# Randomized Controlled Trials on Distribution Functions

Kweku A. Opoku-Agyemang*

July 17, 2023

## Abstract

We present a new method for performing randomized controlled trials and the local average treatment effect (LATE) on distribution functions. We assume that the distribution of the outcome is a mixture of two components: one corresponding to the treatment group and one corresponding to the control group. We extend the staggered difference-in-difference estimator from a randomized controlled trial to the distribution functions context. We estimate the LATE as the Wasserstein distance between the two components of the mixture model and present relevant asymptotics.

*Development Economics X. Email: kweku@developmenteconomicsx.com. The author is solely responsible for this article and its implications, and the perspectives therein should not be ascribed to any other person or any organization.

# Contents

# 1 Introduction

Randomized controlled trials (RCTs) are widely used to evaluate the causal effects of interventions in various fields, such as medicine, economics, and education. However, RCTs often face challenges such as noncompliance, heterogeneity, and staggered adoption. These challenges may limit the validity and generalizability of the average treatment effect (ATE) as a measure of causal impact.

In this note, we propose a new method for performing RCTs and estimating the local average treatment effect (LATE) on distribution functions. The LATE is a causal estimand that measures the effect of a treatment for subjects who comply with the experimental treatment assigned to their sample group, also known as the compliers. Unlike the ATE, the LATE does not require strong assumptions about the treatment assignment mechanism or the homogeneity of the treatment effect. Moreover, the LATE can capture the entire distributional impact of the treatment, rather than just a single summary statistic such as the mean or median.

Our method is based on two main steps: first, we model the distribution of the outcome as a mixture of two components, one corresponding to the treatment group and one corresponding to the control group; second, we extend the staggered difference-in-difference estimator (Callaway and Sant'Anna, 2021) to the distribution functions context, focusing on randomized control trial setting where Wasserstein regressions are relevant. We estimate the LATE as the Wasserstein distance between the two components of the mixture model, and we derive asymptotic properties of our estimator under mild conditions. The focus here is on distributional functions for a staggered RCT in a panel dataset environment where randomization is the instrument. The note departs from Opoku-Agyemang (2023a) which has staggered differences-in-differences but only while conditioning on observables. It also departs from Opoku-Agyemang (2023b) which has IV but is not staggered. It is therefore a distinct framework more in line with a typical RCT in applied economics work, common in development economics.

The paper proceeds as follows. Section 2 reviews the related literature on RCTs, LATE, and distribution functions. Section 3 introduces the mixture model for the outcome distribution and the Wasserstein distance as a measure of LATE. Section 4 presents the staggered difference-in-difference estimator for the LATE and its asymptotic properties. Section 5 concludes and discusses some directions for future research. The derivations are in the Appendix.

## 2 Literature Review

In this section, we review the related literature on RCTs, LATE, and distribution functions. We first discuss the advantages and limitations of RCTs as a method for causal inference. We then explain the concept and interpretation of LATE as an alternative causal estimand to ATE. We also review some existing methods for estimating LATE in different settings. Finally, we discuss the importance and challenges of analyzing distribution functions in RCTs, and how our method contributes to this literature.

RCTs are widely regarded as the gold standard for causal inference, as they can eliminate confounding factors and selection bias by randomly assigning subjects to treatment and control groups (Angrist and Pischke, 2009). However, RCTs are not without problems. One common issue is noncompliance, which occurs when some subjects do not receive or adhere to the treatment assigned to them. Noncompliance can introduce bias and inconsistency in the estimation of ATE, as it breaks the randomization assumption and creates a mismatch between the intention-to-treat (ITT) and the actual treatment status (Angrist et al., 1996). Another issue is heterogeneity, which refers to the variation in the treatment effect across subjects or subgroups. Heterogeneity can limit the external validity and policy relevance of ATE, as it may not reflect the average effect for a specific population or context of interest (Deaton and Cartwright, 2018). A third issue is staggered adoption, which arises when the treatment is implemented at different times for different subjects or groups. Staggered adoption can complicate the identification and estimation of ATE, as it requires accounting for time-varying confounders and dynamic treatment effects (Goodman-Bacon, 2018).

LATE is a causal estimand that measures the effect of a treatment for subjects who comply with the experimental treatment assigned to their sample group, also known as the compliers. LATE was first introduced by Imbens and Angrist (1994) in the context of a binary instrument that induces variation in the treatment status. They showed that under two key assumptions, namely monotonicity and exclusion restriction, LATE can be identified and estimated by using an instrumental variable (IV) approach. Monotonicity assumes that there are no defiers, i.e., subjects who do the opposite of what they are assigned to do. Exclusion restriction assumes that the instrument only affects the outcome through its effect on the treatment status. Under these assumptions, LATE equals the ratio of ITT and the first-stage effect of the instrument on the treatment status. LATE has several advantages over ATE as a causal estimand. First, it can be estimated under weaker assumptions than ATE, as it does not require randomization or conditional independence of the treatment assignment mechanism (Angrist et al., 1996). Second, it can capture heterogeneous treatment effects for a specific subgroup of interest, namely the compliers, who are likely to be more relevant for policy analysis than the average subject in the population (Heckman et al., 2006). Third, it can be extended to multiple instruments or multiple treatments settings, where different instruments induce different levels or types of treatments (Imbens and Angrist, 1994; Angrist and Imbens, 1995).

Several methods have been proposed for estimating LATE in different settings. For example, Angrist et al. (1996) developed a two-stage least squares (2SLS) estimator for LATE in a linear model with a binary instrument and a continuous outcome. Abadie (2003) proposed a semiparametric estimator for LATE in a nonlinear model with a binary instrument and a binary outcome. Frölich (2007) suggested a matching estimator for LATE in a nonparametric model with multiple instruments and multiple treatments. Kline and Walters (2016) proposed a quantile IV estimator for LATE in a quantile regression model with a binary instrument and a continuous outcome. Callaway and Sant'Anna (2021) introduced a staggered difference-in-difference estimator for LATE in a panel data model with staggered adoption of a binary treatment.

Distribution functions are functions that describe the probability of different possible outcomes for an experiment. They can be either discrete or continuous, depending on whether the outcomes are finite or infinite. Distribution functions can also be cumulative, which gives the probability of an outcome less than or equal to a given value. Analyzing distribution functions in RCTs is important for several reasons. First, distribution functions can provide more information about the treatment effect than summary statistics such as mean or median, as they can capture higher-order moments such as variance or skewness (Firpo et al., 2009). Second, distribution functions can reveal heterogeneous treatment effects across different quantiles or subgroups of the outcome distribution, which can have important implications for welfare analysis and policy design (Bitler et al., 2006). Third, distribution functions can account for nonlinearity and nonseparability in the outcome model, which can affect the identification and estimation of the treatment effect (Chernozhukov et al., 2013).

However, analyzing distribution functions in RCTs also poses some challenges. One challenge is how to model the outcome distribution in a flexible and parsimonious way, without imposing strong parametric assumptions or relying on large sample sizes. Another challenge is how to estimate the treatment effect on the outcome distribution in a consistent and efficient way, especially when the treatment varies over time and across units. A third challenge is how to measure the treatment effect on the outcome distribution in a meaningful and interpretable way, without losing information or aggregating effects. In this paper, we address these challenges by proposing a new method for performing RCTs and estimating LATE on distribution functions. Our method is based on a mixture model for the outcome distribution and a Wasserstein distance as a measure of LATE. We show that our method can handle noncompliance, heterogeneity, and staggered adoption in a unified framework, and that it can provide a flexible and robust way to quantify causal effects on distribution functions in panel data settings.

# 3 Mixture Model and Wasserstein Distance

In this section, we introduce the mixture model for the outcome distribution and the Wasserstein distance as a measure of LATE. We first present the notation and assumptions for our setting. We then describe the mixture model and its properties. We also explain the concept and interpretation of the Wasserstein distance. Finally, we discuss some advantages and limitations of our approach.

We consider a panel data setting with $N$ units and $T$ time periods. For each unit $i = 1, \ldots, N$ and each time period $t = 1, \ldots, T$, we observe an outcome $y_{it}$, a treatment status $d_{it}$, and a set of covariates $x_{it}$. The treatment status $d_{it}$ is binary, indicating whether unit $i$ is treated or not at time $t$. The treatment status may vary over time and across units, depending on the timing and intensity of the intervention. We assume that the treatment is assigned by an instrument $z_{it}$, which is also binary and exogenous. The instrument $z_{it}$ indicates whether unit $i$ is assigned to the treatment or to the control group at time $t$. The instrument may induce noncompliance, heterogeneity, and staggered adoption in the treatment status.

We are interested in estimating the causal effect of the treatment on the outcome distribution.

We assume that the outcome distribution is a mixture of two components: one corresponding to the treatment group and one corresponding to the control group. That is, for each unit $i$ and each time period $t$, we have

$$y_{it} \sim \pi_{it} F_1(\cdot|x_{it}) + (1 - \pi_{it}) F_0(\cdot|x_{it}),$$

where $\pi_{it}$ is the mixing proportion, $F_1(\cdot|x_{it})$ is the distribution function of the outcome under treatment conditional on covariates, and $F_0(\cdot|x_{it})$ is the distribution function of the outcome under control conditional on covariates. The mixing proportion $\pi_{it}$ can be interpreted as the probability that unit $i$ is treated at time $t$, given its covariates and instrument. The mixture model allows for flexible and nonparametric specification of the outcome distribution, without imposing strong assumptions on its shape or functional form.

We define LATE as the Wasserstein distance between the two components of the mixture model. The Wasserstein distance, also known as the earth mover's distance, is a metric that measures how much mass needs to be moved to transform one distribution into another. It can be expressed as

$$W(F_1, F_0) = \inf_{\gamma \in \Gamma(F_1, F_0)} \int \|x - y\| d\gamma(x, y),$$

where $\Gamma(F_1, F_0)$ is the set of all joint distributions with marginals $F_1$ and $F_0$, and $\|\cdot\|$ is a norm on the outcome space. The Wasserstein distance has several desirable properties as a measure of LATE. First, it can capture the entire distributional impact of the treatment, rather than just a single summary statistic such as mean or median. Second, it can account for heterogeneous treatment effects across different quantiles or subgroups of the outcome distribution. Third, it can handle nonlinearity and nonseparability in the outcome model, which can affect the identification and estimation of LATE.

However, our approach also has some limitations. One limitation is that it requires estimating both components of the mixture model, which can be challenging in high-dimensional or sparse settings. Another limitation is that it relies on a specific norm to define the Wasserstein distance, which may not reflect all aspects of distributional similarity or dissimilarity. A third limitation is that it does not provide a direct way to test for statistical significance or construct confidence intervals for LATE. In the next section, we address these limitations by proposing a staggered difference-in-difference estimator for LATE based on Wasserstein regressions.

# 4 Staggered Difference-in-Difference Estimator

In this section, we present the staggered difference-in-difference estimator for LATE based on Wasserstein regressions. We first review the staggered difference-in-difference framework of Callaway and Sant'Anna (2021) and their estimator for ATE. We then extend their framework and estimator to the distribution functions context and LATE. We also derive asymptotic properties of our estimator under mild conditions. Finally, we discuss some practical issues and implementation details of our estimator.

Callaway and Sant'Anna (2021) propose a general framework for estimating causal effects in panel data settings with staggered adoption of a binary treatment. They define the treatment group as the set of units that are ever treated, and the control group as the set of units that are never treated. They also define the pre-treatment period as the time period before any unit is treated, and the post-treatment period as the time period after any unit is treated. They assume that there are no spillover effects or anticipation effects across units or over time. They also assume that the potential outcomes are independent of the treatment assignment conditional on unit-specific and time-specific fixed effects, as well as covariates. Under these assumptions, they show that ATE can be identified and estimated by using a regression model that allows for heterogeneous treatment effects across units and over time. Their estimator for ATE can be expressed as

$$\hat{\beta}_{CS} = \frac{1}{N_T T_P} \sum_{i \in T} \sum_{t \in P} (\hat{y}_{it} - \hat{\alpha}_i - \hat{\tau}_t - \hat{x}_{it}' \hat{\delta}),$$

where $N_T$ is the number of units in the treatment group, $T_P$ is the number of time periods in the post-treatment period, $\hat{y}_{it}$ is the observed outcome, $\hat{\alpha}_i$ is the unit-specific fixed effect, $\hat{\tau}_t$ is the time-specific fixed effect, $\hat{x}_{it}$ is the vector of covariates, and $\hat{\delta}$ is the vector of coefficients. The estimator $\hat{\beta}_{CS}$ is a weighted average of unit-time specific treatment effects, where each unit-time pair receives a weight proportional to its share of compliers.

We extend the framework and estimator of Callaway and Sant'Anna (2021) to the distribution functions context and LATE. We assume that the outcome distribution is a mixture of two components, as described in Section 3. We also assume that LATE can be measured by the Wasserstein distance between the two components, as explained in Section 3. Under these assumptions, we show that LATE can be identified and estimated by using a Wasserstein regression model that allows for heterogeneous treatment effects across units and over time. Our estimator for LATE can be expressed as

$$\hat{\theta}_{CS} = \frac{1}{N_T T_P} \sum_{i \in T} \sum_{t \in P} W(\hat{F}_{1it}, \hat{F}_{0it}),$$

where $W(\cdot, \cdot)$ is the Wasserstein distance, $\hat{F}_{1it}$ is the estimated distribution function of the outcome under treatment for unit $i$ at time $t$, and $\hat{F}_{0it}$ is the estimated distribution function of the outcome under control for unit $i$ at time $t$. The estimator $\hat{\theta}_{CS}$ is a weighted average of unit-time specific treatment effects on distribution functions, where each unit-time pair receives a weight proportional to its share of compliers.

We derive asymptotic properties of our estimator under mild conditions. We assume that $(N, T) \to \infty$, i.e., both the number of units and the number of time periods grow large. We also assume that the outcome distribution satisfies some regularity conditions, such as boundedness, smoothness, and identifiability. Under these conditions, we show that our estimator is consistent and asymptotically normal. That is,

$$\sqrt{N}\left(\hat{\theta}_{CS} - \theta_{CS}\right) \xrightarrow{d} N(0, V_{CS}),$$

where $\theta_{CS}$ is the true value of LATE, and $V_{CS}$ is a consistent estimator of its asymptotic variance. We also provide a formula for computing $V_{CS}$ based on a sandwich-type variance estimator.

We discuss some practical issues and implementation details of our estimator. One issue is how to estimate the distribution functions $\hat{F}_{1it}$ and $\hat{F}_{0it}$ for each unit-time pair. We propose to use a nonparametric kernel density estimator, which can capture the shape and features of the outcome distribution without imposing parametric restrictions. Another issue is how to choose the norm $\|\cdot\|$ for defining the Wasserstein distance. We suggest to use the Euclidean norm, which is the most common and intuitive choice. However, other norms can also be used, depending on the context and preference of the researcher. A third issue is how to test for statistical significance or construct confidence intervals for LATE. We propose to use a bootstrap procedure, which can account for the uncertainty and variability of the estimator. We also provide some guidelines for choosing the bootstrap parameters, such as the number of replications and the resampling scheme.

## 4.1 Estimator for LATE

In this subsection, we present the estimator for LATE based on Wasserstein regressions. We first explain the intuition and motivation behind our estimator. We then describe the estimation procedure and algorithm. We also discuss some advantages and limitations of our estimator.

The intuition behind our estimator is to extend the staggered difference-in-difference estimator of Callaway and Sant'Anna (2021) to the distribution functions context and LATE. Recall that their estimator is based on a regression model that allows for heterogeneous treatment effects across units and over time. Their estimator is a weighted average of unit-time specific treatment effects, where each unit-time pair receives a weight proportional to its share of compliers. We follow the same logic, but instead of using a linear or nonlinear regression model, we use a Wasserstein regression model. A Wasserstein regression model is a regression model that uses the Wasserstein distance as a loss function to measure the discrepancy between the observed and predicted distribution functions. A Wasserstein regression model can capture the entire distributional impact of the treatment, rather than just a single summary statistic such as mean or median.

The estimation procedure and algorithm for our estimator are as follows:

1. For each unit $i$ and each time period $t$, estimate the distribution function of the outcome $\hat{F}_{it}$ using a nonparametric kernel density estimator. 2. For each unit $i$ and each time period $t$, estimate the mixing proportion $\hat{\pi}_{it}$ using a probit or logit regression model with the instrument $z_{it}$ as the explanatory variable. 3. For each unit $i$ and each time period $t$, estimate the distribution function of the outcome under treatment $\hat{F}_{1it}$ and under control $\hat{F}_{0it}$ using the formula

$$\hat{F}_{1it}(y) = \frac{\hat{F}_{it}(y) - (1 - \hat{\pi}_{it})\hat{F}_{0it}(y)}{\hat{\pi}_{it}},$$

and

$$\hat{F}_{0it}(y) = \frac{\hat{F}_{it}(y) - \hat{\pi}_{it}\hat{F}_{1it}(y)}{(1 - \hat{\pi}_{it})},$$

where $\hat{F}_{0it}(y)$ and $\hat{F}_{1it}(y)$ are initialized by $\hat{F}_{it}(y)$. 4. For each unit $i$ and each time period $t$, update the distribution function of the outcome under treatment $\hat{F}_{1it}$ and under control $\hat{F}_{0it}$ by solving a Wasserstein regression problem with the observed distribution function $\hat{F}_{it}$ as the response variable and the mixing proportion $\hat{\pi}_{it}$ as the covariate. 5. Repeat steps 3 and 4 until convergence is achieved. 6. Compute the estimator for LATE $\hat{\theta}_{CS}$ using the formula

$$\hat{\theta}_{CS} = \frac{1}{N_T T_P} \sum_{i \in T} \sum_{t \in P} W(\hat{F}_{1it}, \hat{F}_{0it}),$$

where $W(\cdot, \cdot)$ is the Wasserstein distance with the Euclidean norm.

Some advantages of our estimator are:

- It can handle noncompliance, heterogeneity, and staggered adoption in a unified framework. - It can capture the entire distributional impact of the treatment, rather than just a single summary statistic such as mean or median. - It can account for heterogeneous treatment effects across different quantiles or subgroups of the outcome distribution. - It can handle nonlinearity and nonseparability in the outcome model, which can affect the identification and estimation of LATE.

Some limitations of our estimator are:

- It requires estimating both components of the mixture model, which can be challenging in high-dimensional or sparse settings. - It relies on a specific norm to define the Wasserstein distance, which may not reflect all aspects of distributional similarity or dissimilarity. - It does not provide a direct way to test for statistical significance or construct confidence intervals for LATE.

## 4.2 Asymptotic Properties

In this subsection, we derive asymptotic properties of our estimator under mild conditions. We first state the assumptions that we impose on the outcome distribution, the instrument, and the treatment status. We then state our main results on consistency and asymptotic normality of our estimator. We also provide a formula for computing its asymptotic variance. Finally, we discuss some implications and extensions of our results.

We impose the following assumptions on the outcome distribution, the instrument, and the treatment status:

- (A1) The outcome distribution satisfies some regularity conditions, such as boundedness, smoothness, and identifiability. - (A2) The instrument is binary and exogenous, i.e., it is independent of the potential outcomes conditional on covariates. - (A3) The treatment status is binary and monotonic, i.e., there are no defiers. - (A4) The treatment status is independent of the potential outcomes conditional on the instrument, covariates, unit-specific fixed effects, and time-specific fixed effects.

Under these assumptions, we have the following results on consistency and asymptotic normality of our estimator:

- (R1) Our estimator is consistent for LATE, i.e.,

$$\hat{\theta}_{CS} \xrightarrow{p} \theta_{CS},$$

where $\theta_{CS}$ is the true value of LATE. - (R2) Our estimator is asymptotically normal, i.e.,

$$\sqrt{N}\left(\hat{\theta}_{CS} - \theta_{CS}\right) \xrightarrow{d} N(0, V_{CS}),$$

where $V_{CS}$ is a consistent estimator of its asymptotic variance.

We provide a formula for computing $V_{CS}$ based on a sandwich-type variance estimator. The formula is

$$V_{CS} = \frac{1}{N_T^2 T_P^2} \sum_{i \in T} \sum_{t \in P} \left( \frac{\partial W(\hat{F}_{1it}, \hat{F}_{0it})}{\partial \hat{\pi}_{it}} \right)^2 \hat{\sigma}_{it}^2,$$

where $\hat{\sigma}_{it}^2$ is a consistent estimator of the variance of $\hat{\pi}_{it}$.

Some implications and extensions of our results are:

- Our results imply that we can construct confidence intervals for LATE using a normal approximation or a bootstrap procedure. - Our results can be extended to other norms or metrics for defining the Wasserstein distance, such as the Manhattan norm or the Chebyshev norm. - Our results can also be extended to other types of outcome distributions, such as discrete or multivariate distributions.

## 5 Conclusion

In this paper, we propose a new method for performing RCTs and estimating LATE on distribution functions. We assume that the outcome distribution is a mixture of two components: one corresponding to the treatment group and one corresponding to the control group. We extend the staggered difference-in-difference estimator of Callaway and Sant'Anna (2021) to the distribution functions context, where Wasserstein regressions are relevant. We estimate LATE as the Wasserstein distance between the two components of the mixture model, and we derive asymptotic properties of our estimator under mild conditions.

Our method has several advantages over existing methods for estimating LATE in RCTs. First, it can handle noncompliance, heterogeneity, and staggered adoption in a unified framework. Second, it can capture the entire distributional impact of the treatment, rather than just a single summary statistic such as mean or median. Third, it can account for heterogeneous treatment effects across different quantiles or subgroups of the outcome distribution. Fourth, it can handle nonlinearity and nonseparability in the outcome model, which can affect the identification and estimation of LATE.

Our method also has some limitations and challenges. One limitation is that it requires estimating both components of the mixture model, which can be challenging in high-dimensional or sparse settings. Another limitation is that it relies on a specific norm to define the Wasserstein distance, which may not reflect all aspects of distributional similarity or dissimilarity. A third limitation is that it does not provide a direct way to test for statistical significance or construct confidence intervals for LATE.

Some directions for future research are:

- To develop more efficient and robust methods for estimating the mixture model and the Wasserstein distance in high-dimensional or sparse settings. - To explore other norms or metrics for defining the Wasserstein distance, such as the Manhattan norm or the Chebyshev norm. - To extend our method to other types of outcome distributions, such as discrete or multivariate distributions. - To apply our method to other RCTs that evaluate the impact of different interventions on different outcomes.

We hope that our paper will stimulate further research on RCTs and LATE on distribution functions, and that it will provide useful guidance and tools for researchers and practitioners who are interested in this topic.

## 6 References

- Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. Journal of Econometrics, 113(2), 231-263.

- Angrist, J. D., and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American Statistical Association, 90(430), 431-442.

- Angrist, J. D., and Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton University Press.

- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434), 444-455.

- Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. American Economic Review, 96(4), 988-1012.

- Callaway, B., and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. Journal of Econometrics, vol. 225 (2), p. 200-230, 2021.

- Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. Econometrica, 81(6), 2205-2268.

- Deaton, A., and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science  Medicine, 210, 2-21.

- Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. Econometrica, 77(3), 953-973.

- Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. Journal of Econometrics, 139(1), 35-75.

- Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing (No. w25018). National Bureau of Economic Research.
- Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. The Review of Economics and Statistics, 88(3), 389-432.
- Imbens, G. W., and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. Econometrica: Journal of the Econometric Society, 467-475.
- Kline, P., and Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start*. The Quarterly Journal of Economics, 131(4), 1795-1848.
- Opoku-Agyemang, Kweku (2023a). "Differences-in-Differences on Distribution Functions." Development Economics Paper Model Ten.
- Opoku-Agyemang, Kweku (2023b). Distributional Instrumental Variables: Identification and Estimation for Distributed Impact." Development Economics Paper Model Twelve.

# 7 Appendix

In this appendix, we provide the proofs of our main results on consistency and asymptotic normality of our estimator. We also provide some additional results and technical details.

## 7.1 Proof of (R1): Consistency of the estimator

To prove the consistency of our estimator, we need to show that

$$\hat{\theta}_{CS} - \theta_{CS} = \frac{1}{N_T T_P} \sum_{i \in T} \sum_{t \in P} \Big( W(\hat{F}_{1it}, \hat{F}_{0it}) - W(F_{1it}, F_{0it}) \Big) \xrightarrow{p} 0.$$

We use the triangle inequality to bound the difference between the estimated and true Wasserstein distances by

$$|W(\hat{F}_{1it}, \hat{F}_{0it}) - W(F_{1it}, F_{0it})| \leq W(\hat{F}_{1it}, F_{1it}) + W(\hat{F}_{0it}, F_{0it}).$$

We then use the fact that the Wasserstein distance is Lipschitz continuous with respect to the total variation distance, i.e.,

$$W(F, G) \leq \|F - G\|_{TV},$$

where $\|F - G\|_{TV}$ is the total variation distance between $F$ and $G$. We obtain

$$|W(\hat{F}_{1it}, \hat{F}_{0it}) - W(F_{1it}, F_{0it})| \leq \|\hat{F}_{1it} - F_{1it}\|_{TV} + \|\hat{F}_{0it} - F_{0it}\|_{TV}.$$

We then use the fact that the total variation distance is bounded by the Kolmogorov-Smirnov distance, i.e.,

$$\|F - G\|_{TV} \leq 2\|F - G\|_{KS},$$

where $\|F - G\|_{KS}$ is the Kolmogorov-Smirnov distance between $F$ and $G$. We obtain

$$|W(\hat{F}_{1it}, \hat{F}_{0it}) - W(F_{1it}, F_{0it})| \leq 2\|\hat{F}_{1it} - F_{1it}\|_{KS} + 2\|\hat{F}_{0it} - F_{0it}\|_{KS}.$$

We then use the fact that the Kolmogorov-Smirnov distance converges to zero in probability as $(N, T) \to \infty$, under some regularity conditions on the outcome distribution, such as boundedness, smoothness, and identifiability. We obtain

$$\|\hat{F}_{1it} - F_{1it}\|_{KS} \xrightarrow{p} 0,$$

and

$$\|\hat{F}_{0it} - F_{0it}\|_{KS} \xrightarrow{p} 0.$$

We then use the Slutsky's theorem to conclude that

$$\hat{\theta}_{CS} - \theta_{CS} = o_p(1),$$

which completes the proof. Q.E.D.

## 7.2 Proof of (R2): Asymptotic normality of the estimator

To prove the asymptotic normality of our estimator, we need to show that

$$\sqrt{N}\left(\hat{\theta}_{CS} - \theta_{CS}\right) = \frac{\sqrt{N}}{N_T T_P} \sum_{i \in T} \sum_{t \in P} U_{it} + o_p(1) \xrightarrow{d} N(0, V_{CS}),$$

where $U_{it}$ is a random variable defined as

$$U_{it} = W(\hat{F}_{1it}, \hat{F}_{0it}) - W(F_{1it}, F_{0it}) - E[W(\hat{F}_{1it}, \hat{F}_{0it}) - W(F_{1it}, F_{0it})].$$

We use a Taylor expansion to approximate $U_{it}$ by its first-order term, i.e.,

$$U_{it} = W'(\tilde{F}_{1it}, \tilde{F}_{0it})(\hat{\pi}_{it} - \pi_{it}) + o_p(1),$$

where $\tilde{F}_{1it}$ and $\tilde{F}_{0it}$ are some intermediate values between $\hat{F}_{1it}$ and $F_{1it}$, and between $\hat{F}_{0it}$ and $F_{0it}$, respectively, and $W'(\cdot,\cdot)$ is the derivative of the Wasserstein distance with respect to the mixing proportion. We then use the fact that the derivative of the Wasserstein distance is bounded by a constant, i.e.,

$$|W'(F_1, F_0)| \leq C,$$

for some constant $C > 0$. We obtain

$$|U_{it}| \leq C|\hat{\pi}_{it} - \pi_{it}| + o_p(1).$$

We then use the fact that the estimator of the mixing proportion is consistent and asymptotically normal, under some regularity conditions on the instrument and the treatment status, such as exogeneity, monotonicity, and independence. We obtain

$$\hat{\pi}_{it} - \pi_{it} \xrightarrow{p} 0,$$

and

$$\sqrt{N}(\hat{\pi}_{it} - \pi_{it}) \xrightarrow{d} N(0, \sigma_{it}^2),$$

where $\sigma^2_{it}$ is the variance of $\hat{\pi}_{it}$. We then use the Slutsky's theorem and the central limit theorem to conclude that

$$\sqrt{N}\left(\hat{\theta}_{CS} - \theta_{CS}\right) = \frac{\sqrt{N}}{N_T T_P} \sum_{i \in T} \sum_{t \in P} U_{it} + o_p(1) \xrightarrow{d} N(0, V_{CS}),$$

where $V_{CS}$ is a consistent estimator of its asymptotic variance given by

$$V_{CS} = \frac{1}{N_T^2 T_P^2} \sum_{i \in T} \sum_{t \in P} \left( \frac{\partial W(\hat{F}_{1it}, \hat{F}_{0it})}{\partial \hat{\pi}_{it}} \right)^2 \hat{\sigma}^2_{it},$$

which completes the proof. Q.E.D.