mathbb{E}^{(k)}[\Gamma_{wy}^{(k)}]\mathbb{E}[\Gamma'_{\theta}]-\mathbb{E}[\boldsymbol{\Gamma}'_{w}]^{\top}\mathbb{E}^{(k)}[\boldsymbol{\Gamma}_{y}^{(k)}]\right|=O\left(\sqrt{\mathbb{E}[\bar{\Delta}_{\pi}^{2}]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}_{y}^{2}]}\right).\tag{B.1.49}$$

By definition, we can write

$$\Delta_{\pi i}=\sqrt{\mathbb{E}^{(k)}[(\hat{\boldsymbol{\pi}}_{i}(\boldsymbol{W}_{i})-\boldsymbol{\pi}_{i}(\boldsymbol{W}_{i}))^{2}]},\quad \Delta_{yi}=\sqrt{\mathbb{E}^{(k)}[\|\hat{\boldsymbol{m}}_{i}-\boldsymbol{m}_{i}\|_{2}^{2}]}+\|\boldsymbol{\tau}_{i}-\tau^{*}\mathbf{1}_{T}\|_{2},\quad \forall i\in\mathcal{I}_{k}.$$

By Assumption B.1.1 and B.1.4,

$$\begin{aligned}
\mathbb{E}^{(k)}[\Gamma_{wy}^{(k)}]&=\frac{1}{m}\sum_{i\in\mathcal{I}_{k}}\mathbb{E}^{(k)}[\Theta_{i}\boldsymbol{W}_{i}^{\top}J\tilde{\boldsymbol{Y}}_{i}^{\mathrm{obs}}]\\
&=\frac{1}{m}\sum_{i\in\mathcal{I}_{k}}\mathbb{E}^{(k)}[\Theta_{i}\boldsymbol{W}_{i}^{\top}J\tilde{\boldsymbol{Y}}_{i}(0)]+\frac{1}{m}\sum_{i\in\mathcal{I}_{k}}\mathbb{E}^{(k)}[\Theta_{i}\boldsymbol{W}_{i}^{\top}J\,\mathrm{diag}(\boldsymbol{W}_{i})\boldsymbol{\tau}_{i}]\\
&=\frac{1}{m}\sum_{i\in\mathcal{I}_{k}}\mathbb{E}^{(k)}[\Theta_{i}\boldsymbol{W}_{i}^{\top}J]\mathbb{E}^{(k)}[\tilde{\boldsymbol{Y}}_{i}(0)]+\frac{1}{m}\sum_{i\in\mathcal{I}_{k}}\mathbb{E}^{(k)}[\Theta_{i}\boldsymbol{W}_{i}^{\top}J\,\mathrm{diag}(\boldsymbol{W}_{i})]\boldsymbol{\tau}_{i}.
\end{aligned}$$

Similarly,

$$\mathbb{E}^{(k)}[\mathbf{\Gamma}_y^{(k)}] = \frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_i J]\mathbb{E}^{(k)}[\tilde{\mathbf{Y}}_i(0)] + \frac{1}{m}\sum_{i\in\mathcal{I}_k}\mathbb{E}^{(k)}[\Theta_i J\,\mathrm{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.$$

Putting the pieces together and using the fact that $\mathbb{E}[\mathbf{\Gamma}_w']^\top J = \mathbb{E}[\mathbf{\Gamma}_w']^\top, \mathbb{E}^{(k)}[\mathbf{\Gamma}_w]^\top J = \mathbb{E}^{(k)}[\mathbf{\Gamma}_w]^\top$,

$$\begin{aligned}
&\mathbb{E}^{(k)}[\Gamma_{wy}^{(k)}]\mathbb{E}[\Gamma_\theta'] - \mathbb{E}[\mathbf{\Gamma}_w']^\top\mathbb{E}^{(k)}[\mathbf{\Gamma}_y^{(k)}] \\
&= \frac{1}{m}\sum_{i\in\mathcal{I}_k}\left\{\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta'] - \mathbb{E}[\mathbf{\Gamma}_w']^\top\mathbb{E}^{(k)}[\Theta_i J\,\mathrm{diag}(\boldsymbol{W}_i)]\right\}\boldsymbol{\tau}_i. \\
&+ \frac{1}{m}\sum_{i\in\mathcal{I}_k}\left\{\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J]\mathbb{E}[\Gamma_\theta'] - \mathbb{E}[\mathbf{\Gamma}_w']^\top\mathbb{E}^{(k)}[\Theta_i]\right\}\mathbb{E}^{(k)}[\tilde{\mathbf{Y}}_i(0)] \\
&\triangleq \frac{1}{m}\sum_{i\in\mathcal{I}_k}\boldsymbol{a}_{i1}^\top\boldsymbol{\tau}_i + \frac{1}{m}\sum_{i\in\mathcal{I}_k}\boldsymbol{a}_{i2}^\top\mathbb{E}^{(k)}[\tilde{\mathbf{Y}}_i(0)] \qquad \text{(B.1.50)}
\end{aligned}$$

As in the proof of Theorem B.1.3. Let

$$\Theta_i^* = \frac{\mathbf{\Pi}(\boldsymbol{W}_i)}{\boldsymbol{\pi}_i(\boldsymbol{W}_i)}, \quad \tilde{\mathbf{Y}}_i^{*\mathrm{obs}} = \mathbf{Y}_i^{\mathrm{obs}} - \boldsymbol{m}_i,$$

and $(\Gamma_\theta^*, \mathbf{\Gamma}_w^*)$ be the counterpart of $(\Gamma_\theta, \mathbf{\Gamma}_w)$ with $(\Theta_i, \tilde{\mathbf{Y}}_i^{\mathrm{obs}})$ replaced by $(\Theta_i^*, \tilde{\mathbf{Y}}_i^{*\mathrm{obs}})$. Recalling (B.1.9) on page 119, there exists a constant C_1 that only depends on c_π and T such that

$$\begin{aligned}
&\left|\mathbb{E}^{(k)}[(\Theta_i - \Theta_i^*)\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\right| + \left\|\mathbb{E}^{(k)}[(\Theta_i - \Theta_i^*)J\boldsymbol{W}_i]\right\|_2 + \left|\mathbb{E}^{(k)}[\Theta_i - \Theta_i^*]\right| \\
&+ \left\|\mathbb{E}^{(k)}[(\Theta_i - \Theta_i^*)J\,\mathrm{diag}(\boldsymbol{W}_i)]\right\|_{\mathrm{op}} \le C_1\Delta_{\pi i}. \qquad \text{(B.1.51)}
\end{aligned}$$

where $\Delta_{\pi i}'$ and $\bar{\Delta}_\pi'$ are defined in (B.1.30) and (B.1.31), respectively. Then,

$$\begin{aligned}
&\left|\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta'] - \left(\mathbb{E}[\Theta_i^*\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*]\right)\right| \\
&\le \left|\mathbb{E}^{(k)}[\Theta_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)] - \mathbb{E}[\Theta_i^*\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\right|\cdot\mathbb{E}[\Gamma_\theta'] \\
&\quad + \mathbb{E}[\Theta_i^*\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\cdot|(\mathbb{E}[\Gamma_\theta'] - \mathbb{E}[\Gamma_\theta^*])| \\
&\le C_1(\mathbb{E}[\Gamma_\theta']\cdot\Delta_{\pi i} + \mathbb{E}[\Theta_i^*\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\cdot\bar{\Delta}_\pi').
\end{aligned}$$

Note that $\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}^\top J \operatorname{diag}(\boldsymbol{W})] \le T$ is a constant. By Assumption B.1.2, $\mathbb{E}[\Gamma_\theta] \le 1/c_\pi$. Thus,

$$\begin{aligned}&\left\|\boldsymbol{a}_{i1} - \left(\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* J \operatorname{diag}(\boldsymbol{W}_i)]\right)\right\|_2 \\ &\le C_2(\Delta_{\pi i} + \bar{\Delta}'_\pi),\end{aligned} \tag{B.1.52}$$

for some constant C_2 that only depends on c_π and T. Let

$$\boldsymbol{a}_{i1}^* = \mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)]\mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Theta_i^* J \operatorname{diag}(\boldsymbol{W}_i)].$$

Since $\|\boldsymbol{\tau}_i\|_2 \le \Delta_{yi}$,

$$\left|\frac{1}{m}\sum_{i\in\mathcal{I}_k} \boldsymbol{a}_{i1}^\top \boldsymbol{\tau}_i - \frac{1}{m}\sum_{i\in\mathcal{I}_k} \boldsymbol{a}_{i1}^{*\top} \boldsymbol{\tau}_i\right| \le \frac{C_3}{m}\sum_{i\in\mathcal{I}_k} (\Delta_{\pi i} + \bar{\Delta}'_\pi)\Delta_{yi}. \tag{B.1.53}$$

On the other hand, by definition of Θ_i^*,

$$\mathbb{E}[\Theta_i^* \boldsymbol{W}_i^\top J \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}^\top J \operatorname{diag}(\boldsymbol{W})], \quad \mathbb{E}[\Theta_i^* J \operatorname{diag}(\boldsymbol{W}_i)] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J \operatorname{diag}(\boldsymbol{W})],$$

and

$$\mathbb{E}[\Gamma_\theta^*] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i^*] = 1, \quad \mathbb{E}[\boldsymbol{\Gamma}_w^*] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Theta_i^* J \boldsymbol{W}_i] = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}].$$

Thus,

$$\begin{aligned}\boldsymbol{a}_{i1}^{*\top} &= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}^\top J \operatorname{diag}(\boldsymbol{W})] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J\boldsymbol{W}]^\top \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[J \operatorname{diag}(\boldsymbol{W}_i)] \\ &= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \operatorname{diag}(\boldsymbol{W})].\end{aligned}$$

By the DATE equation,

$$\boldsymbol{a}_{i1}^{*\top} = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}]\xi^\top.$$

As a consequence,

$$\frac{1}{m}\sum_{i\in\mathcal{I}_k} \boldsymbol{a}_{i1}^{*\top} \boldsymbol{\tau}_i = \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}] \cdot \left(\frac{1}{m}\sum_{i\in\mathcal{I}_k} \xi^\top \boldsymbol{\tau}_i\right). \tag{B.1.54}$$

Now we turn to the second and third terms of (B.1.50). Similar to (B.1.52), we can show that

$$\|\boldsymbol{a}_{i2}\|_2 = \|\boldsymbol{a}_{i2} - (\mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Gamma_\theta^*] - \mathbb{E}[\boldsymbol{\Gamma}_w^*]^\top \mathbb{E}[\Gamma_\theta^*])\|_2 \le C_3(\Delta_{\pi i} + \bar{\Delta}'_\pi),$$

for some constant C_3 that only depends on c_π and T. By (B.1.1), $\|\tilde{\boldsymbol{Y}}_i(0)\|_2 \le \Delta_{yi}$. Therefore,

$$\left| \frac{1}{m} \sum_{i \in \mathcal{I}_k} \boldsymbol{a}_{i2}^\top \mathbb{E}^{(k)}[\tilde{\boldsymbol{Y}}_i(0)] \right| \le \frac{C_3}{m} \sum_{i \in \mathcal{I}_k} (\Delta_{\pi i} + \bar{\Delta}'_\pi)\Delta_{yi}. \tag{B.1.55}$$

Putting (B.1.50), (B.1.53), (B.1.54), and (B.1.55) together, we arrive at

$$\begin{aligned}
&\left| \mathbb{E}^{(k)}[\Gamma_{wy}^{(k)}]\mathbb{E}[\Gamma'_\theta] - \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top \mathbb{E}[\boldsymbol{\Gamma}_y^{(k)}] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])^\top J \boldsymbol{W}] \cdot \left(\frac{1}{m} \sum_{i \in \mathcal{I}_k} \xi^\top \boldsymbol{\tau}_i \right) \right| \\
&\le \frac{C_4}{m} \sum_{i \in \mathcal{I}_k} (\Delta_{\pi i} + \bar{\Delta}'_\pi)\Delta_{yi},
\end{aligned}$$

for some constant C_4 that only depends on c_π and T. Since $\tau^* = 0$,

$$\frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{m} \sum_{i \in \mathcal{I}_k} \xi^\top \boldsymbol{\tau}_i \right) = \frac{1}{n} \sum_{i=1}^{n} \xi^\top \boldsymbol{\tau}_i = \tau^* = 0$$

Therefore, averaging over k and marginalizing over $\mathcal{D}_{-k}$ yields that

$$\frac{1}{K} \sum_{k=1}^{K} \left| \mathbb{E}^{(k)}[\Gamma_{wy}]\mathbb{E}[\Gamma'_\theta] - \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top \mathbb{E}^{(k)}[\boldsymbol{\Gamma}_y^{(k)}] \right| = O\left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(\Delta_{\pi i} + \bar{\Delta}'_\pi)\Delta_{yi} \right] \right).$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}
&\frac{1}{K} \sum_{k=1}^{K} \left| \mathbb{E}^{(k)}[\Gamma_{wy}]\mathbb{E}[\Gamma'_\theta] - \mathbb{E}[\boldsymbol{\Gamma}'_w]^\top \mathbb{E}^{(k)}[\boldsymbol{\Gamma}_y^{(k)}] \right| \\
&= O\left(\frac{1}{n} \sum_{i=1}^{n} \sqrt{\mathbb{E}\left[(\Delta_{\pi i} + \bar{\Delta}'_\pi)^2 \right]} \sqrt{\mathbb{E}\left[\Delta_{yi}^2 \right]} \right) \\
&= O\left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(\Delta_{\pi i} + \bar{\Delta}'_\pi)^2 \right]} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\Delta_{yi}^2 \right]} \right) \\
&= O\left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}[\Delta_{\pi i}^2] + \mathbb{E}[\bar{\Delta}_\pi'^2] \right)} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\Delta_{yi}^2]} \right) \\
&= O\left(\sqrt{\mathbb{E}[\bar{\Delta}_\pi^2] + \mathbb{E}[\bar{\Delta}_\pi'^2]} \cdot \sqrt{\mathbb{E}[\bar{\Delta}_y^2]} \right).
\end{aligned}$$

Therefore, (B.1.49) is proved by (B.1.32) and (B.1.33) on page 133.

Step 6 Next, we shall prove that

$$\frac{1}{K}\sum_{k=1}^{K}\left|\mathbb{E}[\Gamma'_{wy}]\mathbb{E}^{(k)}[\Gamma_\theta^{(k)}-\Gamma_\theta^{'(k)}]-\mathbb{E}^{(k)}[\mathbf{\Gamma}_w^{(k)}-\mathbf{\Gamma}_w^{'(k)}]^\top\mathbb{E}[\mathbf{\Gamma}'_y]\right|=O\left(\sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}_y^2]}\right). \tag{B.1.56}$$

Using the same argument as (B.1.51), we can show that

$$\left\|\mathbb{E}[(\Theta'_i-\Theta_i)J\boldsymbol{W}_i]\right\|_2+\left|\mathbb{E}[\Theta'_i-\Theta_i]\right|\le C_1(\Delta'_{\pi i}+\Delta_{\pi i}).$$

Averaging over $i\in\mathcal{I}_k$, we obtain that

$$|\mathbb{E}^{(k)}[\Gamma_\theta^{(k)}]-\mathbb{E}^{(k)}[\Gamma_\theta^{'(k)}]|+\|\mathbb{E}^{(k)}[\mathbf{\Gamma}_w^{(k)}]-\mathbb{E}^{(k)}[\mathbf{\Gamma}_w^{'(k)}]\|_2\le C_1(\bar{\Delta}'_\pi+\bar{\Delta}_\pi)=O\left(\bar{\Delta}_\pi\right), \tag{B.1.57}$$

where the last step uses Remark B.1.1. On the other hand,

$$\begin{aligned}\mathbb{E}[\Gamma'_{wy}]&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta'_i\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\mathrm{obs}}]\\&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta'_i\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i'(0)]+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta'_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)\boldsymbol{\tau}_i]\\&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta'_i\boldsymbol{W}_i^\top J]\mathbb{E}[\tilde{\boldsymbol{Y}}_i'(0)]+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Theta'_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\boldsymbol{\tau}_i.\end{aligned}$$

Note that

$$\|\mathbb{E}[\Theta'_i\boldsymbol{W}_i^\top J]\|_2\le\frac{\sqrt{T}}{c_\pi},\quad\|\mathbb{E}[\Theta'_i\boldsymbol{W}_i^\top J\,\mathrm{diag}(\boldsymbol{W}_i)]\|_2\le\frac{T}{c_\pi},\quad\|\mathbb{E}[\tilde{\boldsymbol{Y}}_i'(0)]\|_2+\|\boldsymbol{\tau}_i\|_2\le\Delta_{yi}.$$

As a result, there exists a constant C_5 that only depends on c_π and T such that

$$|\mathbb{E}[\Gamma'_{wy}]|\le\frac{C_5}{n}\sum_{i=1}^{n}\Delta_{yi}\le C_5\bar{\Delta}_y.$$

Similarly,

$$\|\mathbb{E}[\mathbf{\Gamma}'_y]\|_2\le C_5\bar{\Delta}_y.$$

Together with (B.1.57), we prove (B.1.56).

Step 7 Consider the following decompositions:

$$\Gamma_{wy}\Gamma_\theta-\Gamma'_{wy}\Gamma'_\theta$$

$$
\begin{aligned}
&= (\Gamma_{wy} - \Gamma'_{wy})(\Gamma_\theta - \Gamma'_\theta) \\
&+ (\Gamma_{wy} - \Gamma'_{wy})(\Gamma'_\theta - \mathbb{E}[\Gamma'_\theta]) + (\Gamma'_{wy} - \mathbb{E}[\Gamma'_{wy}])(\Gamma_\theta - \Gamma'_\theta) \\
&+ \left(\Gamma_{wy} - \Gamma'_{wy} - \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}^{(k)}[\Gamma_{wy}^{(k)} - \Gamma_{wy}^{'(k)}]\right)\mathbb{E}[\Gamma'_\theta] + \mathbb{E}[\Gamma'_{wy}]\left(\Gamma_\theta - \Gamma'_\theta - \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}^{(k)}[\Gamma_\theta^{(k)} - \Gamma_\theta^{'(k)}]\right) \\
&- \frac{1}{K}\left(\sum_{k=1}^{K} \mathbb{E}^{(k)}[\Gamma_{wy}^{'(k)}]\right) \cdot \mathbb{E}[\Gamma'_\theta] \\
&+ \frac{1}{K}\left(\sum_{k=1}^{K} \mathbb{E}^{(k)}[\Gamma_{wy}^{(k)}]\right) \cdot \mathbb{E}[\Gamma'_\theta] \\
&+ \mathbb{E}[\Gamma'_{wy}] \cdot \frac{1}{K}\left(\sum_{k=1}^{K} \mathbb{E}^{(k)}[\Gamma_\theta^{(k)} - \Gamma_\theta^{'(k)}]\right),
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbf{\Gamma}_w^\top \mathbf{\Gamma}_y - \mathbf{\Gamma}_w^{'\top} \mathbf{\Gamma}'_y \\
&= (\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w)^\top (\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y) \\
&+ (\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w)^\top (\mathbf{\Gamma}'_y - \mathbb{E}[\mathbf{\Gamma}'_y]) + (\mathbf{\Gamma}'_w - \mathbb{E}[\mathbf{\Gamma}'_w])^\top (\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y) \\
&+ \left(\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w - \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}^{(k)}[\mathbf{\Gamma}_w^{(k)} - \mathbf{\Gamma}_w^{'(k)}]\right)^\top \mathbb{E}[\mathbf{\Gamma}'_y] + \mathbb{E}[\mathbf{\Gamma}'_w]^\top \left(\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y - \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}^{(k)}[\mathbf{\Gamma}_y^{(k)} - \mathbf{\Gamma}_y^{'(k)}]\right) \\
&- \mathbb{E}[\mathbf{\Gamma}'_w]^\top \frac{1}{K}\left(\sum_{k=1}^{K} \mathbb{E}^{(k)}[\mathbf{\Gamma}_y^{'(k)}]\right) \\
&+ \mathbb{E}[\mathbf{\Gamma}'_w]^\top \frac{1}{K}\left(\sum_{k=1}^{K} \mathbb{E}^{(k)}[\mathbf{\Gamma}_y^{(k)}]\right) \\
&+ \frac{1}{K}\left(\sum_{k=1}^{K} \mathbb{E}^{(k)}[\mathbf{\Gamma}_w^{(k)} - \mathbf{\Gamma}_w^{'(k)}]\right)^\top \mathbb{E}[\mathbf{\Gamma}'_y].
\end{aligned}
$$

Since $(\boldsymbol{\pi}'_i, \boldsymbol{m}'_i)$'s are deterministic,

$$
\frac{1}{K}\left(\sum_{k=1}^{K} \mathbb{E}^{(k)}[\Gamma_{wy}^{'(k)}]\right) \cdot \mathbb{E}[\Gamma'_\theta] = \frac{1}{K}\left(\sum_{k=1}^{K} \mathbb{E}[\Gamma_{wy}^{'(k)}]\right) \cdot \mathbb{E}[\Gamma'_\theta] = \mathbb{E}[\Gamma'_{wy}]\mathbb{E}[\Gamma'_\theta],
$$

and

$$\mathbb{E}[\mathbf{\Gamma}'_w]^\top \frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}^{(k)}[\mathbf{\Gamma}_y^{'(k)}]\right) = \mathbb{E}[\mathbf{\Gamma}'_w]^\top \frac{1}{K}\left(\sum_{k=1}^{K}\mathbb{E}[\mathbf{\Gamma}_y^{'(k)}]\right) = \mathbb{E}[\mathbf{\Gamma}'_w]^\top \mathbb{E}[\mathbf{\Gamma}'_y].$$

By (B.1.41), (B.1.42), (B.1.47), (B.1.48), (B.1.49), (B.1.56), and triangle inequality,

$$\Gamma_{wy}\Gamma_\theta - \mathbf{\Gamma}_w^\top \mathbf{\Gamma}_y - \left\{\Gamma'_{wy}\Gamma'_\theta - \mathbf{\Gamma}_w^{'\top}\mathbf{\Gamma}'_y\right\} = O_{\mathbb{P}}\left(n^{-r} + n^{-(r+1)/2} + n^{-(r'+1)/2} + \sqrt{\mathbb{E}[\bar{\Delta}_\pi^2]}\cdot\sqrt{\mathbb{E}[\bar{\Delta}_y^2]}\right).$$

The proof of (B.1.29) is then completed.

Proof of (B.1.38)**:** let

$$\hat{\mathcal{V}}_i'' = \Theta'_i\left\{\Gamma_{wy} - \mathbf{\Gamma}_y^\top J\boldsymbol{W}_i + \Gamma_\theta \boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \mathbf{\Gamma}_w^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\right\}. \tag{B.1.58}$$

Recalling the definition of $\hat{\mathcal{V}}'_i$ in (B.1.37) on page 134,

$$\begin{aligned}
|\hat{\mathcal{V}}'_i - \hat{\mathcal{V}}''_i| &\le |\Gamma_{wy} - \Gamma'_{wy}|\cdot\Theta'_i + \|\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y\|_2 \cdot \|\Theta'_i J\boldsymbol{W}_i\|_2 \\
&\quad + |\Gamma_\theta - \Gamma'_\theta|\cdot|\Theta'_i\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}| + \|\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w\|_2 \cdot \|\Theta'_i J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2 \\
&\le \left\{|\Gamma_{wy} - \Gamma'_{wy}| + \|\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y\|_2 + |\Gamma_\theta - \Gamma'_\theta| + \|\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w\|_2\right\} \\
&\quad \cdot\left\{\Theta'_i + \|\Theta'_i J\boldsymbol{W}_i\|_2 + |\Theta'_i\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}| + \|\Theta'_i J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2\right\}
\end{aligned}$$

By Jensen's inequality and Cauchy-Schwarz inequality,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}'_i - \hat{\mathcal{V}}''_i)^2 \\
&\le 4\left\{|\Gamma_{wy} - \Gamma'_{wy}| + \|\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y\|_2 + |\Gamma_\theta - \Gamma'_\theta| + \|\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w\|_2\right\}^2 \\
&\quad \cdot\frac{1}{n}\sum_{i=1}^{n}\left\{\Theta_i^{'2} + \|\Theta'_i J\boldsymbol{W}_i\|_2^2 + |\Theta'_i\boldsymbol{W}_i^\top J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}|^2 + \|\Theta'_i J\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2^2\right\} \\
&\le \frac{8T}{c_\pi^2}\left\{|\Gamma_{wy} - \Gamma'_{wy}|^2 + \|\mathbf{\Gamma}_y - \mathbf{\Gamma}'_y\|_2^2 + |\Gamma_\theta - \Gamma'_\theta|^2 + \|\mathbf{\Gamma}_w - \mathbf{\Gamma}'_w\|_2^2\right\}\cdot\frac{1}{n}\sum_{i=1}^{n}\left\{1 + \|\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2^2\right\},
\end{aligned}$$

where the last inequality uses Assumption B.1.2. By (B.1.39) and (B.1.40) on page 136,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}'_i - \hat{\mathcal{V}}''_i)^2 = O_{\mathbb{P}}(n^{-r}) = o_{\mathbb{P}}(1). \tag{B.1.59}$$

On the other hand, recalling the definition of $\hat{\mathcal{V}}_i$ in (B.1.11) on page 122,

$$\begin{aligned}
|\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i| &\le |\Gamma_{wy}| \cdot |\Theta_i' - \Theta_i| + \|\boldsymbol{\Gamma}_y\|_2 \cdot \|(\Theta_i' - \Theta_i) J \boldsymbol{W}_i\|_2 \\
&\quad + |\Gamma_\theta| \cdot |\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}| + \|\boldsymbol{\Gamma}_w\|_2 \cdot \|\Theta_i' J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 \\
&\quad + \Theta_i \hat{\tau} \left\{ |\Gamma_{ww}| + |\boldsymbol{\Gamma}_w^\top J \boldsymbol{W}_i| + |\Gamma_\theta \boldsymbol{W}_i^\top J \boldsymbol{W}_i| + |\boldsymbol{\Gamma}_w^\top J \boldsymbol{W}_i| \right\} \\
&\le \left\{ |\Gamma_{wy}| + \|\boldsymbol{\Gamma}_y\|_2 + |\Gamma_\theta| + \|\boldsymbol{\Gamma}_w\|_2 \right\} \cdot \left\{ |\Theta_i' - \Theta_i| + \|(\Theta_i' - \Theta_i) J \boldsymbol{W}_i\|_2 \right. \\
&\quad \left. + |\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}| + \|\Theta_i' J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 \right\} \\
&\quad + \Theta_i |\hat{\tau}| \left\{ |\Gamma_{ww}| + |\boldsymbol{\Gamma}_w^\top J \boldsymbol{W}_i| + |\Gamma_\theta \boldsymbol{W}_i^\top J \boldsymbol{W}_i| + |\boldsymbol{\Gamma}_w^\top J \boldsymbol{W}_i| \right\}.
\end{aligned}$$

Since $\|J \boldsymbol{W}_i\|_2 \le \sqrt{T}$,

$$\|(\Theta_i' - \Theta_i) J \boldsymbol{W}_i\|_2 \le \sqrt{T} |\Theta_i' - \Theta_i|.$$

By triangle inequality,

$$\begin{aligned}
&|\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}| \\
&\le |\Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{\text{obs}}| + |\Theta_i' \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i \boldsymbol{W}_i^\top J \tilde{\boldsymbol{Y}}_i^{'\text{obs}}| \\
&\le \frac{\sqrt{T}}{c_\pi} \|\tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2 + \sqrt{T} \|\tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 \cdot |\Theta_i' - \Theta_i| \\
&= \frac{\sqrt{T}}{c_\pi} \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2 + \sqrt{T} \|\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2 \cdot |\Theta_i' - \Theta_i|.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\|\Theta_i' J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 \\
&\le \|\Theta_i J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 + \|\Theta_i' J \tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \Theta_i J \tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2 \\
&\le \frac{1}{c_\pi} \|\tilde{\boldsymbol{Y}}_i^{'\text{obs}} - \tilde{\boldsymbol{Y}}_i^{\text{obs}}\|_2 + \|\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2 \cdot |\Theta_i' - \Theta_i| \\
&= \frac{1}{c_\pi} \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2 + \|\tilde{\boldsymbol{Y}}_i^{'\text{obs}}\|_2 \cdot |\Theta_i' - \Theta_i|
\end{aligned}$$

Putting pieces together, we have that

$$(\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i)^2$$

$$\leq C\left\{|\Gamma_{wy}| + \|\mathbf{\Gamma}_y\|_2 + |\Gamma_\theta| + \|\mathbf{\Gamma}_w\|_2\right\}^2 \left\{|\Theta_i' - \Theta_i|^2 \cdot (1 + \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2) + \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2\right\}$$
$$+ C|\hat{\tau}|\left\{|\Gamma_{ww}| + |\mathbf{\Gamma}_w^\top J \boldsymbol{W}_i| + |\Gamma_\theta \boldsymbol{W}_i^\top J \boldsymbol{W}_i| + |\mathbf{\Gamma}_w^\top J \boldsymbol{W}_i|\right\},$$

for some constant C that only depends on c_π and T. By Lemma B.1.2 and Markov's inequality,

$$|\Gamma_{wy}| + \|\mathbf{\Gamma}_y\|_2 + |\Gamma_\theta| + \|\mathbf{\Gamma}_w\|_2 = O_{\mathbb{P}}(1), \quad |\Gamma_{ww}| + |\mathbf{\Gamma}_w^\top J \boldsymbol{W}_i| + |\Gamma_\theta \boldsymbol{W}_i^\top J \boldsymbol{W}_i| + |\mathbf{\Gamma}_w^\top J \boldsymbol{W}_i| = O(1).$$

By the first part of the theorem,

$$|\hat{\tau}| = o_{\mathbb{P}}(1).$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^n (\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i)^2 = O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^n \left\{|\Theta_i' - \Theta_i|^2 \cdot (1 + \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2) + \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2\right\}\right) + o_{\mathbb{P}}(1). \tag{B.1.60}$$

By Assumption B.1.2 and B.1.6,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[|\Theta_i' - \Theta_i|^2] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{\mathbf{\Pi}(\boldsymbol{W}_i)^2}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)^2 \boldsymbol{\pi}_i'(\boldsymbol{W}_i)^2}|\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i)|^2\right]$$
$$\leq \frac{1}{c_\pi^2}\sum_{i=1}^n \mathbb{E}[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^2] = O(n^{-r}) = o(1).$$

By Assumption B.1.6,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2\right] = O(n^{-r}) = o(1).$$

By Markov's inequality, we obtain that

$$\frac{1}{n}\sum_{i=1}^n |\Theta_i' - \Theta_i|^2 + \frac{1}{n}\sum_{i=1}^n \|\hat{\boldsymbol{m}}_i - \boldsymbol{m}_i'\|_2^2 = o_{\mathbb{P}}(1). \tag{B.1.61}$$

By Hölder's inequality,

$$\frac{1}{n}\sum_{i=1}^n |\Theta_i' - \Theta_i|^2 \cdot \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2 \leq \left(\frac{1}{n}\sum_{i=1}^n |\Theta_i' - \Theta_i|^{2(1+2/\omega)}\right)^{\omega/(2+\omega)} \left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^{2+\omega}\right)^{2/(2+\omega)}.$$

By Markov's inequality and Assumption B.1.4,

$$\frac{1}{n}\sum_{i=1}^n \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^{2+\omega} = O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^{2+\omega}]\right) = O_{\mathbb{P}}(1).$$

By Assumption B.1.2,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[|\Theta_i' - \Theta_i|^{2(1+2/\omega)}\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\boldsymbol{\Pi}(\boldsymbol{W}_i)^{2(1+2/\omega)}}{\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i)^{2(1+2/\omega)}\boldsymbol{\pi}_i'(\boldsymbol{W}_i)^{2(1+2/\omega)}}|\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i)|^{2(1+2/\omega)}\right] \\
&\leq \frac{1}{c_\pi^{4(1+2/\omega)}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2(1+2/\omega)}\Big] \\
&\overset{(i)}{\leq} \frac{1}{c_\pi^{4(1+2/\omega)}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[(\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i))^{2}\Big] \\
&= O(n^{-r}) = o(1),
\end{aligned}$$

where (i) uses the fact that $|\hat{\boldsymbol{\pi}}_i(\boldsymbol{W}_i) - \boldsymbol{\pi}_i'(\boldsymbol{W}_i)| \leq 1$. Thus, by Markov's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}|\Theta_i' - \Theta_i|^2 \cdot \|\tilde{\boldsymbol{Y}}_i'^{\text{obs}}\|_2^2 = o_{\mathbb{P}}(1). \tag{B.1.62}$$

Putting (B.1.60), (B.1.61), and (B.1.62) together, we conclude that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i)^2 = o_{\mathbb{P}}(1). \tag{B.1.63}$$

By Jensen's inequality, (B.1.59), and (B.1.63),

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i)^2 \leq \frac{2}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i'')^2 + \frac{2}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i'' - \hat{\mathcal{V}}_i)^2 = o_{\mathbb{P}}(1). \tag{B.1.64}$$

By Lemma B.1.2, it is easy to see that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^2 = O_{\mathbb{P}}(1). \tag{B.1.65}$$

As a result,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i' - \frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right| \leq \frac{1}{n}\sum_{i=1}^{n}|\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i' - \hat{\mathcal{V}}_i)^2} = o_{\mathbb{P}}(1).$$

Together with (B.1.65), it implies that

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2\right| = o_{\mathbb{P}}(1).$$

On the other hand, by triangle inequality, Cauchy-Schwarz inequality, and (B.1.65),

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^{'2}-\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^{2}\right|\le\frac{2}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i(\hat{\mathcal{V}}_i'-\hat{\mathcal{V}}_i)+\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i'-\hat{\mathcal{V}}_i)^2$$
$$\le 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^{2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i'-\hat{\mathcal{V}}_i)^2}+\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathcal{V}}_i'-\hat{\mathcal{V}}_i)^2=o_{\mathbb{P}}(1).$$

Therefore,

$$|\hat{\sigma}^2-\hat{\sigma}^{'2}|\le\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^{2}-\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i^{'2}\right|+\left|\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i\right)^2-\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{V}}_i'\right)^2\right|=o_{\mathbb{P}}(1).$$

□

B.1.6. Miscellaneous.

Proposition B.1.1. *[Petrov [1975], p. 112, Theorem 5] Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $\mathbb{E}[X_j]=0$, for all j. Assume also $\mathbb{E}[X_j^2 g(X_j)]<\infty$ for some function g that is non-negative, even, and non-decreasing in the interval $x>0$, with $x/g(x)$ being non-decreasing for $x>0$. Write $B_n=\sum_j \mathrm{Var}[X_j]$. Then,*

$$d_K\left(\mathcal{L}\left(\frac{1}{\sqrt{B_n}}\sum_{j=1}^{n}X_j\right),N(0,1)\right)\le\frac{A}{B_n g(\sqrt{B_n})}\sum_{j=1}^{n}\mathbb{E}\left[X_j^2 g(X_j)\right],$$

where A is a universal constant, $\mathcal{L}(\cdot)$ denotes the probability law, d_K denotes the Kolmogorov-Smirnov distance (i.e., the ℓ_∞-norm of the difference of CDFs)

Proposition B.1.2 (Theorem 2 of von Bahr and Esseen [1965]). *Let $\{Z_i\}_{i=1,\ldots,n}$ be independent mean-zero random variables. Then for any $a\in[0,1)$,*

$$\mathbb{E}\left|\sum_{i=1}^{n}Z_i\right|^{1+a}\le 2\sum_{i=1}^{n}\mathbb{E}|Z_i|^{1+a}.$$

B.2. Solving the DATE equation

For notational convenience, denote by $h(\mathbf{\Pi})=(h_1(\mathbf{\Pi}),\ldots,h_T(\mathbf{\Pi}))$ the LHS of the DATE equation. We start by a simple but useful observation that, for any $\mathbf{\Pi}$,

$$\mathbf{1}_T^\top h(\mathbf{\Pi})=\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[(\mathbf{1}_T^\top\mathrm{diag}(\boldsymbol{W})-\mathbf{1}_T^\top\xi\boldsymbol{W}^\top)J(\boldsymbol{W}-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\right]$$
$$=\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[(\boldsymbol{W}^\top-\boldsymbol{W}^\top)J(\boldsymbol{W}-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\right]=0. \tag{B.2.1}$$

Thus, there is at least one redundant equation and for any matrix $V \in \mathbb{R}^{T\times(T-1)}$ with $V^\top \mathbf{1}_T = 0$,

$$h(\mathbf{\Pi}) = 0 \Longleftrightarrow V^\top h(\mathbf{\Pi}) = 0. \tag{B.2.2}$$

B.2.1. Proof of equation (3.4.1)**.** Set $V = (1,-1)^\top$ in (B.2.2). Then

$$V^\top h(\mathbf{\Pi}) = 0 \Longleftrightarrow h_1(\mathbf{\Pi}) - h_2(\mathbf{\Pi}) = 0.$$

As a result,

$$\begin{aligned}
0 &= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[((W_1,-W_2)-(\xi_1-\xi_2)(W_1,W_2))\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\begin{bmatrix}W_1-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_1]\\ W_2-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_2]\end{bmatrix}\right]\\
&= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[(W_1+W_2-(\xi_1-\xi_2)(W_1-W_2))(W_1-W_2-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}(W_1-W_2))]\\
&= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_1^2-W_2^2-(\xi_1-\xi_2)(W_1-W_2)^2]-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_1+W_2-(\xi_1-\xi_2)(W_1-W_2)]\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}(W_1-W_2)\\
&= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_1-W_2-(\xi_1-\xi_2)(W_1-W_2)^2]-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_1+W_2-(\xi_1-\xi_2)(W_1-W_2)]\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}(W_1-W_2)\\
&= (\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))-(\xi_1-\xi_2)(\mathbf{\Pi}(1,0)+\mathbf{\Pi}(0,1))\\
&\quad -\{\mathbf{\Pi}(1,0)+\mathbf{\Pi}(0,1)+2\mathbf{\Pi}(1,1)-(\xi_1-\xi_2)(\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))\}\{\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1)\}\\
&= (\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))-(\xi_1-\xi_2)(\mathbf{\Pi}(1,0)+\mathbf{\Pi}(0,1))\\
&\quad -\{1+\mathbf{\Pi}(1,1)-\mathbf{\Pi}(0,0)-(\xi_1-\xi_2)(\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))\}\{\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1)\}.
\end{aligned}$$

Rearranging the terms yields

$$\{\mathbf{\Pi}(1,1)-\mathbf{\Pi}(0,0)\}\{\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1)\} = (\xi_1-\xi_2)\Big\{(\mathbf{\Pi}(1,0)-\mathbf{\Pi}(0,1))^2-(\mathbf{\Pi}(1,0)+\mathbf{\Pi}(0,1))\Big\}. \tag{B.2.3}$$

B.2.2. Proof of Theorem 3.4.1. Let $\boldsymbol{e}_j$ denote the j-th canonical basis in $\mathbb{R}^T$. Then

$$h_j(\mathbf{\Pi}) = \boldsymbol{e}_j^\top \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\Big[(\mathrm{diag}(\boldsymbol{W})-\xi\boldsymbol{W}^\top)J(\boldsymbol{W}-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\Big].$$

We can decompose $h_j(\mathbf{\Pi})$ into $h_{j1}(\mathbf{\Pi})-\xi_j h_2(\mathbf{\Pi})$ where

$$h_{j1}(\mathbf{\Pi}) = \boldsymbol{e}_j^\top \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\mathrm{diag}(\boldsymbol{W})J(\boldsymbol{W}-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])], \quad h_2(\mathbf{\Pi}) = \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\Big[\boldsymbol{W}^\top J(\boldsymbol{W}-\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\Big].$$

Then

$$
\begin{aligned}
h_{j1}(\mathbf{\Pi}) &= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\Big[W_j \boldsymbol{e}_j^\top J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}])\Big] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\Big[W_j \boldsymbol{e}_j^\top J \boldsymbol{W}\Big] - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\Big[W_j \boldsymbol{e}_j^\top J\Big]\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[W_j\left(W_j - \frac{\mathbf{1}_T^\top \boldsymbol{W}}{T}\right)\right] - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\Big[W_j\Big]\boldsymbol{e}_j^\top J \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\boldsymbol{W}] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[W_j\left(W_j - \frac{\mathbf{1}_T^\top \boldsymbol{W}}{T}\right)\right] - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\Big[W_j\Big]\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}\left[W_j - \frac{\mathbf{1}_T^\top \boldsymbol{W}}{T}\right] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_j] - (\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_j])^2 + \frac{\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_j]\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[\mathbf{1}_T^\top \boldsymbol{W}]}{T} - \frac{\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_j(\mathbf{1}_T^\top \boldsymbol{W})]}{T},
\end{aligned}
$$

where the last equality follows from the fact that $W_j^2 = W_j$. By (B.2.2), it is equivalent to find $\mathbf{\Pi}$ satisfying

$$
\Delta h_j(\mathbf{\Pi}) = h_{j+1}(\mathbf{\Pi}) - h_j(\mathbf{\Pi}) = 0, \quad j = 1, 2, \ldots, T-1.
$$

In this case, $\xi_{j+1} = \xi_j$ for any j, and thus,

$$
h_{(j+1)1}(\mathbf{\Pi}) - h_{j1}(\mathbf{\Pi}) = 0, \quad j = 1, 2, \ldots, T-1. \tag{B.2.4}
$$

By definition,

$$
W_{j+1} - W_j = I(\boldsymbol{W} = \boldsymbol{w}_{(T-j)}). \tag{B.2.5}
$$

As a consequence, we have

$$
\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_{j+1}] - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_j] = \mathbf{\Pi}(\boldsymbol{w}_{(T-j)}),
$$

$$
(\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_{j+1}])^2 - (\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_j])^2 = \mathbf{\Pi}(\boldsymbol{w}_{(T-j)})^2 + 2\mathbf{\Pi}(\boldsymbol{w}_{(T-j)})\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_j],
$$

and

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_{j+1}(\mathbf{1}_T^\top \boldsymbol{W})] - \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[W_j(\mathbf{1}_T^\top \boldsymbol{W})] \\
&= \mathbb{E}_{\boldsymbol{W}\sim\mathbf{\Pi}}[I(\boldsymbol{W} = \boldsymbol{w}_{(T-j)})(\mathbf{1}_T^\top \boldsymbol{w}_{(T-j)})] = (T-j)\mathbf{\Pi}(\boldsymbol{w}_{(T-j)}).
\end{aligned}
$$

As a result,

$$
h_{(j+1)1}(\mathbf{\Pi}) - h_{j1}(\mathbf{\Pi})
$$

$$= \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)})\left\{1 - \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)}) - 2\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j] + \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\mathbf{1}_T^\top \boldsymbol{W}]}{T} - \frac{T-j}{T}\right\}$$

$$= \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)})\left\{\frac{j}{T} - \boldsymbol{\Pi}(\boldsymbol{w}_{(T-j)}) - 2\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_j] + \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\mathbf{1}_T^\top \boldsymbol{W}]}{T}\right\}. \tag{B.2.6}$$

Let

$$g_j(\boldsymbol{\Pi}) = \frac{T-j}{T} - \boldsymbol{\Pi}(\boldsymbol{w}_{(j)}) - 2\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_{T-j}] + \frac{\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\mathbf{1}_T^\top \boldsymbol{W}]}{T}. \tag{B.2.7}$$

Thus, (B.2.4) can be reformulated as

$$\boldsymbol{\Pi}(\boldsymbol{w}_{(j)}) = 0 \text{ or } g_j(\boldsymbol{\Pi}) = 0, \quad j = 1, 2, \ldots, T-1. \tag{B.2.8}$$

Since $\mathbb{S}^* = \{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(j_1)}, \ldots, \boldsymbol{w}_{(j_r)}, \boldsymbol{w}_{(T)}\}$, $\boldsymbol{\Pi}(\boldsymbol{w}_{(j_k)}) > 0$ for each $k = 1, \ldots, r$. As a result, (B.2.8) is equivalent to

$$g_{j_r}(\boldsymbol{\Pi}) = 0, \quad g_{j_k}(\boldsymbol{\Pi}) - g_{j_{k+1}}(\boldsymbol{\Pi}) = 0, \quad k = 1, \ldots, r-1. \tag{B.2.9}$$

Note that

$$W_{T-j_k} = 1 \Longleftrightarrow \boldsymbol{W} \in \{\boldsymbol{w}_{(j_k+1)}, \ldots, \boldsymbol{w}_{(T)}\}.$$

The first equation is equivalent to

$$\frac{T-j_r}{T} - \boldsymbol{\Pi}(\boldsymbol{w}_{(j_r)}) - 2\boldsymbol{\Pi}(\boldsymbol{w}_{(T)}) + \frac{1}{T}\left(\sum_{k=1}^{r} j_k \boldsymbol{\Pi}(\boldsymbol{w}_{(j_k)}) + T\boldsymbol{\Pi}(\boldsymbol{w}_{(T)})\right) = 0$$

$$\Longleftrightarrow \boldsymbol{\Pi}(\boldsymbol{w}_{(T)}) = \frac{T-j_r}{T} - \boldsymbol{\Pi}(\boldsymbol{w}_{(j_r)}) + \frac{1}{T}\sum_{k=1}^{r} j_k \boldsymbol{\Pi}(\boldsymbol{w}_{(j_k)}). \tag{B.2.10}$$

By (B.2.5),

$$\begin{aligned}\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_{T-j_k}] - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[W_{T-j_{k+1}}] &= \mathbb{P}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left(\boldsymbol{W} \in \{\boldsymbol{w}_{(j_k+1)}, \boldsymbol{w}_{(j_k+2)}, \ldots, \boldsymbol{w}_{(j_{k+1})}\}\right) \\ &= \mathbb{P}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left(\boldsymbol{W} = \boldsymbol{w}_{(j_{k+1})}\right) = \boldsymbol{\Pi}(\boldsymbol{w}_{(j_{k+1})}).\end{aligned}$$

Therefore, the second equation of (B.2.8) can be simplified to

$$\boldsymbol{\Pi}(\boldsymbol{w}_{(j_{k+1})}) + \boldsymbol{\Pi}(\boldsymbol{w}_{(j_k)}) = \frac{j_{k+1} - j_k}{T}, \quad k = 1, \ldots, r-1. \tag{B.2.11}$$

Finally the simplex constraint determines $\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(0)})$ as

$$\mathbf{\Pi}(\boldsymbol{w}_{(0)}) = 1 - \mathbf{\Pi}(\boldsymbol{w}_{(T)}) - \sum_{k=1}^{r} \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}). \tag{B.2.12}$$

Clearly, $\mathbf{\Pi}(\boldsymbol{w}_{(j_1)})$ determines all other $\mathbf{\Pi}(\boldsymbol{W}_{(j_k)})$'s. Therefore, the solution set of (B.2.10) - (B.2.12) is a one-dimensional linear subspace. The solution set of the DATE equation is empty if it has no intersection with the set $\{\mathbf{\Pi} : \mathbf{\Pi}(\boldsymbol{w}_{(j_k)}) > 0, r = 1, \ldots, r\}$; otherwise, it must be a segment which can be characterized as $\{\lambda\mathbf{\Pi}^{(1)} + (1-\lambda)\mathbf{\Pi}^{(2)} : \lambda \in (0,1)\}$.

B.2.3. Proof of Theorem 3.4.2. Let $\boldsymbol{\eta} = (\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(1)}), \ldots, \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(T)})) \in \mathbb{R}^T$. Then the DATE equation can be equivalently formulated as

$$\sum_{j=1}^{T} \Big(\operatorname{diag}(\tilde{\boldsymbol{w}}_{(j)}) - \xi\tilde{\boldsymbol{w}}_{(j)}^\top\Big) J(\tilde{\boldsymbol{w}}_{(j)} - \boldsymbol{\eta})\eta_j = 0.$$

Since $\tilde{\boldsymbol{w}}_{(j)} = \boldsymbol{e}_j$, $\operatorname{diag}(\tilde{\boldsymbol{w}}_{(j)}) = \boldsymbol{e}_j\boldsymbol{e}_j^\top$ and we can reformulate the above equation as

$$\sum_{j=1}^{T} \Big(\boldsymbol{e}_j - \xi\Big)\boldsymbol{e}_j^\top J(\boldsymbol{e}_j - \boldsymbol{\eta})\eta_j = 0 \Longleftrightarrow \sum_{j=1}^{T} f_j(\boldsymbol{\eta})\boldsymbol{e}_j = \left\{\sum_{j=1}^{T} f_j(\boldsymbol{\eta})\right\}\xi.$$

where $f_j(\boldsymbol{\eta}) = \boldsymbol{e}_j^\top J(\boldsymbol{e}_j - \boldsymbol{\eta})\eta_j$. It can be equivalently formulated as an equation on $\boldsymbol{\eta}$ and a scalar b:

$$\sum_{j=1}^{T} f_j(\boldsymbol{\eta})\boldsymbol{e}_j = b\xi. \tag{B.2.13}$$

This is because for any $\boldsymbol{\eta}$ that satisfies (B.2.13), multiplying $\mathbf{1}_T^\top$ on both sides implies that

$$b = b(\xi^\top \mathbf{1}_T) = \sum_{j=1}^{T} f_j(\boldsymbol{\eta}).$$

Taking the j-th entry of both sides, (B.2.13) yields that

$$f_j(\boldsymbol{\eta}) = \xi_j b. \tag{B.2.14}$$

By definition,

$$f_j(\boldsymbol{\eta}) = \eta_j\Big(\boldsymbol{e}_j^\top J\boldsymbol{e}_j - \boldsymbol{e}_j^\top J\boldsymbol{\eta}\Big) = \eta_j\left(1 - \frac{1}{T} - \eta_j + \frac{1}{T}\sum_{j=1}^{T}\eta_j\right).$$

Since $\mathbf{\Pi}$ should be supported on $\{\tilde{\boldsymbol{w}}_{(0)}, \tilde{\boldsymbol{w}}_{(1)}, \ldots, \tilde{\boldsymbol{w}}_{(T)}\}$,

$$\sum_{j=1}^{T} \eta_j = \sum_{j=1}^{T} \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(j)}) = 1 - \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(0)}).$$

Therefore, (B.2.14) is equivalent to

$$\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(j)}) \left(1 - \mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(j)}) - \frac{\mathbf{\Pi}(\tilde{\boldsymbol{w}}_{(0)})}{T}\right) = \xi_j b.$$

B.2.4. Proof of Theorem 3.4.3. Let $\|\boldsymbol{w}\|_1$ be the L_1 norm of $\boldsymbol{w}$, i.e., $\|\boldsymbol{w}\|_1 = \sum_{i=1}^{n} w_i$. For given $\mathbf{\Pi}$ such that

$$\mathbf{\Pi}(\cdot \mid \|\boldsymbol{w}\|_1 = k') \sim \mathrm{Unif}(\mathcal{W}^{\mathrm{tra}}_{T,k'} \setminus \mathcal{W}^{\mathrm{tra}}_{T,k'-1}), \quad k' = 1, \ldots, k,$$

By symmetry,

$$\mathbb{E}[\boldsymbol{W} \mid \|\boldsymbol{W}\|_1] = \frac{\|\boldsymbol{W}\|_1}{T} \mathbf{1}_T.$$

By the iterated law of expectation,

$$\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}}[\boldsymbol{W}] = \mathbb{E}_{\|\boldsymbol{W}\|_1} \mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}}[\boldsymbol{W} \mid \|\boldsymbol{W}\|_1] = \frac{\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}}[\|\boldsymbol{W}\|_1]}{T} \mathbf{1}_T.$$

Since $J\mathbf{1}_T = 0$, the DATE equation with $\xi = \mathbf{1}_T/T$ reduces to

$$\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}} \left[\left(\mathrm{diag}(\boldsymbol{W}) - \frac{\mathbf{1}_T}{T} \boldsymbol{W}^\top \right) J \boldsymbol{W} \right] = 0.$$

We will prove the following stronger claim:

$$\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}} \left[\left(\mathrm{diag}(\boldsymbol{W}) - \frac{\mathbf{1}_T}{T} \boldsymbol{W}^\top \right) J \boldsymbol{W} \mid \|\boldsymbol{W}\|_1 = k' \right] = 0, \quad \forall k' = 1, \ldots, k.$$

Conditional on $\|\boldsymbol{W}\|_1 = k'$,

$$J\boldsymbol{W} = \boldsymbol{W} - \frac{k'}{T} \mathbf{1}_T, \quad \mathrm{diag}(\boldsymbol{W}) \boldsymbol{W} = \boldsymbol{W}, \quad \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{W}^\top \mathbf{1}_T = k'$$

Thus,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}} \left[\left(\mathrm{diag}(\boldsymbol{W}) - \frac{\mathbf{1}_T}{T} \boldsymbol{W}^\top \right) J \boldsymbol{W} \mid \|\boldsymbol{W}\|_1 = k' \right] \\ &= \mathbb{E}_{\boldsymbol{W} \sim \mathbf{\Pi}} \left[\left(\mathrm{diag}(\boldsymbol{W}) - \frac{\mathbf{1}_T}{T} \boldsymbol{W}^\top \right) \left(\boldsymbol{W} - \frac{k'}{T} \mathbf{1}_T \right) \mid \|\boldsymbol{W}\|_1 = k' \right] \end{aligned}$$

$$= \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\left[\boldsymbol{W} - \frac{k'}{T}\boldsymbol{W} - \frac{k'\mathbf{1}_T}{T} + \frac{k'^2\mathbf{1}_T}{T^2} \mid \|\boldsymbol{W}\|_1 = k'\right]$$

$$= 0.$$

B.2.5. A general solver via nonlinear programming. When $\mathbb{S}_{\text{design}} = \{\check{\boldsymbol{w}}_{(1)}, \ldots, \check{\boldsymbol{w}}_{(K)}\}$, the DATE equation can be formulated as a quadratic system. The j-th equation of DATE equation is

$$\mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}\Big[(e_j\boldsymbol{W}_j - \boldsymbol{W}\xi_j)^T J(\boldsymbol{W} - \mathbb{E}_{\boldsymbol{W}\sim\boldsymbol{\Pi}}[\boldsymbol{W}])\Big] = 0, \tag{B.2.15}$$

Let $\boldsymbol{p} = (\boldsymbol{\Pi}(\check{\boldsymbol{w}}_{(1)}), \ldots, \boldsymbol{\Pi}(\check{\boldsymbol{w}}_{(K)})) \in \mathbb{R}^T$, $A = (\check{\boldsymbol{w}}_{(1)}, \ldots, \check{\boldsymbol{w}}_{(K)}) \in \mathbb{R}^{T\times K}$, $B^{(j)} = (B_1^{(j)}, \ldots, B_K^{(j)}) \in \mathbb{R}^{T\times K}$, and $\boldsymbol{b}^{(j)} = (b_1^{(j)}, \ldots, b_K^{(j)})^\top \in \mathbb{R}^K$, where

$$B_k^{(j)} = J(\boldsymbol{e}_j\check{\boldsymbol{w}}_{(k),j} - \check{\boldsymbol{w}}_{(k)}\xi_j) \in \mathbb{R}^T, \quad b_k^{(j)} = \check{\boldsymbol{w}}_{(k)}^\top B_k^{(j)} \in \mathbb{R}.$$

It is easy to see that $B^{(j)} = J(\boldsymbol{e}_j\boldsymbol{e}_j^\top - \xi_j I)A$ and $\boldsymbol{b}^{(j)} = \text{diag}(A^\top B^{(j)})$. Then (B.2.15) can be reformulated as

$$\boldsymbol{p}^\top\boldsymbol{b}^{(j)} - \boldsymbol{p}^\top(A^\top B^{(j)})\boldsymbol{p} = 0.$$

As a result, the DATE equation has a solution iff the minimal value of the following optimization problem is 0:

$$\min \sum_{j=1}^{T}\{\boldsymbol{p}^\top\boldsymbol{b}^{(j)} - \boldsymbol{p}^\top(A^\top B^{(j)})\boldsymbol{p}\}^2, \quad \text{s.t., } \boldsymbol{p}^\top\mathbf{1} = 1, \boldsymbol{p} \geq 0. \tag{B.2.16}$$

We can optimize (B.2.16) via the standard BFGS algorithm, with the uniform distribution being the initial value. When the minimal value with a given initial value is bounded away from zero, we will try other randomly generated initial values to ensure a thorough search. If none of the initial values yields a zero objective, we claim that the DATE equation has no solution. Note that (B.2.16) is a nonconvex problem, the BFGS algorithm is not guaranteed to find the global minimum. Therefore, it should be viewed as an attempt to find a solution of the DATE equation instead of a trustable solver.

On the other hand, when the DATE equation has multiple solutions, it is unclear which solution can be found. In principle, we can add different constraints or regularizers to (B.2.16) in order to obtain a "well-behaved" solution. For instance, it is reasonable to find the most dispersed reshaped function to maximize the sample efficiency. For this purpose, we can find the solution

that maximizes $\min_k \mathbf{\Pi}(\check{\boldsymbol{w}}_{(k)})$. This can be achieved by replacing the constraint $\boldsymbol{p} \geq 0$ in (B.2.16) by $\boldsymbol{p} \geq c\mathbf{1}$ and find the largest c for which the minimal value is zero.

B.3. Aggregated AIPW estimator is not doubly robust in the presence of fixed effects

We are not aware of other doubly robust estimators for DATE when the treatment and outcome models are defined as in our paper. In absence of dynamic treatment effects, it is tempting to treat each period as a cross-sectional data, estimate the time-specific ATE τ_t by an aggregated AIPW estimator, and aggregate these estimates. To the best of our knowledge, this estimator has not been proposed in the literature. However, perhaps surprisingly, we show in this section that the aggregated AIPW estimator is not doubly robust because of the fixed effect terms in the outcome model (3.3.6).

Specifically, for time period t, the AIPW estimator for τ_t is defined as

$$\hat{\tau}_t = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(Y_{it} - \hat{\mathbb{E}}[Y_{it}(1) \mid \boldsymbol{X}_i])W_{it}}{\hat{\mathbb{P}}(W_{it} = 1 \mid \boldsymbol{X}_i)} - \frac{(Y_{it} - \hat{\mathbb{E}}[Y_{it}(0) \mid \boldsymbol{X}_i])(1 - W_{it})}{\hat{\mathbb{P}}(W_{it} = 0 \mid \boldsymbol{X}_i)} + \hat{\mathbb{E}}[Y_{it}(1) \mid \boldsymbol{X}_i] - \hat{\mathbb{E}}[Y_{it}(0) \mid \boldsymbol{X}_i]\right).$$

Then the aggregated AIPW estimator is defined as

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\tau}_t.$$

It is known that $\hat{\tau}_t$ is doubly robust in the sense that $\hat{\tau}_t$ is consistent if either $\hat{\mathbb{P}}(W_{it} = 1)$ or $(\hat{\mathbb{E}}[Y_{it}(1) \mid X_i], \hat{\mathbb{E}}[Y_{it}(0) \mid X_i])$ is consistent for all i and t. Importantly, the requirement on the outcome model for the AIPW estimator is strictly stronger than that for the RIPW estimator; the former requires both m_{it} and the fixed effects to be consistently estimated while the latter only requires m_{it} to be consistent. It turns out that the extra requirement leads to tricky problems of the AIPW estimator.

To demonstrate the failure of the AIPW estimator, we only consider the case with a large sample size $n = 10000$ and a constant treatment effect to highlight that the failure is not driven by small samples or effect heterogeneity. In particular, we consider a standard TWFE model

$$Y_{it}(0) = \alpha_i + \gamma_t + m_{it} + \epsilon_{it}, \quad m_{it} = X_i\beta_t, \quad \tau_{it} = \tau,$$

where $\sum_{i=1}^{n}\alpha_i = \sum_{t=1}^{T}\gamma_t = 0$. The other details are the same as Section 3.5.1.

Both the RIPW and the aggregated AIPW estimators require estimates of the treatment and outcome models. First, we consider a wrong and a correct treatment model:

- (Wrong treatment model): set $\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) = |\{j : \boldsymbol{W}_j = \boldsymbol{w}\}|/n$, i.e., the empirical distribution of $\boldsymbol{W}_i$'s that ignores the covariate;
- (Correct treatment model): set $\hat{\boldsymbol{\pi}}_i(\boldsymbol{w}) = |\{j : \boldsymbol{W}_j = \boldsymbol{w}, X_j = X_i\}|/|\{j : X_j = X_i\}|$, i.e., the empirical distribution of $\boldsymbol{W}_i$'s stratified by the covariate.

With a large sample, $\hat{\boldsymbol{\pi}}_i$ in the second setting is a consistent estimator of $\boldsymbol{\pi}_i$. For the aggregated AIPW estimator, we use the marginal distributions of $\hat{\boldsymbol{\pi}}_i$ as the estimates of marginal propensity scores. Similarly, we consider a wrong and a correct outcome model:

- (Wrong outcome model): $\hat{m}_{it} = 0$ for every i and t;
- (Correct outcome model): run unweighted TWFE regression adjusting for interaction between X_i and time fixed effects, i.e., $X_i I(t = t')$ for each $t' = 1, \ldots, t$, and set $\hat{m}_{it} = X_i \hat{\beta}_t$.

With a large sample, the standard theory implies the consistency of $\hat{\beta}_t$, and hence $\hat{m}_{it} \approx m_{it}$. Unlike the RIPW estimator, the aggregated AIPW estimator requires the estimate of full conditional expectations of potential outcomes, instead of merely $\hat{m}_{it}$. In this case, a reasonble estimate of the outcome model can be formulated as

$$\hat{\mathbb{E}}[Y_{it}(0) \mid X_i] = \hat{\alpha}_i + \hat{\gamma}_t + X_i \hat{\beta}_t, \quad \hat{\mathbb{E}}[Y_{it}(1) \mid X_i] = \hat{\mathbb{E}}[Y_{it}(0) \mid X_i] + \hat{\tau}.$$

For short panels with $T = O(1)$, the time fixed effects γ_t's can be estimated via the standard TWFE regression, which are known to be consistent. However, there is no way to consistently estimate the unit fixed effect α_i since only T samples $Y_{i1}, \ldots, Y_{iT}$ can be used for estimation. The central question is how to estimate α_i for the aggregated AIPW estimator. Here we consider three strategies:

(1) using the plug-in estimate of α_i's, even if they are inconsistent;
(2) pretending that α_i does not exist and setting $\hat{\alpha}_i = 0$;
(3) using the Mundlak-type regression estimates proposed by Arkhangelsky and Imbens [2018].

Note that the first strategy cannot be used with cross-fitting because it is impossible to estimate α_i without using the i-th sample.

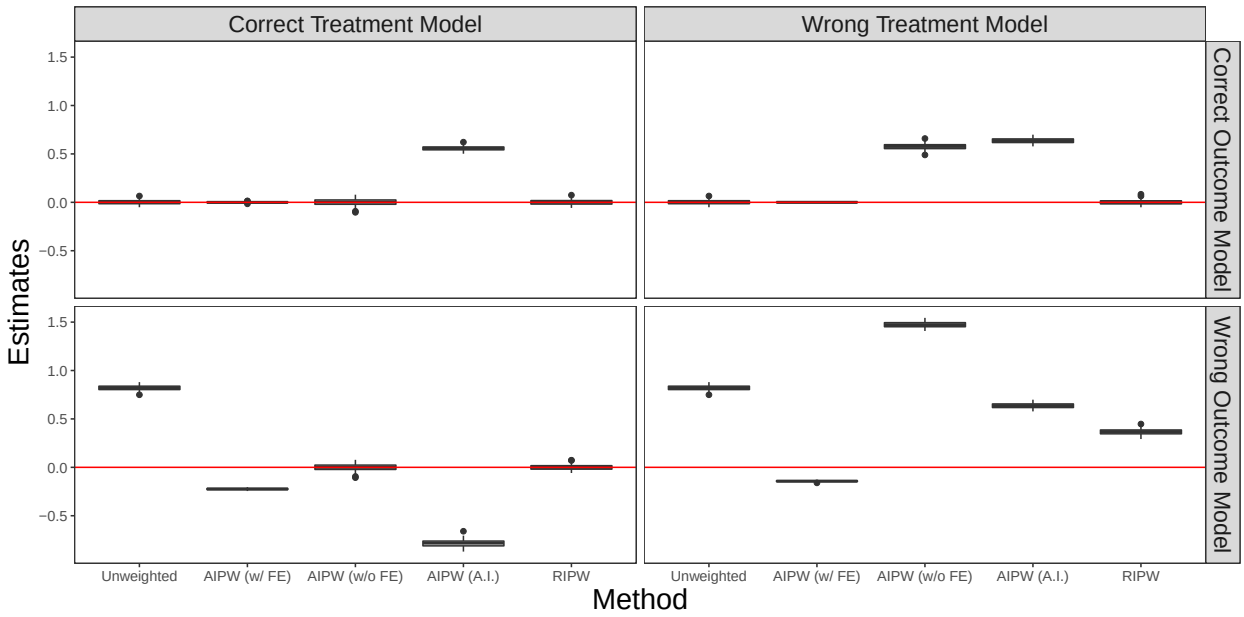


FIGURE B.1. Boxplots of $\hat{\tau} - \tau$ for the RIPW, unweighted TWFE, and three versions of AIPW estimators

"AIPW (w/ FE)" for the one with estimated fixed effects, "AIPW (w/o FE)" for the one that zeros out the fixed effects, and "AIPW (A.I.)" for the one that uses Arkhangelsky and Imbens [2018]'s estimator.

We then consider all four combinations of outcome and treatment modelling. Figure B.1 presents the boxplots of $\hat{\tau} - \tau$ for the three versions of AIPW, RIPW, and unweighted TWFE estimator.

First, we can see that all estimators are unbiased when both models are correct and biased when both models are wrong. As expected, the RIPW estimator is also unbiased when one of the model is correct, and the unweighted estimator is unbiased when the outcome model is correct. However, none of AIPW estimators are doubly robust: the AIPW estimator with estimated fixed effects is biased when the treatment model is correct, and the AIPW estimator that zeros out fixed effects or applies Mundlak-type estimator are biased when the outcome model is correct.

The bias of AIPW that zeros out the fixed effects can be attributed to biased estimates of the outcome model despite including the covariates. The other two AIPW estimators can be attributed to the dependence between the outcome model estimates on the treatment assignment. In fact, when T is small, this dependence is nonvanishing no matter how fixed effects are estimated. On the other hand, the AIPW estimator is valid under a correct treatment model but a wrong outcome model only when the outcome model estimate is asymptotically independent of the assignments.

In sum, there is no simple way to estimate fixed effects to make the resulting aggregated AIPW estimator doubly robust.

Bibliography

Rahi Abouk and Babak Heydari. The immediate effect of covid-19 policies on social-distancing behavior in the united states. *Public health reports*, 136(2):245–252, 2021.

Sarah Abraham and Liyang Sun. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *arXiv preprint arXiv:1804.05785*, 2018.

Richard H Adams Jr and John Page. Do international migration and remittances reduce poverty in developing countries? *World development*, 33(10):1645–1669, 2005.

Rita Afsar. Internal migration and the development nexus: the case of bangladesh. In *Regional conference on migration, development and pro-poor policy choices in Asia*, pages 22–24, 2003.

Mehtap Akgüç, Corrado Giulietti, and Klaus F Zimmermann. The rumic longitudinal survey: Fostering research on labor markets in china. *IZA Journal of Labor & Development*, 3(1):5, 2014.

Carlo Alcaraz, Daniel Chiquiar, and Alejandrina Salcedo. Remittances, schooling, and child labor in mexico. *Journal of Development Economics*, 97(1):156–165, 2012.

Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951.

Francisca M Antman. The intergenerational effects of paternal migration on schooling and work: What can we learn from children's time allocations? *Journal of Development Economics*, 96(2): 200–208, 2011.

Francisca M Antman. 16 the impact of migration on family left behind. *International handbook on the economics of migration*, page 293, 2013.

Manuel Arellano. *Panel data econometrics*. Oxford university press, 2003.

Marie Joy B Arguillas and Lindy Williams. The impact of parents' overseas employment on educational outcomes of filipino children. *International Migration Review*, 44(2):300–319, 2010.

Dmitry Arkhangelsky and Guido Imbens. The role of the propensity score in fixed effect models. Technical report, National Bureau of Economic Research, 2018.

Dmitry Arkhangelsky and Guido W Imbens. Double-robust identification for causal panel data models. *arXiv preprint arXiv:1909.09412*, 2019.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

Fred Arnold and Nasra M Shah. Asian labor migration to the middle east. *International Migration Review*, 18(2):294–318, 1984.

Susan Athey and Guido W Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research, 2018.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research, 2018.

Marcus A Bachhuber, Brendan Saloner, Chinazo O Cunningham, and Colleen L Barry. Medical cannabis laws and opioid analgesic overdose mortality in the united states, 1999-2010. *JAMA internal medicine*, 174(10):1668–1673, 2014.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

Timothy J Bartik. Who benefits from state and local economic development policies? 1991.

Graziano Battistella and Ma Cecilia G Conaco. The impact of labour migration on the children left behind: a study of elementary school children in the philippines. *SOJOURN: Journal of Social Issues in Southeast Asia*, pages 220–241, 1998.

Christopher F Baum, Mark E Schaffer, and Steven Stillman. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal*, 7(4):465–506, 2007.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57 (1):289–300, 1995.

Matthew Blackwell and Soichiro Yamauchi. Adjusting for unmeasured confounding in marginal structural models with propensity-score fixed effects. *arXiv preprint arXiv:2105.03478*, 2021.

Iavor Bojinov, Ashesh Rambachan, and Neil Shephard. Panel experiments and dynamic causal effects: A finite population perspective. *arXiv preprint arXiv:2003.09915*, 2020a.

Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Available at SSRN 3684168*, 2020b.

Margaret Zoller Booth. Children of migrant fathers: The effects of father absence on swazi children's preparedness for school. *Comparative Education Review*, 39(2):195–210, 1995.

Kirill Borusyak and Xavier Jaravel. Revisiting event study designs. *Available at SSRN 2826228*, 2017.

John Bryant. Children of international migrants in indonesia, thailand, and the philippines: A review of evidence and policies, 2005.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods and an application on the minimum wage and employment. *arXiv preprint arXiv:1803.09015*, 2018.

Viviana Celli. Causal mediation analysis in economics: objectives, assumptions, models. Technical report, 2019.

Hongqin Chang, Xiao-yuan Dong, and Fiona MacPhail. Labor migration and time use patterns of the left-behind children and elderly in rural china. *World Development*, 39(12):2199–2210, 2011.

Joyce J Chen. Identifying non-cooperative behavior among spouses: child outcomes in migrant-sending households. *Journal of Development Economics*, 100(1):1–18, 2013.

Xi Chen and William D Nordhaus. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594, 2011.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural

parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*, 2020.

Raj Chetty, John N Friedman, Nathaniel Hendren, Michael Stepner, and The Opportunity Insights Team. *How did COVID-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data*. National Bureau of Economic Research Cambridge, MA, 2020.

Yin-Wong Cheung and Kon S Lai. Lag order and critical values of the augmented dickey–fuller test. *Journal of Business & Economic Statistics*, 13(3):277–280, 1995.

Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368(6489): 395–400, 2020.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Thomas A Croft. Nighttime images of the earth from space. *Scientific American*, 239(1):86–101, 1978.

Janet Currie, Henrik Kleven, and Esmée Zwiers. Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, volume 110, pages 42–48, 2020.

Clément de Chaisemartin and Xavier d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. Technical report, National Bureau of Economic Research, 2019.

Clement De Chaisemartin and Xavier d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96, 2020.

Angus Deaton and Alan Heston. Understanding ppps and ppp-based national accounts. *American Economic Journal: Macroeconomics*, 2(4):1–35, 2010.

Sylvie Démurger. Migration and families left behind. *IZA World of Labor*, 2015.

Christopher NH Doll, Jan-Peter Muller, and Jeremy G Morley. Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics*, 57(1):75–92, 2006.

Alejandra Cox Edwards and Manuelita Ureta. International migration, remittances, and schooling: evidence from el salvador. *Journal of development economics*, 72(2):429–461, 2003.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Christopher D Elvidge, Kimberley E Baugh, Eric A Kihn, Herbert W Kroehl, Ethan R Davis, and Chris W Davis. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*, 18(6): 1373–1379, 1997.

Philip A Ernst, LCG Rogers, and Quan Zhou. The distribution of yule's" nonsense correlation". *arXiv preprint arXiv:1909.02546*, 2019.

Robert C. Feenstra, Robert Inklaar, and Marcel P. Timmer. The next generation of the penn world table. *American Economic Review*, 105(10):3150–3182, 2015. URL http://www.ggdc.net/pwt/.

Matthew S Fritz and David P MacKinnon. Required sample size to detect the mediated effect. *Psychological science*, 18(3):233–239, 2007.

Varuni Ganepola. The psychosocial wellbeing of families left behind by asylum migration–the sri lankan experience. In *WIDER Conference on Poverty, International Migration and Asylum*, pages 27–28, 2002.

Tilottama Ghosh, Sharolyn Anderson, Rebecca L Powell, Paul C Sutton, and Christopher D Elvidge. Estimation of mexico's informal economy and remittances using nighttime imagery. *Remote Sensing*, 1(3):418–444, 2009.

Tilottama Ghosh, Rebecca L Powell, Christopher D Elvidge, Kimberly E Baugh, Paul C Sutton, and Sharolyn Anderson. Shedding light on the global distribution of economic activity. *The Open Geography Journal*, 3(1), 2010.

Paul Goldsmith-Pinkham, Isaac Sorkin, and Henry Swift. Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624, 2020.

Xiaodong Gong, Sherry Tao Kong, Shi Li, and Xin Meng. Rural-urban migrants: A driving force for growth. *China's Dilemma: Economic Growth, the Environment and Climate Change*, pages 110–152, 2008.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research, 2018.

Andrew Goodman-Bacon and Jan Marcus. Using difference-in-differences to identify causal effects of covid-19 policies. 2020.

Elspeth Graham and Lucy P Jordan. Migrant parents and the psychological well-being of left-behind children in southeast asia. *Journal of Marriage and Family*, 73(4):763–787, 2011.

Clive WJ Granger and Paul Newbold. Spurious regressions in econometrics. *Journal of econometrics*, 2(2):111–120, 1974.

Sanjeev Gupta, Catherine A Pattillo, and Smita Wagh. Effect of remittances on poverty and financial development in sub-saharan africa. *World development*, 37(1):104–115, 2009.

Gordon H Hanson and Christopher Woodruff. Emigration and educational attainment in mexico. Technical report, Mimeo., University of California at San Diego, 2003.

Mariaflavia Harari. Cities in bad shape: Urban geometry in india. *American Economic Review*, 110 (8):2377–2421, 2020.

Andrew F Hayes and Michael Scharkow. The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological science*, 24(10):1918–1927, 2013.

Bingyan He, Jingyi Fan, Ni Liu, Huijuan Li, Yanjun Wang, Joshua Williams, and Kaishing Wong. Depression risk of left-behind children in rural china. *Psychiatry research*, 200(2-3):306–312, 2012.

James Heckman, Rodrigo Pinto, and Peter Savelyev. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103 (6):2052–86, 2013.

James J Heckman and Rodrigo Pinto. Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric reviews*, 34(1-2):6–31, 2015.

J Vernon Henderson, Adam Storeygard, and David N Weil. Measuring economic growth from outer space. *American economic review*, 102(2):994–1028, 2012.

J Vernon Henderson, Tim Squires, Adam Storeygard, and David Weil. The global distribution of economic activity: nature, history, and the role of trade. *The Quarterly Journal of Economics*, 133 (1):357–406, 2018.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Roland Hodler and Paul A Raschky. Regional favoritism. *The Quarterly Journal of Economics*, 129 (2):995–1033, 2014.

Wassily Hoeffding. A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, pages 558–566, 1951.

David Holtz, Michael Zhao, Seth G Benzell, Cathy Y Cao, Mohammad Amin Rahimian, Jeremy Yang, Jennifer Allen, Avinash Collis, Alex Moehring, and Tara Sowrirajan. Interdependence and the cost of uncoordinated responses to covid-19. *Proceedings of the National Academy of Sciences*, 117(33):19837–19843, 2020.

Yingyao Hu and Jiaxiong Yao. Illuminating economic growth. *Journal of Econometrics*, 2021.

Martin Huber. Disentangling policy effects into causal channels. *IZA World of Labor*, 2016.

Martin Huber. A review of causal mediation analysis for assessing direct and indirect treatment effects. Technical report, Université de Fribourg, 2019.

Kosuke Imai and In Song Kim. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63(2):467–490, 2019.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press, 2015.

Guido Imbens. The role of the propensity score in estimating dose–response functions. *Biometrika*, 87(0):706–710, 2000.

Institute of Labor Economics (IZA), Australian National University, Pthe University of Queensland, and the Beijing Normal University. Longitudinal survey on rural urban migration in china (rumic) (2008-2009). *International Data Service Center of IZA (IDSC). Version 1.0*, 2014. doi: 10.15185/izadp.7680.1.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

William Kandel and Grace Kao. Shifting orientations: How us labor migration affects children's aspirations in mexican migrant communities. *Social science quarterly*, pages 16–32, 2000.

Joseph Kang and Joseph Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4): 523–539, 2007.

Jeni Klugman. Human development report 2009. overcoming barriers: Human mobility and development. *Overcoming Barriers: Human Mobility and Development (October 5, 2009). UNDP-HDRO Human Development Reports*, 2009.

John E Knodel and Chanpen Saengtienchai. *AIDS and older persons: The view from Thailand*. Population Studies Center, University of Michigan, 2002.

Sherry Tao Kong. Rural–urban migration in china: Survey design and implementation. *Chapters*, 2010.

Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, and William P Hanage. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1-3):159–178, 1992.

Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769, 2017.

Gilberto A Libanio. Unit roots in macroeconomic time series: theory, implications, and evidence. *Nova Economia*, 15(3):145–176, 2005.

MB Maruja and Asis Fabio Baggio. The other face of migration: children and families left behind. 2003.

Bennett T McCallum. Unit roots in macroeconomic time series: Some critical issues, 1993.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1973.

David McKenzie and Hillel Rapoport. Can migration reduce educational attainment? evidence from mexico. *Journal of Population Economics*, 24(4):1331–1358, 2011.

Diana Mendoza. Migrant workers' children left behind, left out. *IPS news*, 2004.

Xin Meng and Chikako Yamauchi. Children of migrants: The impact of parental migration on their children's education and health outcomes. 2015.

Xin Meng, Sherry Tao Kong, and Dandan Zhang. How much do we know about the impact of the economic downturn on the employment of migrants? Technical report, ADBI Working Paper,

2010.

Stelios Michalopoulos and Elias Papaioannou. Pre-colonial ethnic institutions and contemporary african development. *Econometrica*, 81(1):113–152, 2013.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated by Dabrowska, D. M. and Speed, T. P. *Statistical Science*, 5:465–472, 1923/1990.

Liem Nguyen, Brenda SA Yeoh, and Mika Toyota. Migration and the well-being of the left behind in asia: Key themes and trends. *Asian Population Studies*, 2(1):37–44, 2006.

Yoko Niimi, Thai Hung Pham, and Barry Reilly. Determinants of remittances: Recent evidence using data on internal migrants in vietnam. *Asian Economic Journal*, 23(1):19–39, 2009.

Esbjörn Ohlsson. Asymptotic normality for two-stage sampling from a finite population. *Probability theory and related fields*, 81(3):341–352, 1989.

Karen Fog Olwig. Narratives of the children left behind: Home and identity in globalised caribbean families. *Journal of ethnic and migration studies*, 25(2):267–284, 1999.

Nicole E Pashley and Luke W Miratrix. Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, 46(3):271–296, 2021.

Evan Patterson and Matteo Sesia. *knockoff: The Knockoff Filter for Controlled Variable Selection*, 2018. URL `https://CRAN.R-project.org/package=knockoff`. R package version 0.3.2.

VV Petrov. Sums of independent random variables. *Yu. V. Prokhorov. V. StatuleviCius (Eds.)*, 1975.

Peter CB Phillips. New tools for understanding spurious regressions. *Econometrica*, pages 1299–1325, 1998.

Jeremy Proville, Daniel Zavala-Araiza, and Gernot Wagner. Night-time lights: A global, long term look at links to socio-economic trends. *PloS one*, 12(3):e0174610, 2017.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10 (3-4):441–451, 1959.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Jonathan Roth and Pedro HC Sant'Anna. Efficient estimation for staggered rollout designs. *arXiv preprint arXiv:2102.01291*, 2021.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 2020.

David Schoenfeld. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153, 1980.

Azeem Shaikh and Panagiotis Toulis. Randomization tests in observational studies with staggered adoption of treatment. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2019-144), 2019.

Chelsea L Shover, Corey S Davis, Sanford C Gordon, and Keith Humphreys. Association between medical cannabis laws and opioid overdose mortality has reversed over time. *Proceedings of the National Academy of Sciences*, 116(26):12624–12626, 2019.

James H Stock, Mark W Watson, et al. *Introduction to econometrics*, volume 3. Pearson New York, 2012.

Xiaojun Sun, Yuan Tian, Yongxin Zhang, Xiaochun Xie, Melissa A Heath, and Zongkui Zhou. Psychological development and educational problems of left-behind children in rural china. *School Psychology International*, 36(3):227–252, 2015.

Paul C Sutton and Robert Costanza. Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. *Ecological Economics*, 41 (3):509–527, 2002.

Paul C Sutton, Christopher D Elvidge, Tilottama Ghosh, et al. Estimation of gross domestic product at sub-national scales using nighttime satellite imagery. *International Journal of Ecological Economics & Statistics*, 8(S07):5–21, 2007.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the r-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36:299–303, 1965.

Clare Waddington. Livelihood outcomes of migration for poor people. 2003.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Jeffrey M Wooldridge. *Introductory econometrics: A modern approach*. Cengage learning, 2015.

AO Xiang, Dawei Jiang, and ZHAO Zhong. The impact of rural–urban migration on the health of the left-behind parents. *China Economic Review*, 37:126–139, 2016.

Alwyn Young. The african growth miracle. *Journal of Political Economy*, 120(4):696–739, 2012.

Achim Zeileis. *pwt10: Penn World Table (Version 10.x)*, 2021. URL https://CRAN.R-project.org/package=pwt10. R package version 10.0-0.

Qiran Zhao, Xiaohua Yu, Xiaobing Wang, and Thomas Glauben. The impact of parental migration on children's school performance in rural china. *China Economic Review*, 31:43–54, 2014.