# A Stochastic Frontier Model with an Endogenous Treatment Status and a Mediator[*]

**Yi-Ting Chen**

Institute of Economics

Academia Sinica

**Yu-Chin Hsu**

Institute of Economics

Academia Sinica

**Hung-Jen Wang**

Department of Economics

National Taiwan University, and

Institute of Economics

Academia Sinica

This version: March 4, 2017

[*] We thank Editor Rajeev Dehejia, an Associate Editor, three referees, Martin Huber and the participants of the 2015 European Workshop on Efficiency and Productivity Analysis and the 2015 Annual Conference of the International Association of Applied Econometrics for very helpful comments and suggestions. Yu-Chin Hsu gratefully acknowledges the research support from the Ministry of Science and Technology of Taiwan (MOST103-2628-H-001-001-MY4) and the Career Development Award of Academia Sinica, Taiwan.

**Abstract**

Government policies are frequently used to promote productivity. Some policies are designed to enhance production technology, while others are meant to improve production efficiency. An important issue to consider when designing and evaluating policies is whether a mediator is required or effective at achieving the desired final outcome. To better understand and evaluate the policies, we propose a new stochastic frontier model with a treatment status and a mediator, both of which are allowed to be endogenous. The model would allow us to decompose the total program (treatment) effect into technology and efficiency components, and investigate whether the effect is derived directly from the program or indirectly through a particular mediator.

**JEL Classification:** C310, C510

**Keywords:** stochastic frontier models, treatment effect, mediation analysis

# 1 Introduction

Government policies are often instituted to increase productivity at both an industry and firm level in developing and developed countries, and the purposes are served either by promoting better technology or by improving production efficiency. There are ample examples of policies that are designed specifically for either (or both) such purposes. For instance, the Argentinean Technological Fund and Chile's National Productivity and Technological Development Fund were set up to help to improve competitiveness through technological innovation (Chudnovsky et al. 2006, Benavente et al. 2007). In Mexico, the Program for Training the Industrial Workforce sought to increase workforce efficiency for small and medium enterprises (Lopez-Acevedo and Tinajero-Bravo 2011). In the U.S., the Job Corps education and training program was implemented to help disadvantaged youths at work and in life (Schochet et al. 2008 and the references therein). Large dams were built in India to improve agricultural production (Duflo and Pande 2007a).

An important issue to consider when designing and/or evaluating a policy is whether a mediator is required/effective for achieving the final outcome. For example, the Job Corps program may increase participants' future labor income by improving their work skills. Building dams may improve agricultural production efficiency (while also changing the production method) through irrigation. In these two examples, work skills and irrigation systems are policy program *mediators*, and such policies may exert all or partial effects on the final outcome through the mediator.

Given that such policies are widely adopted by governments throughout the world, it is important to have methods with which to evaluate policy effectiveness. Does the policy increase productivity? In which productivity component –technology or efficiency– does the effect take place? Is there an important mediator for exerting the effect? Finally, does the policy work the way it was designed to? Answering the above questions requires that we propose a new stochastic frontier model with a treatment status and a mediator.

Stochastic frontier models are widely used in policy-related studies (e.g., Kleit and Terrell 2001, Greene 2005, and Wang et al. 2008). The model allows researchers to decompose observed outputs into the maximum possible output given the technology and the level of input (the "technology frontier") and the shortfall relative to the frontier (the "inefficiency" component). However, few of these studies are conducted in the treatment effect framework. An exception is Crespo-Cebada et al. (2014) in which the efficiency of Spanish schools was estimated using the propensity score matching method, but no endogeneity problem was mentioned. Discussions of the endogeneity issue in this literature usually center on the feedback between the choice of input and the model residuals. For example, Amsler et al. (2016) and Griffiths and Hajargasht (2016) both considered cases in which one or more of the inputs correlated with the model residuals and discussed estimation procedures modified from standard approaches. Glass et al. (2016) developed a spatial autoregressive model in the stochastic frontier context where the autoregressive term was endogenous and estimated by the maximum likelihood.

In this paper, we propose a new stochastic frontier model for which *endogenous* treatment could separately and simultaneously affect the model's technology frontier and the production inefficiency. Furthermore, the treatment is allowed to exert effects either directly on the two outcome components or indirectly via a mediator. The paper thus makes a contribution in that it brings together the literature on intermediation analysis in treatment effect models and the literature on stochastic frontier models. The new model allows us to examine whether and how a program might affect the frontier and the efficiency components of the outcome, and enables us to investigate whether influence is exerted directly from the program or indirectly via a mediator; i.e., it allows us to analyze a treatment effect from four different angles: direct effects on the technology frontier, indirect effects on the technology frontier that go through a mediator, direct effects on technical inefficiency, and indirect effects on inefficiency that go through a mediator.

A feature of our model is that it allows for endogenous treatment status, which is important because program participation is often voluntary. Following Imbens and Angrist (1994), we measured and identified the total effect in the subpopulation of compliers in the presence of the endogenous treatment status using a binary instrumental variable (IV). Abadie (2003), Frölich (2007), Hong and Nekipelov (2010), and Donald et al. (2014a, 2014b) extended the model of Imbens and Angrist (1994) to allow for the presence of covariates. However, none of these studies included a mediator, so mediation analysis is not allowed. In comparison, our model includes an endogenous binary mediator in addition to the endogenous treatment status. We then apply a recent method by Frölich and Huber (2014, 2017), who extended the aforementioned method to the context of mediation analysis, to identify the program's total effect, the direct effect, and the indirect effect on both of the frontier and inefficiency components of the model. Section 2.1 provides more discussions of the relationship between our model and the mediation analysis literature.

We show that the model parameters can be identified and estimated via a two-stage weighted non-linear least squares (WNLS) method. We also establish the asymptotic properties of the WNLS estimator (WNLSE). A Monte Carlo analysis is provided to show the performance of the WNLSE in finite samples. Attention is paid to include endogenous treatment effects and mediation effects in the data-generating process in a structural manner. The results show that the mean square error (MSE) of the WNLSE reasonably reduces as the sample size increases in various settings.

## The Dam Example

As an empirical illustration, we apply the proposed model to evaluate the effects of large dam constructions on local (the area in the vicinity of dams) agricultural production in India. A similar (but different) policy evaluation is studied by Duflo and Pande (2007a) and we use the same dataset as theirs. We use the application as a running example throughout the paper to illustrate the model.

India has one of the largest numbers of dams in the world, over 90% of which were built for agricultural irrigation. Proposals of dam construction are made in the state level governments. While large dams provide irrigation and flood control, the benefit is more likely captured by downstream districts (administrative units below the state in India) but not the upstream (local) districts. The differential effects give rise to the distributional issue discussed in Duflo and Pande (2007a). In this empirical example, our focus is on the upstream districts on which the impact of dams appears to be more controversial.

The mixed impacts of dams on local agricultural production are well discussed in the literature. Irrigation, flood control, and better infrastructure (paved roads, electricity, etc.) brought about by dam construction could improve production. However, waterlogging and soil salinization, increased incidence of water-borne and water-related diseases in residents, and problems with resettlement or changes in the lifestyle of local populations could all negatively impact production. In addition, the use of pesticides and chemical fertilizers is sometimes restricted in the vicinity of water reservoirs to protect the water quality; see Frenken and Faures (1997), McCully (2001), Singh (2002), and Bauder et al. (2010). In this empirical example, we assess the impact of building dams (treatment) on the agricultural production via their indirect effects through irrigation (the mediator) and the direct effect through all of the other mechanisms. Assessments are done with respect to both the technology frontier and the production inefficiency of the local agricultural industry.

The rest of the paper is organized as follows. Section 2 lays out the proposed model and defines the total, direct, and indirect effects. Identification results and estimation methods are respectively in Sections 3 and 4. We use constructions of dams in India and the effects on agricultural productions as an example to illustrate the model's empirical application. The application is provided as a running example throughout the paper to better connect the model and the example, and the estimation results are presented in Section 5. Finally, Section 6 concludes the paper. We collect the asymptotics of the estimators, a bootstrap procedure for statistical inference, a simulation study, and the mathematical proofs and derivations in a supplementary appendix for the sake of brevity. This appendix and the programs used in this paper are available upon request.

# 2 Model and Effects

Let $Y$ be the observable output of an individual or firm, and $X$ be a vector of the observable covariates. The stochastic frontier model for $Y|X$ has been widely applied to various empirical studies. This model is conventionally of the form:

$$\begin{aligned} Y &= Y^* - u, \quad u \geq 0, \\ Y^* &= h(X, \beta^h) + v, \end{aligned} \tag{2.1}$$

where $Y^*$ is the unobserved stochastic frontier, $u \geq 0$ is the unobserved production inefficiency, $h(\cdot)$ is a frontier function with parameter vector $\beta^h$, and $v$ is a pure random error. Equation (2.1) indicates that $Y^*$ and $u$ are two unobserved and distinct random components (as opposed to mediators) of output, in the framework of a stochastic frontier analysis.

The conditional means of the random components are specified as:

$$E[v|X] = 0, \tag{2.2}$$

$$E[Y^*|X] = h(X, \beta^h), \tag{2.3}$$

$$E[u|X] = g(X, \beta^g). \tag{2.4}$$

Here, $g(\cdot)$ is a non-negative inefficiency function with the parameter vector $\beta^g$. The theoretical framework allows $h(\cdot)$ and $g(\cdot)$ to have the same or different covariates, and we use a general $X$ in these two functions in the theoretical discussion for notational simplicity. It is well known in the stochastic frontier literature that the parameters $\beta^h$ and $\beta^g$ can be separately identified by either imposing restrictions on the vectors of $\beta^h$ and $\beta^g$ (so that $h(\cdot)$ and $g(\cdot)$ have different covariates) or by employing distinctive parametric forms of $h(\cdot)$ and $g(\cdot)$, or both. The conditional mean representation of (2.1) is thus

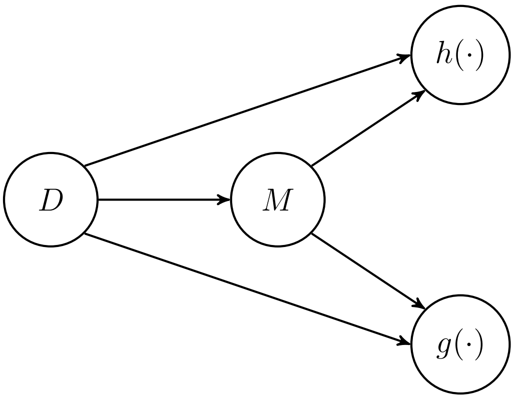
$$E[Y|X] = h(X, \beta^h) - g(X, \beta^g). \tag{2.5}$$

In the following, we discuss a modeling framework that allows us to extend this traditional model to a new model for program evaluation and mediation analysis with an endogenous treatment status and mediator.

## 2.1 Modeling Framework

Corresponding to $(Y, X^\top)^\top$, where "$X^\top$" is the transpose of $X$, we let $D$ be a binary random variable that represents the treatment status regarding the program being evaluated; $D$ has outcome $d = 1$ for treatment and $d = 0$ for no treatment. We also let $M$ be a binary random variable, with the outcomes $\{0, 1\}$, which represents a post-treatment mediator between $D$ and $Y$. We assume that the treatment assignment $D$ is endogenous, in that we cannot find a set of covariates available to us such that conditional on the covariates, $D$ is as good as a random assignment. Otherwise, we say that $D$ is exogenous. We also assume that $M$ is endogenous. We are interested in the program-evaluation problem - evaluating the effectiveness of $D$ on $Y$ (i.e., $h(\cdot)$ and $g(\cdot)$) and the mediation analysis - analyzing whether and how $D$ directly influences $Y$ and indirectly affects $Y$ via $M$. Since both tasks involve counterfactual assessments, we conduct the potential-outcome setting of $M$ and $Y$ to facilitate these analyses; see Frölich and Huber (2014) and Huber (2014) for examples of the setting.

We present the causal pathway considered by our modeling framework in Figure 1 for specificity. $M$ is a post-treatment variable in this causal pathway, and $h(\cdot)$ and



Note: $E[Y^*|X] = h(X, \beta^h)$ and $E[u|X] = g(X, \beta^g)$.

Figure 1: Causal pathway of the proposed model.

$g(\cdot)$ are two post-mediation outcomes that serve as key components of $Y$. This modeling framework allows us to combine treatment-and-mediation analysis for the sub-pathways of $(D, M, h)$ and $(D, M, g)$, and stochastic frontier analysis for the decomposition of $(h, g, Y)$.

In this framework, the mediator $M$ is considered a function of $D$, denoted by $M(D)$, and the output $Y$ is set as a function of $D$ and $M$, as denoted by $Y(D, M)$, so that $M$ would potentially vary with $D$, and $Y$ would potentially change with $D$ and $M$, respectively. Given the event $D = d$, $M(d)$ would be the same as the actual $M$, and $M(1-d)$ is a counterfactual $M$ with regard to what would happen to $M$ if the outcome of $D$ was $1-d$; $Y(d, M(d))$ is the same as the actual $Y$, and $Y(d, M(1-d))$, $Y(1-d, M(d))$ and $Y(1-d, M(1-d))$ are respectively counterfactual $Y$ regarding what would happen to $Y$ when $(D, M) = (d, M(1-d))$, $(1-d, M(d))$ or $(1-d, M(1-d))$. By construction, $Y$ and $Y(d, M(d))$ have the relationship:

$$Y = D \cdot Y(1, M(1)) + (1 - D) \cdot Y(0, M(0)); \tag{2.6}$$

similarly, $M$ and $M(d)$ have the relationship:

$$M = D \cdot M(1) + (1 - D) \cdot M(0). \tag{2.7}$$

We let $Z_1$ be an exogenous binary IV with outcome $z_1 \in \{0, 1\}$ to implement program evaluation with the endogenous $D$. $D(z_1)$ is a potential treatment status for $z_1 \in \{0, 1\}$. Given the event $Z_1 = z_1$, $D(z_1)$ would be the same as the actual $D$, and $D(1-z_1)$ is a counterfactual $D$. It is customary to think of $Z_1$ as a variable that indicates whether an "exogenous incentive" to obtain treatment is present. Similar to (2.6) and (2.7), $D$ and $D(z_1)$ have the relationship:

$$D = Z_1 \cdot D(1) + (1 - Z_1) \cdot D(0).$$

As in Imbens and Angrist (1994), we use the four outcomes of $(D(1), D(0))$ to divide the whole population into four mutually disjoint subpopulations:

$$(D(1), D(0)) = \begin{cases} (1, 1), & \text{always takers,} \\ (1, 0), & \text{compliers (hereafter, denoted by } \mathcal{C}), \\ (0, 1), & \text{defiers,} \\ (0, 0), & \text{never takers.} \end{cases} \tag{2.8}$$

It is known that, by focusing on the subpopulation of compliers, we can identify the conditional local average treatment effect (CLATE):

$$CLATE(x) \equiv E[Y(1, M(1))|X = x, \mathcal{C}] - E[Y(0, M(0))|X = x, \mathcal{C}] \tag{2.9}$$

and the local average treatment effect (LATE):

$$LATE = E[CLATE(X)|\mathcal{C}] = E[Y(1, M(1))|\mathcal{C}] - E[Y(0, M(0))|\mathcal{C}]$$

when $D$ is endogenous; see, e.g., Abadie (2003, Theorem 3.1).

We decompose the CLATE into a conditional direct LATE (CDLATE) and a conditional indirect LATE (CILATE) for mediation analysis. This not only needs $E[Y(0,M(0))|X,\mathcal{C}]$ and $E[Y(1,M(1))|X,\mathcal{C}]$ but also two additional conditional means $E[Y(0,M(1))|X,\mathcal{C}]$ and $E[Y(1,M(0))|X,\mathcal{C}]$ to facilitate the decomposition of (2.9):

$$CLATE(x) = CDLATE(x) + CILATE(x), \tag{2.10}$$

where

$$CDLATE(x) \equiv E[Y(1,M(1))|X=x,\mathcal{C}] - E[Y(0,M(1))|X=x,\mathcal{C}] \tag{2.11}$$

and

$$CILATE(x) \equiv E[Y(0,M(1))|X=x,\mathcal{C}] - E[Y(0,M(0))|X=x,\mathcal{C}]; \tag{2.12}$$

see Frölich and Huber (2014) for such a decomposition of LATE. The CDLATE of $D$ on $Y$ is defined by holding the mediator $M$ at $M(1)$, and the CILATE of $D$ on $Y$ is generated through changing $M$ from $M(0)$ to $M(1)$ by holding the treatment status $D$ at $D=0$. It is known that the decomposition is path-dependent; see, e.g., Fortin, Lemieux, and Firpo (2011). One may also define CDLATE by fixing $M$ at $M(0)$ instead and define CILATE by holding the treatment status at $D=1$.

This modeling framework is similar to that of Frölich and Huber (2014), which has a non-parametric structure and a single mediator that can be either continuous or binary. The main difference is that we adopt a parametric structure to accommodate the fact that stochastic frontier analysis typically relies on parametric specification to separate the frontier function from the inefficiency component. Meanwhile, we focus on a binary mediator for simplicity. The use of a continuous mediator or multiple mediators would considerably complicate our model and estimation method, so we leave it for a future study. Other mediation analysis studies in economics include

Flores and Flores-Lagunes (2009) and Huber (2014), who considered the exogenous treatment status. Huber et al. (2016) conducted comprehensive empirical Monte-Carlo simulations to investigate the finite sample properties of different classes of estimators of direct and indirect causal effects. Flores and Flores-Lagunes (2010) provided non-parametric bounds for direct and indirect average treatment effects without using any instrument variables. Flores and Flores-Lagunes (2010) and Bampasidou et al. (2016) further applied the bound analysis to evaluate the impact of a job training program. We can refer to the references in MacKinnon (2008), Frölich and Huber (2014), and Huber (2014) for mediation analysis in other fields.

## 2.2 A New Stochastic Frontier Model

For program evaluation and mediation analysis in the stochastic frontier context, we propose a new stochastic frontier model for potential outputs: For $d, d' \in \{0, 1\}$,

$$Y(d, M(d')) = \tilde{h}(d, M(d'), X) + v(d, M(d')) - u(d, M(d'), X), \tag{2.13}$$

where $\tilde{h}(d, M(d'), X)$ is a potential frontier function with outcomes:

$$\tilde{h}(d, M(d'), X) = \begin{cases} h(X, \beta^h_{11}), & (d, M(d')) = (1, 1), \\ h(X, \beta^h_{10}), & (d, M(d')) = (1, 0), \\ h(X, \beta^h_{01}), & (d, M(d')) = (0, 1), \\ h(X, \beta^h_{00}), & (d, M(d')) = (0, 0), \end{cases} \tag{2.14}$$

$v(d, M(d'))$ is a pure potential random error with $E[v(d, M(d'))|X, \mathcal{C}] = 0$, and $u(d, M(d'), X)$ is a non-negative potential production inefficiency such that

$$u(d, M(d'), X) = \tilde{g}(d, M(d'), X) + \tilde{u}(d, M(d')).$$

Here, $E[\tilde{u}(d, M(d'))|X, \mathcal{C}] = 0$, and $\tilde{g}(d, M(d'), X) = E[u(d, M(d'), X)|X, \mathcal{C}]$ is a non-negative potential inefficiency function with outcomes:

$$\tilde{g}(d, M(d'), X) = \begin{cases} g(X, \beta^g_{11}), & (d, M(d')) = (1, 1), \\ g(X, \beta^g_{10}), & (d, M(d')) = (1, 0), \\ g(X, \beta^g_{01}), & (d, M(d')) = (0, 1), \\ g(X, \beta^g_{00}), & (d, M(d')) = (0, 0). \end{cases} \tag{2.15}$$

This model has the parameter vector: $\beta_d \equiv (\beta^{h\top}_{d1}, \beta^{h\top}_{d0}, \beta^{g\top}_{d1}, \beta^{g\top}_{d0})^\top$ for $d \in \{0, 1\}$. Theoretically $h(X, \beta^h_{dj})$ may take different functional forms for different $(d, M(d'))$ combinations, where $j = M(d')$. Here, we assume that the functional forms are the same for simplicity. Similarly, the form of $g(X, \beta^g_{dj})$ would be related to the distributional assumption of $u$ and we assume that it has the same functional form for different combinations of $(d, M(d'))$.

Deriving the conditional mean $E[Y(d, M(d'))|X, \mathcal{C}]$ in this new model requires that we use the law of total probability to show that for $d \in \{0, 1\}$,

$$\begin{aligned} E[\tilde{h}(d, M(d'), X)|X, \mathcal{C}] &= P(M(d') = 1|X, \mathcal{C})\tilde{h}(d, 1, X) + P(M(d') = 0|X, \mathcal{C})\tilde{h}(d, 0, X) \\ &= E[M(d')|X, \mathcal{C}]h(X, \beta^h_{d1}) + (1 - E[M(d')|X, \mathcal{C}])\, h(X, \beta^h_{d0}) \end{aligned}$$

and

$$\begin{aligned} E[\tilde{g}(d, M(d'), X)|X, \mathcal{C}] &= P(M(d') = 1|X, \mathcal{C})\tilde{g}(d, 1, X) + P(M(d') = 0|X, \mathcal{C})\tilde{g}(d, 0, X) \\ &= E[M(d')|X, \mathcal{C}]g(X, \beta^g_{d1}) + (1 - E[M(d')|X, \mathcal{C}])\, g(X, \beta^g_{d0}). \end{aligned}$$

As will be shown in (3.10), we can obtain the following potential-mediator model under suitable assumptions: for $d' \in \{0, 1\}$,

$$E[M(d')|X, \mathcal{C}] = m_{d'}(X, \alpha_m), \tag{2.16}$$

where $m_{d'}(\cdot)$ is a non-negative function in $(0, 1)$ with a parameter vector $\alpha_m$. Combining the potential-output model with this potential-mediator model allows us to obtain the following conditional mean representation for (2.13): for $d, d' \in \{0, 1\}$,

$$E[Y(d, M(d'))|X, \mathcal{C}] = h_{d'}(X, \alpha_m, \beta_{d1}^h, \beta_{d0}^h) - g_{d'}(X, \alpha_m, \beta_{d1}^g, \beta_{d0}^g), \tag{2.17}$$

where

$$h_{d'}(X, \alpha_m, \beta_{d1}^h, \beta_{d0}^h) \equiv m_{d'}(X, \alpha_m) h(X, \beta_{d1}^h) + (1 - m_{d'}(X, \alpha_m))\, h(X, \beta_{d0}^h) \tag{2.18}$$

and

$$g_{d'}(X, \alpha_m, \beta_{d1}^g, \beta_{d0}^g) \equiv m_{d'}(X, \alpha_m) g(X, \beta_{d1}^g) + (1 - m_{d'}(X, \alpha_m))\, g(X, \beta_{d0}^g). \tag{2.19}$$

This representation is of the parameter vector: $(\alpha_m^\top, \beta_d^\top)^\top$ for $d \in \{0, 1\}$.

Upon comparing (2.17) with (2.5), we note that our model is substantially different from traditional stochastic frontier models in two important aspects. First, our model comprises a system of stochastic frontier models for latent variables: $Y(0, M(0))$, $Y(0, M(1))$, $Y(1, M(0))$ and $Y(1, M(1))$ and a pair of models for $M(0)$ and $M(1)$, and its conditioning set is based on the subpopulation of compliers and thus is also latent. Therefore, it needs a new parameter identification method. Second, our model allows the parameters of the frontier function $h(\cdot)$ and the inefficiency function $g(\cdot)$ to vary with $(d, M(d'))$.

Compared to traditional treatment effect models, our model can be applied to further decompose CLATE in (2.9), and CDLATE and CILATE in (2.10) into their production-frontier components and production-inefficiency components, respectively. Note that this extension is facilitated when using the parametric specification in (2.17).

## 2.3 Conditional Local Average Treatment Effects

Suppose that for $d, d' \in \{0, 1\}$, (2.17) is correctly specified for $E[Y(d, M(d'))|X, \mathcal{C}]$, and the parameter vector $(\alpha_m^\top, \beta_d^\top)^\top$ is identifiable. Let $x$ denote a value of $X$ that represents a particular type of individual's observable features.

For program evaluation in the stochastic frontier context, according to (2.17), (2.18) and (2.19), we decompose (2.9) into:

$$CLATE(x) = CLATE_h(x) - CLATE_g(x), \tag{2.20}$$

with the CLATE for the production frontier:

$$\begin{aligned} CLATE_h(x) &\equiv h_1(x, \alpha_m, \beta^h_{11}, \beta^h_{10}) - h_0(x, \alpha_m, \beta^h_{01}, \beta^h_{00}) \\ &= m_1(x, \alpha_m) h(x, \beta^h_{11}) + (1 - m_1(x, \alpha_m))\, h(x, \beta^h_{10}) \\ &- \big(m_0(x, \alpha_m) h(x, \beta^h_{01}) + (1 - m_0(x, \alpha_m))\, h(x, \beta^h_{00})\big) \end{aligned}$$

and the CLATE for the production inefficiency:

$$\begin{aligned} CLATE_g(x) &\equiv g_1(x, \alpha_m, \beta^g_{11}, \beta^g_{10}) - g_0(x, \alpha_m, \beta^g_{01}, \beta^g_{00}) \\ &= m_1(x, \alpha_m) g(x, \beta^g_{11}) + (1 - m_1(x, \alpha_m))\, g(x, \beta^g_{10}) \\ &- (m_0(x, \alpha_m) g(x, \beta^g_{01}) + (1 - m_0(x, \alpha_m))\, g(x, \beta^g_{00}))\,. \end{aligned}$$

For mediation analysis, we further decompose (2.11) and (2.12) into:

$$CDLATE(x) = CDLATE_h(x) - CDLATE_g(x) \tag{2.21}$$

and

$$CILATE(x) = CILATE_h(x) - CILATE_g(x), \tag{2.22}$$

respectively, with the CDLATE for the production frontier:

$$\begin{aligned} &CDLATE_h(x) \equiv h_1(x, \alpha_m, \beta^h_{11}, \beta^h_{10}) - h_1(x, \alpha_m, \beta^h_{01}, \beta^h_{00}) \\ &= m_1(x, \alpha_m) \big(h(x, \beta^h_{11}) - h(x, \beta^h_{01})\big) + (1 - m_1(x, \alpha_m)) \big(h(x, \beta^h_{10}) - h(x, \beta^h_{00})\big)\,, \end{aligned}$$

the CDLATE for the production inefficiency:

$$\begin{aligned} &CDLATE_g(x) \equiv g_1(x, \alpha_m, \beta^g_{11}, \beta^g_{10}) - g_1(x, \alpha_m, \beta^g_{01}, \beta^g_{00}) \\ &= m_1(x, \alpha_m) (g(x, \beta^g_{11}) - g(x, \beta^g_{01})) + (1 - m_1(x, \alpha_m)) (g(x, \beta^g_{10}) - g(x, \beta^g_{00}))\,, \end{aligned}$$

the CILATE for the production frontier:

$$CILATE_h(x) \equiv h_1(x, \alpha_m, \beta_{01}^h, \beta_{00}^h) - h_0(x, \alpha_m, \beta_{01}^h, \beta_{00}^h)$$
$$= (m_1(x, \alpha_m) - m_0(x, \alpha_m)) \left(h(x, \beta_{01}^h) - h(x, \beta_{00}^h)\right),$$

and the CILATE on the production inefficiency:

$$CILATE_g(x) \equiv g_1(x, \alpha_m, \beta_{01}^g, \beta_{00}^g) - g_0(x, \alpha_m, \beta_{01}^g, \beta_{00}^g)$$
$$= (m_1(x, \alpha_m) - m_0(x, \alpha_m)) \left(g(x, \beta_{01}^g) - g(x, \beta_{00}^g)\right).$$

Studying these conditional effects is useful for exploring whether and how treatment influences the outputs of the individuals in $\mathcal{C}$ with features characterized by $X = x$.

We can establish a set of parameter restrictions corresponding to these decompositions for program evaluation:

$$H_o^{\mathrm{p}} \ : \ \beta_{11}^h = \beta_{10}^h = \beta_{01}^h = \beta_{00}^h \text{ and } \beta_{11}^g = \beta_{10}^g = \beta_{01}^g = \beta_{00}^g \text{ for no } CLATE(x), \forall x;$$
$$H_o^{\mathrm{p}h} \ : \ \beta_{11}^h = \beta_{10}^h = \beta_{01}^h = \beta_{00}^h \text{ for no } CLATE_h(x), \forall x;$$
$$H_o^{\mathrm{p}g} : \ \beta_{11}^g = \beta_{10}^g = \beta_{01}^g = \beta_{00}^g \text{ for no } CLATE_g(x), \forall x.$$

In addition, as will be seen in (3.8) and (3.10), the difference between the functions $m_1(\cdot)$ and $m_0(\cdot)$ is fully determined by parameter $\alpha_d$, which represents the partial effect of $D$ on $M$ such that $m_1(\cdot) = m_0(\cdot)$ if $\alpha_d = 0$; otherwise, $m_1(\cdot) \neq m_0(\cdot)$. Correspondingly, we can also establish another set of parameter restrictions for mediation analysis:

$$\begin{aligned}
H_o^{\mathrm{d}} &: (\beta_{11}^h, \beta_{10}^h, \beta_{11}^g, \beta_{10}^g) = (\beta_{01}^h, \beta_{00}^h, \beta_{01}^g, \beta_{00}^g) \text{ for no } CDLATE(x), \forall x;\\
H_o^{\mathrm{d}h} &: (\beta_{11}^h, \beta_{10}^h) = (\beta_{01}^h, \beta_{00}^h) \text{ for no } CDLATE_h(x), \forall x;\\
H_o^{\mathrm{d}g} &: (\beta_{11}^g, \beta_{10}^g) = (\beta_{01}^g, \beta_{00}^g) \text{ for no } CDLATE_g(x), \forall x;\\
H_o^{\mathrm{i}} &: (\beta_{01}^h, \beta_{01}^g) = (\beta_{00}^h, \beta_{00}^g) \text{ for no } CILATE(x), \forall x;\\
H_o^{\mathrm{i}h} &: \beta_{01}^h = \beta_{00}^h \text{ or } \alpha_d = 0 \text{ for no } CILATE_h(x), \forall x;\\
H_o^{\mathrm{i}g} &: \beta_{01}^g = \beta_{00}^g \text{ or } \alpha_d = 0 \text{ for no } CILATE_g(x), \forall x.
\end{aligned}$$

Note that the hypotheses: $H_o^{\mathrm{i}h}$ and $H_o^{\mathrm{i}g}$ can be simplified as the parameter restrictions: $\beta_{01}^h = \beta_{00}^h$ and $\beta_{01}^g = \beta_{00}^g$, respectively, when $\alpha_d \neq 0$.

These hypotheses are all conditional on the subpopulation of compliers. Among them, $H_o^{\mathrm{p}}$ means that the potential frontier function $\tilde{h}(d, M(d'), X)$ and the potential inefficiency function $\tilde{g}(d, M(d'), X)$ are both invariant to the treatment status outcome $d$ and the potential mediator $M(d')$, so that there is no CLATE for all possible $x$. In comparison, $H_o^{\mathrm{d}}$ ($H_o^{\mathrm{i}}$) means that, fixing the outcome of potential mediator (treatment status), these two potential functions are both invariant to the outcome of treatment status (potential mediator), so there is no CDLATE (CILATE) for all possible $x$. Furthermore, $H_o^{\mathrm{p}h}$ and $H_o^{\mathrm{p}g}$ are weaker than $H_o^{\mathrm{p}}$, and are respectively focused on the restriction of $\tilde{h}(d, M(d'), X)$ and of $\tilde{g}(d, M(d'), X)$. Similarly, $H_o^{\mathrm{d}h}$ and $H_o^{\mathrm{d}g}$ ($H_o^{\mathrm{i}h}$ and $H_o^{\mathrm{i}g}$) are weaker than $H_o^{\mathrm{d}}$ ($H_o^{\mathrm{i}}$), and are respectively focused on the restriction of $\tilde{h}(d, M(d'), X)$ and that of $\tilde{g}(d, M(d'), X)$. In empirical applications, researchers may assess the significance of the effectiveness of $D$ on $Y$ and disentangle the underlying effects of the influences of $D$ on $Y$ for the subpopulation of compliers conditional on $X = x$ by testing these hypotheses and comparing results.

## 2.4 Local Average Treatment Effects

In addition, we may extend (2.10) to a decomposition of LATE:

$$LATE = DLATE + ILATE, \tag{2.23}$$

with the direct LATE (DLATE):

$$DLATE = E[CDLATE(X)|\mathcal{C}] = E[Y(1, M(1))|\mathcal{C}] - E[Y(0, M(1))|\mathcal{C}]$$

and the indirect LATE (ILATE):

$$ILATE = E[CILATE(X)|\mathcal{C}] = E[Y(0, M(1))|\mathcal{C}] - E[Y(0, M(0))|\mathcal{C}],$$

and use (2.17) to further show the decompositions:

$$LATE = LATE_h - LATE_g, \tag{2.24}$$

$$DLATE = DLATE_h - DLATE_g, \tag{2.25}$$

$$ILATE = ILATE_h - ILATE_g, \tag{2.26}$$

where

$$LATE_h \equiv E[CLATE_h(X)|\mathcal{C}], \quad LATE_g \equiv E[CLATE_g(X)|\mathcal{C}],$$

$$DLATE_h \equiv E[CDLATE_h(X)|\mathcal{C}], \quad DLATE_g \equiv E[CDLATE_g(X)|\mathcal{C}],$$

$$ILATE_h \equiv E[CILATE_h(X)|\mathcal{C}], \quad ILATE_g \equiv E[CILATE_g(X)|\mathcal{C}].$$

Analyzing these effects is useful for implementing the program evaluation and mediation analysis without focusing on a particular type of individual feature.

Similar to the hypotheses: $H_o^{\mathrm{p}}$, $H_o^{\mathrm{p}h}$, $H_o^{\mathrm{p}g}$, $H_o^{\mathrm{d}}$, $H_o^{\mathrm{d}h}$, $H_o^{\mathrm{d}g}$, $H_o^{\mathrm{i}}$, $H_o^{\mathrm{i}h}$ and $H_o^{\mathrm{i}g}$, we also define a set of moment restrictions for program evaluation:

$$\begin{aligned}
&\tilde{H}_o^{\mathrm{p}} \;:\; LATE = 0,\\
&\tilde{H}_o^{\mathrm{p}h} \;:\; LATE_h = 0,\\
&\tilde{H}_o^{\mathrm{p}g} :\; LATE_g = 0,
\end{aligned}$$

and another set of moment restrictions for mediation analysis:

$$\begin{aligned}
&\tilde{H}_o^{\mathrm{d}} \;: DLATE = 0,\\
&\tilde{H}_o^{\mathrm{d}h} :\; DLATE_h = 0,\\
&\tilde{H}_o^{\mathrm{d}g} :\; DLATE_g = 0,\\
&\tilde{H}_o^{\mathrm{i}} \;: ILATE = 0,\\
&\tilde{H}_o^{\mathrm{i}h} :\; ILATE_h = 0,\\
&\tilde{H}_o^{\mathrm{i}g} :\; ILATE_g = 0.
\end{aligned}$$

These hypotheses are respectively of similar interpretations to $H_o^{\mathrm{p}}$, $H_o^{\mathrm{p}h}$, $H_o^{\mathrm{p}g}$, $H_o^{\mathrm{d}}$, $H_o^{\mathrm{d}h}$, $H_o^{\mathrm{d}g}$, $H_o^{\mathrm{i}}$, $H_o^{\mathrm{i}h}$ and $H_o^{\mathrm{i}g}$, while they are not conditional on $X = x$. In applications, researchers can evaluate the effectiveness of $D$ on $Y$ and disentangle its underlying effects for the subpopulation of compliers by testing these hypotheses and comparing the testing results.

We summarized the CLATE and LATE components in Table 1. This table clearly shows that the proposed method connects the program evaluation (the mediation analysis) to the stochastic frontier analysis by further decomposing the CLATE and LATE (the associated direct and indirect effects) on the output into their frontier and inefficiency components.

Table 1: CLATE and LATE Components

| | stochastic frontier analysis: CLATE | | | stochastic frontier analysis: LATE | | |
|---|---|---|---|---|---|---|
| | output | frontier | inefficiency | output | frontier | inefficiency |
| program evaluation | $CLATE(x)$ | $CLATE_h(x)$ | $CLATE_g(x)$ | $LATE$ | $LATE_h$ | $LATE_g$ |
| mediation | $DCLATE(x)$ | $DCLATE_h(x)$ | $CLATE_g(x)$ | $DLATE$ | $DLATE_h$ | $DLATE_g$ |
| analysis | $ICLATE(x)$ | $ICLATE_h(x)$ | $CLATE_g(x)$ | $ILATE$ | $ILATE_h$ | $ILATE_g$ |

## 2.5 The Dam Example: Variables and Effects

Here, we link the proposed model to the dam example, which we introduced in the Introduction, by giving the variables, the endogeneity issues, and the effects in our empirical illustration. Details of the empirical dataset will be discussed in Section 5.1. For ease of exposition, we summarize the empirical definitions of the observable variables that will be discussed below in Table 2, and provide empirical meanings of the potential variables: $Y(d, M(d'))$, $M(d')$ and $D(z_1)$ in the same table. However, we leave the discussion of the empirical validity of IVs in Section 3.3. In addition, we wait until Section 5.1 to demonstrate the empirical dataset, the construction of the treatment and evaluation period, and the details of $X$.

In this example, we let $Y$ be the log of the agricultural production of an Indian district during the post-treatment evaluation period, $D$ be a binary indicator of the dam construction in the same district during the treatment period, and $M$ be a binary indicator of irrigation intensity, which is measured from the ratio of the district's irrigated area to its total cultivated area in the evaluation period. We consider this a mediator because enabling irrigation for agricultural production is an

Table 2: Empirical Definitions of Variables in the Dam Example

| Variable | definition | empirical definition in the dam example |
|---|---|---|
| $Y$ | observable output | the log of agricultural production of an Indian district during the evaluation period. |
| $D$ | binary treatment:<br>$D = 1$, treated<br>$D = 0$, untreated | dam-construction indicator:<br>$D = 1$, one or more dams were built in the district during the treatment period;<br>$D = 0$, otherwise. |
| $M$ | binary mediator:<br>$M = 1$, high intensity<br>$M = 0$, low intensity | irrigation-intensity indicator:<br>$M = 1$, the average ratio of irrigated land in the district is larger than the median of the sample during the evaluation period;<br>$M = 0$, otherwise. |
| $Z_1$ | binary IV for $D$:<br>$Z_1 = 1$, incentive to treat<br>$Z_0 = 0$, otherwise | river gradient (for the geographical suitability of dam construction):<br>$Z_1 = 1$, the fraction of the district through which “gentle-gradient” or “steep-gradient” rivers flow is above the median of the sample;<br>$Z_1 = 0$, otherwise. |
| $Z_2$ | continuous IV for $M$ | river length (for the geographical suitability of irrigation):<br>the OLS residual obtained by regressing the log of the total river length on $Z_1$ and $X$ for each district. |
| $X$ | vector of observed covariates | environmental variables: rainfall, elevation, geographic variables<br>pre-treatment agricultural inputs: existing number of dams, fertilizer, cultivated area before the treatment period. See Section 5.1 for details. |
| $Y(d, M(d'))$ | potential output when $D = d$ and $M = M(d')$, where $d, d' = 0, 1$ | $Y(0, M(0))$, $Y(0, M(1))$, $Y(1, M(0))$ and $Y(1, M(1))$ are potential agricultural productions during the evaluation period; e.g., $Y(1, M(0))$ means the agricultural production with the dam construction and the counterfactual irrigation intensity without the dam construction. |
| $M(d')$ | potential mediator when $D = d'$, where $d' = 0, 1$ | $M(0)$ is the potential irrigation-intensity indicator during the evaluation period without the dam construction during the treatment period; $M(1)$ is the counterpart of $M(0)$ with the dam construction. |
| $D(z_1)$ | potential treatment when $Z_1 = z_1$, where $z_1 = 0, 1$ | $D(1)$ is the potential dam-constriction indicator during the treatment period when the district's river gradient is suitable for dam construction; $D(0)$ is the counterpart of $D(1)$ without this incentive to treat. |

Note: In Section 5, we demonstrate that the treatment period is from 1976 to 1980, and the evaluation period is from 1981 to 1987.

important potential benefit of building dams. In addition, we set $X$ to be a vector of exogenous environmental variables and pre-treatment agricultural inputs.

As was explained in Duflo and Pande (2007a), the Indian Planning Commission sets water storage and irrigation targets for each state government, and the latter proposes dam projects to the federal government based on the targets and the geographical surveys of potential dam sites. With this mechanism, richer and fast growing states can build more dams, and states anticipating larger increases in agricultural production may also build more dams. The correlation between agricultural production (which is the main economic activities to the majority of states) and dam construction gives rise to the endogeneity of $D$. The mechanism also indicates that geographical properties are important in deciding whether to build dams. One of the most important properties is the river gradient. From engineering point of view, the likelihood of dam construction increases if the river gradient is either gentle (1.5-3%) or very steep (more than 6%, mostly for hydropower dams). Following Duflo and Pande (2007a), we explore the connection between dam construction and the geographical property of the local district to propose the IV for $D$. We let $Z_1$ to be a binary IV that is constructed based on the fraction of the district through which either gentle-gradient or steep-gradient rivers (defined above) flow.

The mediator $M$ may also be endogenous in the dam example because the irrigation construction could also be confounded by the unexplained conditions of agricultural production. As will be explained later, the endogeneity of $M$ requires an additional continuous IV, which is denoted by $Z_2$, for parameter identification. All other things being equal, districts with larger river distributary systems, such as river basins (and hence longer total river lengths), are advantageous in building efficient irrigation systems. We set $Z_2$ to be the ordinary least square (OLS) residual obtained by regressing the log of the district's river length on $Z_1$ and $X$ to account for this consideration and to satisfy certain identification assumptions (see Section 3.3).

Given the empirical definitions of $D$ and $Z_1$, a complier in this example is an India district which build dams because it has favorable river gradients and does not if otherwise. Correspondingly, $CLATE(x)$ represents the total effect of dam construction on agricultural production for a complier with $X = x$; $CILATE(x)$ is the associated indirect effect generated via irrigation, $CDLATE(x)$ is the remaining direct effect, and $CILATE_h(x)$ and $CILATE_g(x)$ ($CDLATE_h(x)$ and $CDLATE_g(x)$) are respectively the indirect (direct) effects on the production frontier and inefficiency. In addition, $LATE$ represents the total effect of dam construction on the agricultural production for a complier; $ILATE$ is the associated indirect effect generated via irrigation, $DLATE(x)$ is the remaining direct effect, and $ILATE_h$ and $ILATE_g$ ($DLATE_h$ and $DLATE_g$) are respectively the indirect (direct) effects on the production frontier and inefficiency.

# 3 Identification

We discuss the identification of the CLATE and LATE components in this section. Recall that these components are implied by the model in (2.17) that depends on the production function $h(\cdot)$, the inefficiency function $g(\cdot)$ and the potential-mediator function $m_{d'}(\cdot)$; meanwhile, $h(\cdot)$ and $g(\cdot)$ include the parameter vector $\beta_d$, for $d \in \{0, 1\}$, and $m_{d'}(\cdot)$ includes the parameter vector $\alpha_m$. Following the stochastic frontier literature, parameters of the model are identified based on the the model's distribution assumptions. In particularly, the parametric function of $g(\cdot)$ follows from the distribution assumption of $u$, and the nonlinearity of the function helps to identify parameters in $g(\cdot)$ and $h(\cdot)$. The following discussions focuses on the identification of $\beta_d$ for $d \in \{0, 1\}$, $m_{d'}(\cdot, \alpha_m)$, and the CLATE and LATE components.

To see our identification problem and strategy, note that (2.17) is a conditional mean model of the potential output $Y(d, M(d'))$ on $X$ for the subpopulation of compliers. Accordingly, given $m_{d'}(X, \alpha_m)$, we may first identify $\beta_d$, for $d \in \{0, 1\}$, as the minimizer of the MSE of this model conditional on the subpopulation of compliers. However, this objective function is not operational because it involves $Y(d, M(d'))$, $m_{d'}(x, \alpha_m)$ and the subpopulation of compliers that are all latent in practice. In the context of program evaluation with an endogenous $D$, but without a mediator, it is known that the conditional mean of a potential outcome variable on the subpopulation of compliers can be transformed to a weighted unconditional mean of an observed outcome for identification by assuming an exclusion restriction for $Z_1$, the monotonicity of $D$ in $Z_1$, and a common support condition; see, e.g., Abadie (2003). This identification strategy needs a suitable extension in our context.

Intuitively, when transforming the conditional MSE into an operational form, we need to account for the causal pathway shown in Figure 1 and the endogeneity of both $D$ and $M$. We need to establish a structural model of $(Y, M, D)$ for characterizing the potential outcomes involved in (2.17), and we need to use another IV, denoted $Z_2$, to generate exogenous variations of the endogenous $M$. In addition, we need to extend the exclusion restriction and the associated auxiliary conditions from the single-IV context to a two-IV context. In this study, we facilitate this extension by using the method of Frölich and Huber (2014), which uses a continuous IV for an endogenous binary mediator. We also set $Z_2$ to be a continuous IV for the endogenous $M$ to apply their method.

In the following, we show that this approach allows us to identify the parameter vector: $\beta_d$, for $d \in \{0, 1\}$, the potential mediator function $m_{d'}(X, \alpha_m)$ and the CLATE and LATE components from the distribution of $\boldsymbol{W} \equiv (Y, M, D, Z_1, Z_2, X^\top)^\top$.

## 3.1 Identification of Parameters

Let $\mathcal{B}_d$ be the parameter space of $\beta_d$ for $d \in \{0,1\}$. Corresponding to $\beta_d = (\beta_{d1}^{h\top}, \beta_{d0}^{h\top}, \beta_{d1}^{g\top}, \beta_{d0}^{g\top})^\top$, we let $b_d \equiv (b_{d1}^{h\top}, b_{d0}^{h\top}, b_{d1}^{g\top}, b_{d0}^{g\top})^\top$ be an arbitrary vector in $\mathcal{B}_d$. We can observe that $\beta_d$ is separable from $\alpha_m$ from (2.17), (2.18), and (2.19). Thus, we let the identification of $\beta_d$ be conditional on that of $\alpha_m$, and make the following assumption:

**Assumption 3.1** *Assume that*

*(a)* $\alpha_m$ *is identifiable,*

*(b) (2.17) holds for some* $\beta_d \in \mathcal{B}_d$ *and for* $d \in \{0,1\}$*, and*

*(c)* $E\Big[\big(h_{d'}(X,\alpha_m,\beta_{d1}^h,\beta_{d0}^h)-g_{d'}(X,\alpha_m,\beta_{d1}^g,\beta_{d0}^g)-\big(h_{d'}(X,\alpha_m,b_{d1}^h,b_{d0}^h)-g_{d'}(X,\alpha_m,b_{d1}^g,b_{d0}^g)\big)\big)^2|\mathcal{C}\Big]$ *is strictly positive, for* $d, d' \in \{0,1\}$ *and for all* $b_d \in \mathcal{B}_d$ *such that* $b_d \neq \beta_d$*.*

This assumption is standard for the parameter identification of a non-linear regression. Assumption 3.1($a$) will be shown to be valid following Theorem 3.1. Assumption 3.1($b$) requires (2.17) to be correctly specified for $E[Y(d,M(d'))|X,\mathcal{C}]$, and Assumption 3.1($c$) means that, given $\alpha_m$, the parameter vector $\beta_d$ is unique.

This assumption implies that, for $d \in \{0,1\}$,

$$\beta_d \equiv \arg\min_{b_d \in \mathcal{B}_d} \sum_{d'=0,1} E\Big[\Big(Y(d,M(d')) - h_{d'}(X,\alpha_m,b_{d1}^h,b_{d0}^h) + g_{d'}(X,\alpha_m,b_{d1}^g,b_{d0}^g)\Big)^2|\mathcal{C}\Big]. \tag{3.1}$$

Note that $E[Y(d,M(1))|X,\mathcal{C}]$ and $E[Y(d,M(0))|X,\mathcal{C}]$ share the same parameters, so the objective function is the sum of two sub-problems. Given $\alpha_m$, we could identify $\beta_d$ based on (3.1), if the potential outputs $Y(d,M(d'))$, the potential mediators $M(d')$ underlying $Y(d,M(d'))$ and the $m_{d'}(X,\alpha_m)$ in (2.18) and (2.19), and the potential treatment status $D(z_1)$ underlying the definition of $\mathcal{C}$ were observed. However, we can only observe $\boldsymbol{W}$ in practice rather than these latent variables. Thus, (3.1) is indeed insufficient for the identification of $\beta_d$. We circumvent this problem in the following by setting a parametric structural model for $(Y, M, D)$ and applying the identification strategy of Frölich and Huber (2014) to our parametric context.

Let $1(\cdot)$ be the indicator function which is equal to one when the associated event is true (otherwise, zero). We make the following assumptions:

**Assumption 3.2** *(***Parametric structural model***) Assume that:*

*(a) A parametric structural model for $(Y, M, D)$:*

$$\begin{aligned} Y &= \tilde{h}(D, M, X) - \tilde{g}(D, M, X) + U_Y, \\ M &= 1(\alpha_d D + \alpha_{z_2} Z_2 + X^\top \alpha_x + U_M \geq 0), \\ D &= 1(\gamma_{z1} Z_1 + X^\top \gamma_x + U_D \geq 0), \end{aligned}$$

*where $\tilde{h}(D, M, X)$ and $\tilde{g}(D, M, X)$ follow (2.14) and (2.15); $U_Y$, $U_M$ and $U_D$ are error terms; $(\alpha_d, \alpha_{z_2}, \alpha_x^\top)^\top$ and $(\gamma_{z1}, \gamma_x^\top)^\top$ are parameter vectors.*

*(b) $U_Y \equiv v(D, M) - \tilde{u}(D, M)$ and $E[U_Y | X, \mathcal{C}] = 0$.*

**Assumption 3.3** *(***Identification of (3.1)***) Assume that*

*(a) (Exclusion restrictions of $Z_1$ and $Z_2$)*

$$\begin{aligned} &(Z_1, Z_2) \perp (U_Y, U_M) | U_D, X, \\ &Z_1 \perp (U_Y, U_M, U_D) | Z_2, X, \end{aligned}$$

*where $\perp$ denotes statistical independence.*

*(b) (Conditional independence of $Z_1$ and $Z_2$) $Z_1 \perp Z_2 | X$.*

*(c) (Weak monotonicity of $D$ in $Z_1$)* $\gamma_{z1} > 0$.

*(d) (Weak monotonicity of $M$ in $Z_2$)*

*(d1)* $\alpha_{z_2} > 0$,

*(d2) $U_M$ is continuously distributed with the cumulative distribution function (CDF) $F_{U_M}(\cdot)$ that is strictly increasing in the support of $U_M$.*

*(e) (Common support of $M$)*

$$0 < P(Z_1 = 1|M, U_M, X, \mathcal{C}) < 1 \quad \textit{almost surely.}$$

Assumption 3.2 constitutes a parametric structural model for $(Y, M, D)$ that accommodates the causal pathway in Figure 1. This structural model encompasses a sub-model for $Y|D, M, \mathcal{C}, X$ that reduces to the potential-output model for $Y(d, M(d'))$ in (2.13) when $(D, M) = (d, M(d'))$, and comprises a potential-mediator model:

$$M(d') = 1(\alpha_d d' + \alpha_{z_2} Z_2 + X^\top \alpha_x + U_M \geq 0) \tag{3.2}$$

and a potential-treatment model:

$$D(z_1) = 1(\gamma_{z1} z_1 + X^\top \gamma_x + U_D \geq 0), \tag{3.3}$$

for $d, d', z_1 \in \{0, 1\}$. Assumption 3.3 is introduced to apply the method of Frölich and Huber (2014) to the parametric structural model.

Assumption 3.3$(a)$ is a pair of exclusion restrictions for $Z_1$ and $Z_2$. It requires that, conditional on $(Z_2, X^\top)^\top$, $Z_1$ can only affect $M$ and $Y$ through its effect on $D$; furthermore, conditional on $(U_D, X^\top)^\top$, $(Z_1, Z_2)$ can only affect $Y$ through their effect on $M$. As mentioned by Frölich and Huber (2014), this assumption is implied by a stricter condition: $(Z_1, Z_2) \perp (U_Y, U_M, U_D)|X$. This condition amounts to a two-IV counterpart to the conventional single-IV exclusion restriction.

Assumption 3.3($b$) requires that, conditional on $X$, $Z_2$ is independent of $Z_1$. As mentioned by Frölich and Huber (2014), this is indeed not restrictive because, if $Z_1$ and $Z_2$ are dependent, we can redefine $Z_2$ by removing the dependence of the original $Z_2$ on $Z_1$.

Assumption 3.3($c$) means that $Z_1$ is effective for generating exogenous variations to $D$, and can be easily examined by estimation. Given (3.3), this has two important implications. First, the sign restriction: $\gamma_{z1} > 0$ implies that

$$D(1) = 1(\gamma_{z1} + X^\top \gamma_x + U_D \geq 0) \geq 1(X^\top \gamma_x + U_D \geq 0) = D(0) \tag{3.4}$$

so there are no defiers in the population. Second, it also implies that

$$P(D(1) = 1|X) > P(D(0) = 1|X), \text{ a.s. in } X,$$

and hence

$$\Delta \equiv E[D(1)] - E[D(0)] > 0; \tag{3.5}$$

that is, the subpopulation of compliers are of positive measure. Conditions (3.4) and (3.5) are basic for defining the CLATE and LATE components, and are standard in treatment effect literature.

Assumption 3.3($d$) is introduced for mediation analysis; see Frölich and Huber (2014, p.18). The sign restriction: $\alpha_{z_2} > 0$ requires $Z_2$ to generate exogenous variations to $M$; it can also be easily examined by estimation. This assumption also requires that $U_M$ be a continuous random variable with a strictly increasing distribution function.

Assumption 3.3($e$) implies that the probabilities of the four combinations of $(D, M)$ are all strictly greater than zero conditional on $(U_M, X, \mathcal{C})$. This means that the weight function defined in (3.12) is neither zero nor infinity; this is required for identifying the conditional mean of $Y(d, M(d'))$ on the subpopulation of compliers. See also Frölich and Huber (2014, p.22) for a similar discussion. This assumption also implies $0 < P(Z_1 = 1|X) < 1$ almost surely; that is, the common support condition for identifying LATE in treatment effect literature, e.g., Abadie (2003, Theorem 3.1). This is a key condition in the identification of

$$E\Big[\Big(Y(d, M(d')) - h_{d'}(X, \alpha_m, b^h_{d1}, b^h_{d0}) + g_{d'}(X, \alpha_m, b^g_{d1}, b^g_{d0})\Big)^2 |\mathcal{C}\Big]$$

if we are only interested in the case where $(d, d') = (1, 1)$ or $(d, d') = (0, 0)$ for program evaluation.

In addition, we make the following assumption:

**Assumption 3.4** *(***Parametric models for** $Z_1$**,** $Z_2$ **and** $(U_M, U_D)$*) Assume that*

*(a) (Model for $Z_1$) $Z_1 = 1(X^\top \alpha_{z_1} + U_{Z_1} \geq 0)$ with $U_{Z_1} \perp X$, and the CDF of $U_{Z_1}$, $F_{U_{Z_1}}(\cdot)$, is known.*

*(b) (Model for $Z_2$) $Z_2 \perp (U_D, X)$ and its CDF, $F_{Z_2}(\cdot, \alpha_f)$, is continuous and known up to the parameter vector $\alpha_f$.*

*(c) (Bivariate model for $(U_M, U_D)$) $(U_M, U_D) \perp (Z_1, Z_2, X)$ and its bivariate CDF, $F_{U_{M,D}}(\cdot, \cdot, \rho_{md})$, is continuous and known up to the parameter vector $\rho_{md}$, and has the marginal distribution functions: $F_{U_M}(\cdot)$ for $U_M$ and $F_{U_D}(\cdot)$ for $U_D$.*

This assumption is introduced for two reasons; it implies that the weight function defined in (3.12) has a parametric form, and it allows us to simplify the asymptotic theory which would have been much more complicated if we had not imposed the parametric models of $Z_1$, $Z_2$, and $(U_M, U_D)$.

Assumption 3.4$(a)$ implies a parametric instrument propensity score model:

$$E[Z_1|X] = Q_{Z_1}(X^\top \alpha_{z_1}), \tag{3.6}$$

where $Q_{Z_1}(s) \equiv 1 - F_{U_{Z_1}}(-s)$ for $s \in R$. This model allows us to present the identification result in Theorem 3.2, implement the statistical inference in a simpler manner, and identify $\alpha_{z_1}$ from the distribution of $(Z_1, X^\top)^\top$.

Assumption 3.4($b$) implies that the probability density function (PDF) of $Z_2|U_D, X$ can be expressed as $f_{Z_2}(\cdot, \alpha_f)$, where $f_{Z_2}(s, \alpha_f) \equiv \partial F_{Z_2}(s, \alpha_f)/\partial s$ for $s \in R$; this model will be applied to the derivation of Theorem 3.1.

Assumption 3.4($c$), together with Assumption 3.2($a$), implies a bivariate binary response model for $(M, D)|Z_1, Z_2, X$, which is denoted by $P(M, D|Z_1, Z_2, X, \eta)$, with the parameter vectors $\eta \equiv (\alpha_q^\top, \gamma^\top, \rho_{md}^\top)^\top$, $\alpha_q \equiv (\alpha_d, \alpha_{z_2}, \alpha_x^\top)^\top$, and $\gamma \equiv (\gamma_{z1}, \gamma_x^\top)^\top$. Recall that $D$ and $M$ are allowed to be endogenous for $Y$ because $U_Y$ is allowed to be dependent on $(U_M, U_D)$. Importantly, $P(M, D|Z_1, Z_2, X, \eta)$ also allows $D$ to be endogenous for $M$ because $U_M$ is allowed to be dependent on $U_D$. In addition, $P(M, D|Z_1, Z_2, X, \eta)$ also encompasses the marginal model for $D|Z_1, X$:

$$E[D|Z_1, X] = Q_D(\gamma_{z1} Z_1 + X^\top \gamma_x), \tag{3.7}$$

where $Q_D(s) \equiv 1 - F_{U_D}(-s)$ for $s \in R$.

Given $P(M, D|Z_1, Z_2, X, \eta)$, we can define the following conditional probability:

$$\begin{aligned} &\Psi_{d'}(Z_2, X, \eta) \\ &\equiv P(U_M \geq -(\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x)| - (\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x). \end{aligned} \tag{3.8}$$

In the supplementary appendix, we prove the following theorem:

**Theorem 3.1** *Suppose that Assumption 3.2*($a$) *and Assumptions 3.4*($b$) *and* ($c$) *hold. For* $d' \in \{0, 1\}$,

$$E[M(d')|Z_2, X, \mathcal{C}] = \Psi_{d'}(Z_2, X, \eta) \tag{3.9}$$

*and*

$$E[M(d')|X, \mathcal{C}] = m_{d'}(X, \alpha_m) \equiv \int_R \Psi_{d'}(z_2, X, \eta) f_{Z_2}(z_2, \alpha_f) \mathrm{d}z_2, \tag{3.10}$$

*with the parameter vector* $\alpha_m \equiv (\eta^\top, \alpha_f^\top)^\top$.

The result in (3.10) is required for us to define $h_{d'}(X, \alpha_m, \beta_{d1}^h, \beta_{d0}^h)$ in (2.18) and $g_{d'}(X, \alpha_m, \beta_{d1}^h, \beta_{d0}^h)$ in (2.19). It shows that, under suitable conditions, the potential-mediator model for $E[M(d')|X, \mathcal{C}]$ in (2.16) is a linear mixture of $\Psi_{d'}(z_2, X, \eta)$ in (3.8) with various $z_2$ weighted by various associated $f_{Z_2}(z_2, \alpha_f)$. Correspondingly, $\alpha_m$ is composed of $\eta$ and $\alpha_f$ that are respectively identifiable from the distribution of $(M, D, Z_1, Z_2, X^\top)^\top$ and that of $(Z_2, X^\top)^\top$. Importantly, this shows the validity of Assumption 3.1($a$).

Denote $\alpha \equiv (\alpha_m^\top, \alpha_{z_1}^\top)^\top$ and recall that $\Delta$ is defined in (3.5). In the supplementary appendix, we prove the following result by matching Assumptions 3.2 and 3.3 with those of Theorem 4 of Frölich and Huber (2014):

**Theorem 3.2** *Suppose that Assumptions 3.2, 3.3, 3.4 hold. For* $d, d' \in \{0, 1\}$,

$$\begin{aligned}&E\Big[\Big(Y(d, M(d')) - h_{d'}(X, \alpha_m, b_{d1}^h, b_{d0}^h) + g_{d'}(X, \alpha_m, b_{d1}^g, b_{d0}^g)\Big)^2 \Big| \mathcal{C}\Big] \\ =&E\Big[w(d, d', \alpha_w)\Big(Y - h_{d'}(X, \alpha_m, b_{d1}^h, b_{d0}^h) + g_{d'}(X, \alpha_m, b_{d1}^g, b_{d0}^g)\Big)^2\Big]\Big/\Delta,\end{aligned} \tag{3.11}$$

*where* $\alpha_w \equiv (\alpha_{z_1}^\top, \alpha_\varphi^\top)^\top$, $\alpha_\varphi \equiv (\alpha_d, \alpha_{z_2}, \alpha_f^\top)^\top$, *and*

$$w(d, d', \alpha_w) \equiv \begin{cases} D\lambda(Z_1, X, \alpha_{z_1}), & (d, d') = (1, 1), \\ D\varphi_1(Z_2, \alpha_\varphi)\lambda(Z_1, X, \alpha_{z_1}), & (d, d') = (1, 0), \\ (D-1)\varphi_2(Z_2, \alpha_\varphi)\lambda(Z_1, X, \alpha_{z_1}), & (d, d') = (0, 1), \\ (D-1)\lambda(Z_1, X, \alpha_{z_1}), & (d, d') = (0, 0), \end{cases} \tag{3.12}$$

*with*

$$\lambda(Z_1, X, \alpha_{z_1}) \equiv \frac{Z_1 - Q_{Z_1}(X^\top \alpha_{z_1})}{Q_{Z_1}(X^\top \alpha_{z_1})(1 - Q_{Z_1}(X^\top \alpha_{z_1}))},$$

$$\varphi_1(Z_2, \alpha_\varphi) \equiv \frac{f_{Z_2}(Z_2 + \alpha_d/\alpha_{z_2}, \alpha_f)}{f_{Z_2}(Z_2, \alpha_f)},$$

*and*

$$\varphi_2(Z_2, \alpha_\varphi) \equiv \frac{f_{Z_2}(Z_2 - \alpha_d/\alpha_{z_2}, \alpha_f)}{f_{Z_2}(Z_2, \alpha_f)}.$$

This theorem allows us to represent the minimizer of the conditional MSE of potential outputs in (3.1) as the minimizer of the weighted MSE of actual outputs:

$$\beta_d \equiv \arg\min_{b_d \in \mathcal{B}_d} \sum_{d'=0,1} E\Big[w(d, d', \alpha_w)\Big(Y - h_{d'}(X, \alpha_m, b^h_{d1}, b^h_{d0}) + g_{d'}(X, \alpha_m, b^g_{d1}, b^g_{d0})\Big)^2\Big], \tag{3.13}$$

for $d \in \{0, 1\}$, where the weights: $w(1, 1, \alpha_w)$, $w(1, 0, \alpha_w)$, $w(0, 1, \alpha_w)$ and $w(0, 0, \alpha_w)$ are functions of $(D, Z_1, Z_2, X^\top)^\top$. Importantly, unlike (3.1), (3.13) is an unconditional mean of observable $Y$, as determined by the distribution of $\boldsymbol{W}$. Therefore, given the identification of $\alpha$, we can identify the parameter vector $\beta_d$ according to (3.13) based on Assumptions 3.1 and 3.2.

The above results demonstrate that the parameter vectors: $\alpha$ and $\beta_d$, for $d \in \{0, 1\}$, are identifiable from the distribution of $\boldsymbol{W}$. In principle, we may replace the role of $w(d, d', \alpha_w)$ in the identification of $\beta_d$ with the non-parametric weight of Frölich and Huber (2014); see the proof of Theorem 3.2 in the supplementary appendix. This replacement requires that we estimate $E[Z_1|X]$ and the conditional distribution of $Z_2|X, \mathcal{C}$ via suitable non-parametric estimators, and the asymptotics of the resulting estimator for $\beta_d$ and the associated assumptions would thus be more complicated.

## 3.2 Identification of CLATE and LATE

Note that the components of the CLATE, CDLATE and CILATE decompositions in (2.20), (2.21) and (2.22) are determined by $m_{d'}(x, \alpha_m)$, $h(x, \beta^h_{dj})$ and $g(x, \beta^g_{dj})$ for

$d, d', j \in \{0,1\}$, and the parameter restrictions in the program-evaluation hypotheses: $H_o^{\mathrm{p}}$, $H_o^{\mathrm{p}h}$ and $H_o^{\mathrm{p}g}$ and the mediation-analysis hypotheses: $H_o^{\mathrm{d}}$, $H_o^{\mathrm{d}h}$, $H_o^{\mathrm{d}g}$, $H_o^{\mathrm{i}}$, $H_o^{\mathrm{i}h}$ and $H_o^{\mathrm{i}g}$ are determined by $\beta_0$, $\beta_1$ and $\alpha_d$. Thus, these conditional effects are also identified following the parameter identification discussed in Section 3.1.

We let $f_{X|\mathcal{C}}(\cdot)$ be the conditional PDF of $X|\mathcal{C}$, and $f_X(\cdot)$ be the unconditional PDF of $X$ to discuss the identification of the unconditional effects. Note that $LATE$ and $CLATE(x)$ have the relationship:

$$\begin{aligned} LATE &= \int CLATE(x) f_{X|\mathcal{C}}(x)\mathrm{d}x = \int \left( CLATE(x) \frac{P(\mathcal{C}|X=x)}{P(\mathcal{C})} \right) f_X(x)\mathrm{d}x \\ &= E\left[ \frac{\Delta(X)}{\Delta} \times CLATE(X) \right], \end{aligned} \tag{3.14}$$

where

$$\Delta(X) \equiv P(\mathcal{C}|X) = P(D(1)=1, D(0)=0|X). \tag{3.15}$$

This relationship also holds for the components of the LATE decompositions in (2.23), (2.24), (2.25), and (2.26).

In accordance with the law of total probability, we have the result:

$$\begin{aligned} P(D(1)=1|X) &= P(D(1)=1, D(0)=1|X) + P(D(1)=1, D(0)=0|X), \\ P(D(0)=1|X) &= P(D(1)=1, D(0)=1|X) + P(D(1)=0, D(0)=1|X). \end{aligned} \tag{3.16}$$

As mentioned before, the sign restriction $\gamma_{z1} > 0$ in Assumption 3.3($c$) implies no defiers. Thus, using (3.15) permits the restriction: $P(D(1)=0, D(0)=1|X) = 0$ to therefore simplify (3.16) as

$$\Delta(X) = P(D(1)=1|X) - P(D(0)=1|X) = E[D(1)|X] - E[D(0)|X]. \tag{3.17}$$

In addition, (3.7) implies a parametric model for the potential treatment status:

$$E[D(z_1)|X] = Q_D(\gamma_{z1} z_1 + X^\top \gamma_x) \tag{3.18}$$

and hence the following models for $\Delta(X)$ and $\Delta$:

$$\Delta^*(X,\gamma) \equiv Q_D(\gamma_{z1} + X^\top\gamma_x) - Q_D(X^\top\gamma_x) \tag{3.19}$$

and

$$\Delta^*(\gamma) \equiv E[Q_D(\gamma_{z1} + X^\top\gamma_x)] - E[Q_D(X^\top\gamma_x)]. \tag{3.20}$$

Accordingly, we can easily transform the components of the CLATE decompositions to their LATE counterparts in (2.24), (2.25), and (2.26):

$$LATE_h = \mathcal{H}_{11}(\nu_1^h) - \mathcal{H}_{00}(\nu_0^h), \quad LATE_g = \mathcal{G}_{11}(\nu_1^g) - \mathcal{G}_{00}(\nu_0^g),$$

$$DLATE_h = \mathcal{H}_{11}(\nu_1^h) - \mathcal{H}_{01}(\nu_0^h), \quad DLATE_g = \mathcal{G}_{11}(\nu_1^g) - \mathcal{G}_{01}(\nu_0^g),$$

$$ILATE_h = \mathcal{H}_{01}(\nu_0^h) - \mathcal{H}_{00}(\nu_0^h), \quad ILATE_g = \mathcal{G}_{01}(\nu_0^g) - \mathcal{G}_{00}(\nu_0^g),$$

where $\nu_d^h \equiv (\alpha_m^\top, \beta_{d1}^{h\top}, \beta_{d0}^{h\top}, \gamma^\top)^\top$, $\nu_d^g \equiv (\alpha_m^\top, \beta_{d1}^{g\top}, \beta_{d0}^{g\top}, \gamma^\top)^\top$,

$$\mathcal{H}_{dd'}(\nu_d^h) \equiv E\Big[\frac{\Delta^*(X,\gamma)}{\Delta^*(\gamma)} \times h_{d'}(X, \alpha_m, \beta_{d1}^h, \beta_{d0}^h)\Big], \tag{3.21}$$

and

$$\mathcal{G}_{dd'}(\nu_d^g) \equiv E\Big[\frac{\Delta^*(X,\gamma)}{\Delta^*(\gamma)} \times g_{d'}(X, \alpha_m, \beta_{d1}^g, \beta_{d0}^g)\Big], \tag{3.22}$$

for $d, d' \in \{0,1\}$. Thus, these unconditional effects are also identified by following the parameter identification discussed in Section 3.1.

## 3.3 The Dam Example: Identification

In the above discussions, Assumptions 3.1, 3.2, and 3.4 involve a set of parametric models that include $h(X, \beta_{dj}^h)$, $g(X, \beta_{dj}^g)$, $\Psi_{d'}(Z_2, X, \eta)$, $Q_{Z_1}(X^\top\alpha_{z_1})$ and $f_{Z_2}(z_2, \alpha_f)$ as key ingredients. Empirical specifications of these ingredients used in the dam example will be provided in Section 5.2. Assumption 3.3 is particularly important for parameter identification because it is required for transforming (3.1) to (3.13). Thus, we assess the validity of the assumption in the dam example in this subsection.

We use the river gradient to construct $Z_1$ (for $D$) and the river length information to establish $Z_2$ (for $M$); see Section 2.5 and Table 2. We argue that a district's river gradient does not impact agricultural production directly, but may indirectly exert effects through the implied feasibility of dam construction (on which basis $Z_1$ is proposed) or through its association with the district's geographic properties that may affect production, such as the elevation of a district and whether it is in the river's upstream (steeper gradient) or downstream (smaller gradient); see, for example, Fisher and Cook (2013). The effects of these geographical properties are controlled for in our analysis using a vector of suitably selected $X$ (see Section 5.1), which is comprised of environmental variables and geographical conditions relevant to agricultural production. The scenario for the effect of river length is similar. After controlling for $X$, the length of the river itself has no effect on the agricultural production other than facilitating irrigation for production (on which basis $Z_2$ is proposed).

In the above scenarios, the river-gradient-based $Z_1$ and the river-length-based $Z_2$ satisfy the exclusion restriction, because after controlling $X$, $Z_1$ and $Z_2$ would only influence $Y$ through their influence on $D$ and $M$. Thus, Assumption 3.3($a$) should hold in the dam example.

In addition, Assumption 3.3($b$) is also plausible, at least in a weaker form of uncorrelatedness in the dam example, because, as stated in Table 2, $Z_2$ is the OLS residual of the regression of the original river-length variable on $Z_1$ and $X$. By construction, this $Z_2$ is uncorrelated with $Z_1$, as implied by Assumption 3.3($b$). The inclusion of $X$ in this regression is motivated by Assumption 3.4(b), which implies that $Z_2$ is uncorrelated with $X$. As will be shown in Section 5.2, the sign restrictions in Assumptions 3.3($c$) and ($d$) also hold for the dam example, and Assumption 3.3($e$) is also satisfied for the models specified in the dam example.

# 4 Estimation

Denote $\boldsymbol{W}_i \equiv (Y_i, M_i, D_i, Z_{1i}, Z_{2i}, X_i^\top)^\top$, and let $\{\boldsymbol{W}_i\}_{i=1}^n$ be a random sample of $\boldsymbol{W}$ with sample size $n$. The identification results in Section 3.1 suggest that we may estimate the parameter vectors $\alpha$ and $\beta_d$, for $d \in \{0, 1\}$, in a two-step manner. In the first step, we estimate the parameter vector $\alpha$. In the second step, given the estimation of $\alpha$, we estimate $\beta_d$ via the WNLSE based on (3.13) for $d \in \{0, 1\}$. In this section, we first introduce this two-stage estimation method for $\alpha$ and $\beta_d$, and then discuss how to estimate the CLATE and LATE components based on the $\alpha$ and $\beta_d$ estimates.

## 4.1 First-Step Estimation

When estimating $\alpha$, recall that $\alpha = (\alpha_m^\top, \alpha_{z_1}^\top)^\top$, where $\alpha_m = (\eta^\top, \alpha_f^\top)^\top$ is composed of the parameter vector $\eta$ of the bivariate binary response model $P(M, D|Z_1, Z_2, X, \eta)$ and the parameter vector $\alpha_f$ of the conditional PDF $f_{Z_2}(\cdot, \alpha_f)$, and $\alpha_{z_1}$ is the parameter vector of the instrument propensity score model in (3.6). Because these three models are separable and all established in the parametric context, we can then estimate $\eta$, $\alpha_f$ and $\alpha_{z_1}$ separately by the Maximum Likelihood (ML) method.

Let $\hat{\eta}$, $\hat{\alpha}_f$ and $\hat{\alpha}_{z_1}$ be respectively the ML estimators (MLEs) for $\eta$, $\alpha_f$ and $\alpha_{z_1}$; see Section 2 of the supplementary appendix for their log-likelihood functions. We can estimate $\alpha$ by $\hat{\alpha} \equiv (\hat{\eta}^\top, \hat{\alpha}_f^\top, \hat{\alpha}_{z_1}^\top)^\top$. In addition, corresponding to the fact that

$\eta$ comprises $\alpha_q = (\alpha_d, \alpha_{z_2}, \alpha_x^\top)^\top$, $\gamma = (\gamma_{z1}, \gamma_x^\top)^\top$ and $\rho_{md}$ as subvectors, we can also split $\hat{\eta}$ into $\hat{\alpha}_d$, $\hat{\alpha}_{z_2}$, $\hat{\alpha}_x$, $\hat{\gamma}_{z1}$, $\hat{\gamma}_x$ and $\hat{\rho}_{md}$ to estimate $\alpha_d$, $\alpha_{z_2}$, $\alpha_x$, $\gamma_{z1}$, $\gamma_x$ and $\rho_{md}$, respectively.

## 4.2 Second-Step Estimation

Let $w_i(d, d', \alpha_w)$ be the $w(d, d', \alpha_w)$ in (3.12) evaluated at $\boldsymbol{W} = \boldsymbol{W}_i$. Given estimator $\hat{\alpha}$, we can define the second-step WNLSE for $\beta_d$ as:

$$\hat{\beta}_d \equiv (\hat{\beta}_{d1}^{h\top}, \hat{\beta}_{d0}^{h\top}, \hat{\beta}_{d1}^{g\top}, \hat{\beta}_{d0}^{g\top})^\top = \arg\min_{b_d \in \mathcal{B}_d} \frac{1}{n} \sum_{i=1}^{n} \varrho_{d,i}(\hat{\alpha}, b_d), \tag{4.1}$$

where for $d \in \{0, 1\}$

$$\varrho_{d,i}(\alpha, b_d) \equiv \sum_{d'=0,1} w_i(d, d', \alpha_w) \Big( Y_i - h_{d'}(X_i, \alpha_m, b_{d1}^h, b_{d0}^h) + g_{d'}(X_i, \alpha_m, b_{d1}^g, b_{d0}^g) \Big)^2 . \tag{4.2}$$

Note that (4.1) is the sample analogue of (3.13).

Because $\hat{\beta}_d$ is the solution of the estimating equation:

$$\frac{1}{n} \sum_{i=1}^{n} \nabla_{b_d} \varrho_{d,i}(\hat{\alpha}, \hat{\beta}_d) = 0, \tag{4.3}$$

we may also interpret $\hat{\beta}_d$ as the method-of-moments estimator for $\beta_d$:

$$\hat{\beta}_d = \arg\min_{b_d \in \mathcal{B}_d} \left[ \frac{1}{n} \sum_{i=1}^{n} \nabla_{b_d} \varrho_{d,i}(\hat{\alpha}, b_d) \right]^\top \left[ \frac{1}{n} \sum_{i=1}^{n} \nabla_{b_d} \varrho_{d,i}(\hat{\alpha}, b_d) \right] . \tag{4.4}$$

In fact, we solve $\hat{\beta}_d$ by (4.4). The main reason is that the weight $w_i(d, d', \alpha_w)$ can be negative, so the problem in (4.1) is not strictly concave in $\beta$, which makes it difficult to solve.

## 4.3 Estimation on CLATE and LATE

Given $\hat{\alpha}$ and $\hat{\beta}_d$, for $d \in \{0, 1\}$, we can immediately estimate the components of the CLATE, CDLATE, and CILATE decompositions in (2.20), (2.21), and (2.22) via plug-in principle. Similarly, we can base the tests for $H_o^{\mathrm{p}}$, $H_o^{\mathrm{p}h}$, $H_o^{\mathrm{p}g}$, $H_o^{\mathrm{d}}$, $H_o^{\mathrm{d}h}$, $H_o^{\mathrm{d}g}$, $H_o^{\mathrm{i}}$, $H_o^{\mathrm{i}h}$ and $H_o^{\mathrm{i}g}$ on $\hat{\beta}_{dj}^h$, $\hat{\beta}_{dj}^g$ and $\hat{\alpha}_d$, for $d, j \in \{0, 1\}$.

In addition, we can estimate the components of the LATE, DLATE and ILATE decompositions in (2.24), (2.25) and (2.26) by

$$\hat{\mathcal{H}}_{dd'}(\hat{\nu}_d^h) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta^*(X_i, \hat{\gamma})}{\Delta^*(\hat{\gamma})} \times h_{d'}(X_i, \hat{\alpha}_m, \hat{\beta}_{d1}^h, \hat{\beta}_{d0}^h) \tag{4.5}$$

and

$$\hat{\mathcal{G}}_{dd'}(\hat{\nu}_d^g) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta^*(X_i, \hat{\gamma})}{\Delta^*(\hat{\gamma})} \times g_{d'}(X_i, \hat{\alpha}_m, \hat{\beta}_{d1}^g, \hat{\beta}_{d0}^g), \tag{4.6}$$

where $\hat{\nu}_d^h \equiv (\hat{\alpha}_m^\top, \hat{\beta}_{d1}^{h\top}, \hat{\beta}_{d0}^{h\top}, \hat{\gamma}^\top)^\top$ and $\hat{\nu}_d^g \equiv (\hat{\alpha}_m^\top, \hat{\beta}_{d1}^{g\top}, \hat{\beta}_{d0}^{g\top}, \hat{\gamma}^\top)^\top$, for $d, d' \in \{0, 1\}$. Meanwhile, we can test $\tilde{H}_o^{\mathrm{p}}$, $\tilde{H}_o^{\mathrm{p}h}$, $\tilde{H}_o^{\mathrm{p}g}$, $\tilde{H}_o^{\mathrm{d}}$, $\tilde{H}_o^{\mathrm{d}h}$, $\tilde{H}_o^{\mathrm{d}g}$, $\tilde{H}_o^{\mathrm{i}}$, $\tilde{H}_o^{\mathrm{i}h}$ and $\tilde{H}_o^{\mathrm{i}g}$ based on $\hat{\mathcal{H}}_{dd'}(\hat{\nu}_d^h)$ and $\hat{\mathcal{G}}_{dd'}(\hat{\nu}_d^g)$, for $d, d' \in \{0, 1\}$.

In the supplementary appendix, we discuss the asymptotic normality of the estimators, and provide a bootstrap method for statistical inference. We also perform a Monte Carlo simulation to assess the finite-sample performance of our WNLSEs for the potential output model. We used the structural model as outlined in Assumption 3.2 to generate the data, and the parameters are tuned to mimic the empirical data of the dam example. The parameters in the first-stage estimation are set to be known for simplicity. Results show that the MSE of the estimators reasonably decreases as the sample size increases, and that the MSE appears to come primarily from the variance of the estimate, rather than the bias. Once the endogeneity is controlled for in the model, it does not have an appreciable impact on the estimates.

# 5 Empirical Illustration

We complete the empirical illustration of the dam example in this section. We first discuss the dataset and details of covariates $X$, then specify the parametric models, and finally present our empirical findings.

## 5.1 Dataset and Covariates

The dataset is based on that used in Duflo and Pande (2007a), although our empirical question differs from theirs. Duflo and Pande (2007a) were interested in the differential output effects of dams in the downstream versus the upstream districts. In comparison, we focus on the upstream districts, and we analyze the total, direct, and indirect effects of dam construction on agricultural production and its frontier and inefficiency components in these districts.

The dataset contains information on agricultural production, geographic characteristics, and the number of dams in India at the district level for 1976-1987. Because of data limitations, we focused on the effect of dams built in 1976-1980 (the treatment period) on agricultural production in 1981-1987 (the evaluation period). Because our estimator requires that each observation be independent, the requirement would be better served by converting the panel data to a cross-section structure. We would then take the sample average for variables over time for each given district. After deleting observations with missing values for key variables, the dataset contained 266 observations in total ($n = 266$). Recall that Table 2 provides definitions of key empirical variables. Additional information on the variables can be found in the Data Appendix of Duflo and Pande (2007a).

For district $i$, the definitions of $Y_i$ ($production_i$), $D_i$, $M_i$, $Z_{1i}$ and $Z_{2i}$ follow

Table 2. Vector $X_i$ is defined as:

$$X_i = (1, rain_i, elevation_i, pre\_dam_i, pre\_fert_i, pre\_land_i, rain2_i, pc_{1i}, pc_{2i})^\top , \quad (5.1)$$

where $rain_i$ is the rainfall variable measured as the fractional deviation of the district's rainfall from the district mean in the evaluation period (1981-1987), $elevation_i$ is the log of the mean elevation of district $i$, and $rain2_i$ is the average annual rainfall variation for 1981-1987 of district $i$. The other three covariates, $pre\_dam_i$, $pre\_fert_i$, and $pre\_land_i$, are used to control for the locality's agricultural production conditions prior to the dam's construction. The variable $pre\_dam_i$ represents the number of dams that existed in district $i$ at the beginning of the treatment period. The variable $pre\_fert_i$ is the fertilizer used prior to the dam's construction, and is measured as the log of the average fertilizer in district $i$ three years prior to the dam's construction. The variable $pre\_land_i$ is constructed in a similar manner, and it measures the log of the average of the gross cultivated area prior to the construction. Similar to the former two covariates, these three pre-treatment covariates are exogenous in our empirical context.

The variables $pc_{1i}$ and $pc_{2i}$ are respectively the first and second principal components of a list of geographic-and-environment related variables. Following Duflo and Pande (2007a), the list of variables includes India state dummies, the predicted number of dams per state in 1970, and the interactions of that predicted number with variables such as district elevations, district slope, and district size as a list of geographical variables that may affect a district's agricultural production, the decision to build dams, and the size of the irrigation area. However, because our model is highly non-linear and the sample size is not large, it is difficult to include the entire list of variables in the analysis. Instead, we can conduct a principal component analysis and employ the first two components in our models. The first two components,

$pc_{1i}$ and $pc_{2i}$, account for almost 70% of the data co-movement.

Here we report the summary statistics of $\boldsymbol{W}_i = (Y_i, M_i, D_i, Z_{1i}, Z_{2i}, X_i^\top)^\top$ in Table 3. As we can see, the means of most variables are statistically different between the treated and the untreated groups. Districts that built more dams in the treatment period (i.e., districts with $D = 1$) appear to be at higher elevations (*elevation*). Rainfall and fertilizer were lower in these districts, and the agricultural production was similarly low. The statistics paint a familiar picture of dams being built in poorer mountainous regions.

## 5.2 Model Specification and Estimation

We let $X^h$ and $X^g$ be two subvectors of $X$, and adopted a popular half-normal specification of the stochastic production frontier model (Aigner et al. 1977, Caudill and Ford 1993): for $d, j \in \{0, 1\}$,

$$Y = h(X, \beta^h_{dj}) + v - u, \tag{5.2}$$

$$h(X, \beta^h_{dj}) = X^{h\top} \beta^h_{dj}, \tag{5.3}$$

$$u \sim N^+(0, \sigma_u^2), \tag{5.4}$$

$$\sigma_u = \exp(X^{g\top} \beta^g_{dj}), \tag{5.5}$$

where we also assume that $v$ is a zero-mean normal random variable. In accordance with (5.4) and (5.5), the function of the conditional mean of $u$ is

$$g(X, \beta^g_{dj}) = \sqrt{\frac{2}{\pi}} \exp(X^{g\top} \beta^g_{dj}). \tag{5.6}$$

More specifically, we set

$$X^h_i = (1, rain_i, elevation_i, pre\_dam_i, pre\_fert_i, pre\_land_i, pc_{1i}, pc_{2i})^\top, \tag{5.7}$$

$$X^g_i = (1, rain2_i)^\top; \tag{5.8}$$

with $\beta_{dj}^h$ and $\beta_{dj}^g$ composed of the following parameters:

$$\beta_{dj}^h = \left(\beta_{dj(0)}^h,\, \beta_{dj(r)}^h,\, \beta_{dj(e)}^h,\, \beta_{dj(p)}^h,\, \beta_{dj(f)}^h,\, \beta_{dj(L)}^h, \beta_{dj(c1)}^h, \beta_{dj(c2)}^h\right)^\top, \tag{5.9}$$

$$\beta_{dj}^g = \left(\beta_{dj(0)}^g,\, \beta_{dj(r2)}^g\right)^\top. \tag{5.10}$$

To justify the choice of $X_i^h$ and $X_i^g$, first note that $rain_i$ is included in the frontier function $h(\cdot)$ because it is a direct input to agricultural production. We also add several environmental variables to $h(\cdot)$ to control for the production conditions: $elevation_i$, $pre\_dam_i$, $pre\_fert_i$, and $pre\_land_i$. Geographical properties are often important to agricultural production and elevation is one of them. For instance, precipitation and temperature may change with the elevation, which in turn may affect agricultural yields. The variable $elevation_i$ is included to control for the effect. The other three variables are included as controls of the initial conditions in the program being evaluated. Given the technology, inputs, and geographical conditions, climate variability and uncertainty may hamper the extent to which the production potential may be realized. Thus, variability and uncertainty affect production efficiency. We can therefore include $rain2_i$ in the inefficiency function $g(\cdot)$. This variable measures the volatility of rainfall in district $i$.

In addition, we specify $P(M, D|Z_1, Z_2, X, \eta)$ as a bivariate probit model that is comprised of sub-models for $M$ and $D$ in Assumption 3.2$(a)$ and sets the distribution $F_{U_{M,D}}(\cdot, \cdot, \rho_{md})$ in Assumption 3.4$(c)$ as a bivariate normal distribution:

$$\left.\begin{bmatrix} U_M \\ U_D \end{bmatrix}\right| (Z_1, Z_2, X) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{md} \\ \rho_{md} & 1 \end{bmatrix}\right). \tag{5.11}$$

As in Wooldridge (2010, p.596), this specification is useful because the MLE $\hat{\eta}$ can be computed using an econometrics package that permits ML estimation for a bivariate probit model. Let $\Phi(\cdot)$ and $\phi(\cdot)$ be respectively the distribution function and the

PDF of $N(0,1)$. In the supplementary appendix, we further show that this bivariate probit model implies the following formula of (3.8):

$$\begin{aligned}\Psi_{d'}(Z_2, X, \eta) = &\frac{1}{\Phi(\gamma_{z1} + X^{\top}\gamma_x) - \Phi(X^{\top}\gamma_x)} \\ &\times \int_{X^{\top}\gamma_x}^{\gamma_{z1}+X^{\top}\gamma_x} \Phi\left(\frac{\alpha_d d' + \alpha_z Z_2 + X^{\top}\alpha_x + \rho_{md}u}{\sqrt{1-\rho_{md}^2}}\right)\phi(u)\mathrm{d}u.\end{aligned} \tag{5.12}$$

Furthermore, it implies the following specification of (3.7):

$$Q_D(\gamma_{z1}Z_1 + X^{\top}\gamma_x) = \Phi(\gamma_{z1}Z_1 + X^{\top}\gamma_x). \tag{5.13}$$

Correspondingly, we specify the instrument propensity score model in (3.6) as a probit model:

$$Q_{Z_1}(X^{\top}\alpha_{z_1}) = \Phi(X^{\top}\alpha_{z_1}). \tag{5.14}$$

In this empirical example, the histogram of $Z_2$ indicates a bimodal distribution, so we use a mixture distribution of two normals to approximate the distribution and set

$$f_{Z_2}(z_2, \alpha_f) = \frac{\alpha_{f(p)}}{\alpha_{f(s1)}}\phi\left(\frac{z_2 - \alpha_{f(m1)}}{\alpha_{f(s1)}}\right) + \frac{1-\alpha_{f(p)}}{\alpha_{f(s2)}}\phi\left(\frac{z_2 - \alpha_{f(m2)}}{\alpha_{f(s2)}}\right), \tag{5.15}$$

where $(\alpha_{f(m1)},\ \alpha_{f(s1)})$ and $(\alpha_{f(m2)},\ \alpha_{f(s2)})$ are respectively the means and the standard deviations of the two normal distributions, and $\alpha_{f(p)}$ and $(1-\alpha_{f(p)})$ are the distributions' weights. Accordingly, the potential mediator model $m_{d'}(\cdot)$ in (3.10) is specified using (5.12) and (5.15). The integration in (3.10) is numerically computed using the Gauss-Hermite quadrature with 60 quadrature points.

Given the specifications of $P(M, D|Z_1, Z_2, X, \eta)$, $f_{Z_2}(z_2, \alpha_f)$ and $Q_{Z_1}(X^{\top}\alpha_{z_1})$, we obtain $\hat{\eta} = (\hat{\alpha}_d, \hat{\alpha}_{z_2}, \hat{\alpha}_x^{\top}, \hat{\gamma}_{z1}, \hat{\gamma}_x^{\top}, \hat{\rho}_{md})^{\top}$, $\hat{\alpha}_f = (\hat{\alpha}_{f(m1)},\ \hat{\alpha}_{f(s1)}, \hat{\alpha}_{f(m2)},\ \hat{\alpha}_{f(s2)},\ \hat{\alpha}_{f(p)})^{\top}$ and $\hat{\alpha}_{z_1}$ as the first-stage MLEs in Section 4.1. To save space, we present the estimates of auxiliary parameters ($\hat{\alpha}_d$, $\hat{\alpha}_{z_2}$, $\hat{\alpha}_{f(m1)}$, $\hat{\gamma}_{z1}$, etc.) in Table A.4 – Table A.6 of the supplementary appendix.

Given $h(X, \beta^h_{dj})$, $g(X, \beta^g_{dj})$, (5.14), $\hat{\alpha} = (\hat{\eta}^\top, \hat{\alpha}_f^\top, \hat{\alpha}_{z_1}^\top)^\top$, and $\hat{\alpha}_w$ denoting the $\alpha_w$-associated subvector of $\hat{\alpha}$, we estimate the weight parameter $w(d, d', \alpha_w)$ in (3.12) using $w(d, d', \hat{\alpha}_w)$ and estimate the parameter vector $\beta_d = (\beta^{h\top}_{d1}, \beta^{h\top}_{d0}, \beta^{g\top}_{d1}, \beta^{g\top}_{d0})^\top$ using the second-stage WNLSE $\hat{\beta}_d = (\hat{\beta}^{h\top}_{d1}, \hat{\beta}^{h\top}_{d0}, \hat{\beta}^{g\top}_{d1}, \hat{\beta}^{g\top}_{d0})^\top$ in (4.4), for $d \in \{0, 1\}$.

Given $\hat{\alpha}$ and $\hat{\beta}_d$, we can obtain the CLATE component estimators as in Section 4.3. Given (5.13) and $\hat{\gamma} = (\hat{\gamma}_{z1}, \hat{\gamma}_x^\top)^\top$, we estimate $\Delta^*(X, \gamma)$ in (3.19) and $\Delta^*(\gamma)$ in (3.20) using $\Delta^*(X, \hat{\gamma})$ and $\Delta^*(\hat{\gamma})$, respectively. Accordingly, the LATE component estimators are based on the $\Delta^*(X_i, \hat{\gamma})$s and $\Delta^*(\hat{\gamma})$, as in Section 4.3. The $\Delta^*(\hat{\gamma})$ also gives us the estimated percentage of compliers in the data. In our example, the estimate is 16% which amounts to 43 districts in the population. Although this is not a large portion and we could not determine the average effect for the entire sample, our analysis nevertheless provides informative results for a well-defined subpopulation that may be useful for policy evaluation. See also Imbens (2010) for a discussion of the relevance of LATE.

Before presenting the main empirical results, we first confirm that Assumption 3.3 holds for this empirical example. Recall that we have presented plausible scenarios for Assumption 3.3($a$) and ($b$) in Section 3.3. Assumption 3.3($c$) and ($d1$) are supported by $\hat{\gamma}_{z1}(= 0.384) > 0$ and $\hat{\alpha}_{z_2}(= 0.224) > 0$ as shown in Table A.4 of the supplementary appendix, and Assumption 3.3($d2$) is supported by the bivariate probit model. In the supplementary appendix, we further show that Assumption 3.3($e$) is supported by our empirical specifications.

A general critique of LATE is that this measure pertains to the group of compliers, which is an unobservable and instrument-dependent subpopulation. In the example of Indian dam construction, this subpopulation constitutes districts that build dams by following technology principles (viz., in accordance with river gradients). Although these dams may not be representative of average dams, the LATE estimates may nevertheless measure the "best case scenario", i.e., the economic impact of technologically appropriate dams (Duflo and Pande 2007a, p.623).

Finally, as pointed out by a referee, the thought experiment underlying the compliers considered in the dam example involves altering the geography of Earth ($Z_1$), while $Z_1$ is unable to be manipulated in practice. This means that although we can evaluate the effect of building dam on agricultural production for the subpopulation of compliers by estimating LATE, the policy applicability of evaluation result may be restricted when $Z_1$ is a geographical IV.

## 5.3 Empirical Results

Table 4 shows the WNLSEs for the $\beta^h_{dj}$ and $\beta^g_{dj}$, for $d, j \in \{0, 1\}$. The results of $(d, M(d')) = (1, M(d'))$ are for the treated group ($D_i = 1$), and those of $(d, M(d')) = (0, M(d'))$ are for the untreated group ($D_i = 0$).

Unsurprisingly, the rainfall variable has strong impacts on agricultural production. In particular, districts with dams and large portions of irrigated areas benefit the most from additional rainfall, when other factors are fixed: 1% more of rainfall above the district's average would increase the frontier production by 1.086% ($\hat{\beta}^h_{11(r)}$) and the effect is large and significant. The benefit is also significant for the untreated districts *with* large portions of irrigated areas ($\hat{\beta}^h_{01(r)} = 1.061\%$). However, additional rainfall appears to adversely affect the frontier production in districts with low irrigation intensity ($M(d') = 0$), especially when the district belongs to the untreated group ($\hat{\beta}^h_{00(r)} = -1.541\%$). These results show the important roles of dams and irrigation systems in managing and regulating water supply.

Regarding the elevation variable, it appears to be insignificant in all of the models. The rest of the variables, including $pre\_land_i$ ($\beta^h_{dj(L)}$), $pre\_fert_i$ ($\beta^h_{dj(f)}$), $pre\_dam_i$ ($\beta^h_{dj(p)}$), $pc_{i1}$ ($\beta^h_{dj(c1)}$), and $pc_{i2}$ ($\beta^h_{dj(c2)}$), are included to control for the initial conditions of district $i$ before the treatment. Since the variables may approximate the local area's initial conditions in regard to economic prosperity, the history of agricultural cultivation, traditional agricultural infrastructures, natural endowments, etc., the signs of the coefficients could be ambiguous.

As for the rainfall volatility variable in the inefficiency function, the estimated coefficients of $\beta^g_{11(r2)}$, $\beta^g_{10(r2)}$, $\beta^g_{01(r2)}$, and $\beta^g_{00(r2)}$ are all positively significant. Though the non-linearity of the inefficiency function does not permit a direct marginal effect interpretation of the coefficients, the positive signs of the coefficients indicate that volatility in the rainfall hinders the production potential from being fully realized and thus causes inefficiency to increase.

To formally evaluate the effects of the dams and mediation channel, we perform hypothesis tests on CLATE and LATE, as shown in Sections 2.2 and 2.3:

$$\begin{aligned} CLATE(x) &= E[Y(1, M(1))|X = x, \mathcal{C}] - E[Y(0, M(0))|X = x, \mathcal{C}], \\ LATE &= E[Y(1, M(1))|\mathcal{C}] - E[Y(0, M(0))|\mathcal{C}]. \end{aligned}$$

The $x$ is a value vector of $X$ specified in (5.1). A different $x$ is referred to as a different type of district in the example. A rejection of the null hypothesis of CLATE($x$) $= 0$ for all $x$s, therefore, indicates that potential outputs with and without treatment are not equal for at least some of the districts in the subpopulation of compliers. LATE, on the other hand, is a test of the effect over the averaged compliers.

For both CLATE and LATE, we decompose them into various sub-effects (see Table 1). Thus, a total of nine testable hypotheses/effects for each CLATE and LATE are discussed in Sections 2.3 and 2.4.

Table 5 presents the results of the tests of CLATE. The first row reports the results of testing the null hypotheses that building dams (the treatment) has no effects on the output ($H_o^{\mathrm{p}} : CLATE(x) = 0,\ \forall x$), frontier function ($H_o^{\mathrm{p}h} : CLATE_h(x) = 0,\ \forall x$), and inefficiency ($H_o^{\mathrm{p}g} : CLATE_g(x) = 0,\ \forall x$). The second and third rows show the results of the mediation analysis (i.e., the tests of the direct and indirect effects of building dams on the output, frontier function, and inefficiency). We report the test statistics and associated $p$-values based on $\chi_k^2$ distributions, where the degrees of freedom, $k$s, are equal to the number of test restrictions. For instance, there are two coefficients in the $g(\cdot)$ function, so the null hypothesis $H_o^{\mathrm{p}g}$ would contain six coefficient restrictions across the four vectors of $\beta_{dj}^g$ for $d, j \in (0, 1)$. The null distribution of this test statistic is thus $\chi_6^2$.

The results indicate that we can convincingly reject all but one of the nine null hypotheses with no effect. The implication is that, at least for some districts among the compliers, building dams (the treatment) has a significant total and direct impact on agricultural production and a significant indirect impact on production through the irrigated lands (the mediator).

The results of CLATE, however, do not tell us which type of district would have experienced strong impacts from the dams. We conduct a simple analysis to obtain insights. Given that the impact is likely to be negative (see the LATE results), we look for differences between the group of districts with the lowest 10% of CLATE and the group of districts in the rest of the sample. We conduct $t$ tests on the equality of the means of variables in $X$ between the two groups. The tests show that the districts that are most likely to be adversely affected by the dams are those that have less rainfall ($rain_i$), are located in higher-elevation areas ($elevation_i$), and have smaller pre-existing cultivation land ($pre\_land_i$) and fewer pre-existing dams ($pre\_dam_i$). The two other geographical-related principal components ($pc_{1i}$ and $pc_{2i}$) are also statistically different between the two groups. The results paint a picture of districts with less rainfall and on high-elevation terrain where the agricultural activity had been less developed prior to the dam construction. In these places, the adverse effects of dams out-weigh the benefits. It is plausible that the decision to build a dam in this type of district is mainly to benefit areas downstream of the dam rather than the area in the vicinity of the dam.

Table 6 presents the estimates of LATE and its components, and the content is arranged in a manner similar to that in Table 5. The results indicate that dam construction may negatively impact the local's agricultural production with an effect of $-4.3\%$ of total output (LATE$= -0.043$). The mediation effect through irrigation appears to be positive, contributing to 2.8% increase of total output (ILATE$= 0.028$). The remaining direct effect is negative (DLATE$= -0.07$). Nevertheless, all of the three effects on overall output are statistically insignificant.

If we decompose the output effect into the technology frontier and technical inefficiency components, we find that dams' effects on each of the individual components are much greater. In particular, local dams cause a reduction in the technology frontier and the effect amounts to $-18.5\%$ of total output (LATE$_h = -0.185$). As we mentioned in Section 2.1, technology regression may result from waterlogging and soil salinization after dam construction, and environmental protection actions may also limit the use of certain farming methods (e.g., the use of pesticides and chemical fertilizers) in the neighborhoods of the dams.

As for the impact on the technical inefficiency, we have LATE$_g = -0.142$. The figure indicates efficiency improvement (inefficiency reduction) which may happen because of better infrastructure and flood control brought about by dam constructions. The effect amounts to a 14.2% increase in output which is statistically significant at the 1% level. Therefore, for the subpopulation of compliers, building a dam in the locality may be generally helpful in improving the production efficiency (for the given technology). The mediation analysis further shows that the inefficiency-reducing effect mainly derives directly from the dams: $\text{DLATE}_g = -0.081$ which is statistically significant.

It is interesting to note that dams' effects on the technology frontier and the technical inefficiency are nontrivial but in opposite directions. When adding up, the effects offset each other. Even though the total output effect is small, compositions of the effects change quite a bit as indicated by the changes in the frontier and the efficiency. The agricultural production activities, therefore, are nontrivially affected by the the dam construction. One of the implications to policy makers is to find ways to mitigate the adverse frontier effect and re-enforce the positive efficiency effect when planning dam constructions in the future.

Taken together, building dams *in the local areas* may have significant effects on some types of districts with certain $x$ among the compliers, but the average effects on the compliers (when $X$ is integrated out) are mostly insignificant albeit negative. The latter result is consistent with that of Duflo and Pande (2007a), whose regression analysis also shows an insignificant change in agricultural production in the district where a dam is located. Our stochastic frontier analysis, nevertheless, sheds further light on the results: building dams may actually improve the local area's production efficiency directly (as opposed to indirectly through the mediator), though the positive impact due to the efficiency improvement is offset by the negative impact resulting from the deterioration in the frontier function of production.

We should note at this point that while dams would most likely benefit downstream districts, our focus in this study is on upstream districts (areas in the vicinity of dams) on which the impact of dams appears to be controversial. Our results help to shed light into this issue. The results, on the other hand, shall not be taken as an assessment on the overall value of dams, for such an assessment should account for dams' effect on the downstream districts as well.

## 6 Conclusion

We propose a new fully parametric stochastic frontier model with an endogenous treatment status and an endogenous mediator in which we use a two-IV method, as in Frölich and Huber's study (2014), to handle the endogeneity issue. This model is then used to estimate a program's total productivity effect and decompose it into technology and efficiency components. The decomposition allows policy-makers to identify whether and how the policy affects production technology and efficiency. The mediation analysis, which is also part of the model, further allows us to test whether the aforementioned effects occur directly as a result of the program or indirectly via a mediator. The results of the tests may help policy-makers design better policy mechanisms, and the mediator itself may also serve as an indicator in program evaluation.

We illustrate the application of the model using data from India to estimate the effects of building large dams on upstream districts' agricultural production. Our results show that dams may directly improve the local area's production efficiency, but the overall effect on output was insignificant. This result is generally consistent with that of Duflo and Pande (2007a), who found that a dam would significantly increase the agricultural production in districts downstream from the dam, but its effect on the production of the local district was rather limited.

Table 3: Summary Statistics of Main Variables

| | All | D=0 | D=1 | diff |
|---|---|---|---|---|
| *production* | 9.771 | 9.998 | 9.519 | 0.479 |
| | (0.856) | (0.925) | (0.694) | (0.101)*** |
| $M_i$ | 0.500 | 0.686 | 0.294 | 0.392 |
| | (0.501) | (0.466) | (0.457) | (0.057)*** |
| $Z_{1i}$ | 0.500 | 0.400 | 0.611 | -0.211 |
| | (0.501) | (0.492) | (0.489) | (0.060)*** |
| $Z_{2i}$ | 0 | 0.134 | -0.149 | 0.283 |
| | (1.592) | (1.533) | (1.648) | (0.195) |
| *rain* | -0.020 | 0.009 | -0.052 | 0.061 |
| | (0.099) | (0.106) | (0.077) | (0.012)*** |
| *elevation* | 5.552 | 5.283 | 5.852 | -0.569 |
| | (0.663) | (0.619) | (0.579) | (0.074)*** |
| *pre_dam* | 0.053 | 0.013 | 0.097 | -0.084 |
| | (0.085) | (0.028) | (0.104) | (0.009)*** |
| *pre_fert* | 16.145 | 16.336 | 15.934 | 0.402 |
| | (1.249) | (1.337) | (1.111) | (0.152)*** |
| *pre_land* | 6.250 | 6.205 | 6.300 | -0.095 |
| | (0.523) | (0.556) | (0.481) | (0.064) |
| $pc_1$ | 0 | 0.694 | -0.771 | 1.465 |
| | (6.809) | (7.394) | (6.030) | (0.833)* |
| $pc_2$ | 0 | -1.564 | 1.737 | -3.301 |
| | (3.173) | (1.540) | (3.596) | (0.333)*** |
| $rain2$ | 0.062 | 0.062 | 0.061 | 0.001 |
| | (0.062) | (0.059) | (0.065) | (0.008) |
| # of obs. | 266 | 140 | 126 | |

Note 1: Figures in the 2nd to the 4th columns are the means and the standard deviations (in the parentheses) of the respective variables. The 5th column reports the difference between the variables in $D = 0$ and $D = 1$ and its standard error (in the parentheses). "*" represents the statistical significance from two-sample $t$-tests on the differences at 10% level; "***" indicates the significance at the 1% level.

Note 2: $Z_{2i}$ is obtained from OLS residuals (see Table 2) and $pc_1$ and $pc_2$ are principal components from a vector of geographical and environmental variables. The means of these variables are 0 by construction.

Table 4: Estimation Results

|  | $(d, M(d')) = (1, 1)$ |  |  | $(d, M(d')) = (0, 1)$ |  |
|---|---|---|---|---|---|
| $h$ function | est. | std.err. | $h$ function | est. | std.err. |
| $\beta^h_{11(0)}$ | -0.345 | 0.248 | $\beta^h_{01(0)}$ | -0.531* | 0.291 |
| $\beta^h_{11(r)}$ | 1.086* | 0.596 | $\beta^h_{01(r)}$ | 1.061* | 0.636 |
| $\beta^h_{11(e)}$ | -0.203 | 0.810 | $\beta^h_{01(e)}$ | -0.617 | 0.640 |
| $\beta^h_{11(p)}$ | 0.847 | 0.677 | $\beta^h_{01(p)}$ | -0.284 | 0.179 |
| $\beta^h_{11(f)}$ | 0.740 | 0.503 | $\beta^h_{01(f)}$ | 0.775** | 0.361 |
| $\beta^h_{11(L)}$ | -0.162 | 1.051 | $\beta^h_{01(L)}$ | 0.231 | 0.782 |
| $\beta^h_{11(c1)}$ | -0.011 | 0.141 | $\beta^h_{01(c1)}$ | 0.014 | 0.061 |
| $\beta^h_{11(c2)}$ | 0.132 | 0.354 | $\beta^h_{01(c2)}$ | -0.013 | 0.421 |
| $g$ function |  |  | $g$ function |  |  |
| $\beta^g_{11(0)}$ | -3.121*** | 0.136 | $\beta^g_{01(0)}$ | -1.472*** | 0.211 |
| $\beta^g_{11(r2)}$ | 4.397*** | 0.066 | $\beta^g_{01(r2)}$ | 3.159*** | 0.168 |
| # of obs. | 126 |  | # of obs. | 140 |  |

|  | $(d, M(d')) = (1, 0)$ |  |  | $(d, M(d')) = (0, 0)$ |  |
|---|---|---|---|---|---|
| $h$ function | est. | std.err. | $h$ function | est. | std.err. |
| $\beta^h_{10(0)}$ | -0.685 | 0.725 | $\beta^h_{00(0)}$ | -0.814* | 0.489 |
| $\beta^h_{10(r)}$ | -0.152 | 1.014 | $\beta^h_{00(r)}$ | -1.541** | 0.637 |
| $\beta^h_{10(e)}$ | -0.467 | 0.814 | $\beta^h_{00(e)}$ | 0.512 | 0.617 |
| $\beta^h_{10(p)}$ | 0.275 | 0.909 | $\beta^h_{00(p)}$ | -2.744*** | 0.216 |
| $\beta^h_{10(f)}$ | 0.219 | 0.411 | $\beta^h_{00(f)}$ | 0.449 | 0.352 |
| $\beta^h_{10(L)}$ | 1.576* | 0.924 | $\beta^h_{00(L)}$ | 0.098 | 0.630 |
| $\beta^h_{10(c1)}$ | -0.046 | 0.123 | $\beta^h_{00(c1)}$ | -0.001 | 0.104 |
| $\beta^h_{10(c2)}$ | 0.122 | 0.230 | $\beta^h_{00(c2)}$ | -0.025 | 0.333 |
| $g$ function |  |  | $g$ function |  |  |
| $\beta^g_{10(0)}$ | -2.739*** | 0.142 | $\beta^g_{00(0)}$ | -2.521*** | 0.124 |
| $\beta^g_{10(r2)}$ | 2.613*** | 0.085 | $\beta^g_{00(r2)}$ | 2.568*** | 0.087 |
| # of obs. | 126 |  | # of obs. | 140 |  |

Note: Significance: ***: 1% level, **: 5% level; *: 10% level. The standard errors are bootstrapped from 1,000 re-samples.

Table 5: Hypothesis Tests on Conditional Local Average Treatment Effects (CLATE)

| | output | frontier | inefficiency |
|---|---|---|---|
| program | $H_o^{\mathrm{p}} : CLATE(x) = 0, \forall x$ | $H_o^{\mathrm{p}h} : CLATE_h(x) = 0, \forall x$ | $H_o^{\mathrm{p}g} : CLATE_g(x) = 0, \forall x$ |
| evaluation | $\mathcal{W}$=2816.58 | $\mathcal{W}$=222.03 | $\mathcal{W}$=1921.23 |
| | $k = 30$ | $k = 24$ | $k = 6$ |
| | $p$-value=0.00 | $p$-value=0.00 | $p$-value=0.00 |
| mediation | $H_o^{\mathrm{dp}} : DCLATE(x) = 0, \forall x$ | $H_o^{\mathrm{d}h} : DCLATE_h(x) = 0, \forall x$ | $H_o^{\mathrm{d}g} : DCLATE_g(x) = 0, \forall x$ |
| analysis | $\mathcal{W}$=597.89 | $\mathcal{W}$=17.75 | $\mathcal{W}$=464.68 |
| | $k = 20$ | $k = 16$ | $k = 4$ |
| | $p$-value=0.00 | $p$-value=0.34 | $p$-value=0.00 |
| | $H_o^{\mathrm{i}p} : ICLATE(x) = 0, \forall x$ | $H_o^{\mathrm{i}h} : ICLATE_h(x) = 0, \forall x$ | $H_o^{\mathrm{i}g} : ICLATE_g(x) = 0, \forall x$ |
| | $\mathcal{W}$=215.68 | $\mathcal{W}$=193.80 | $\mathcal{W}$= 22.18 |
| | $k = 10$ | $k = 8$ | $k = 2$ |
| | $p$-value=0.00 | $p$-value=0.00 | $p$-value=0.00 |

The set of parameter restrictions for each of the tests is on page 15.

The $\mathcal{W}$ statistic has a $\chi^2$ distribution with $k$ degrees of freedom which is the number of parameter restrictions for the test.

Table 6: Estimated Local Average Treatment Effects (LATE)

| | output | frontier | inefficiency |
|---|---|---|---|
| program | $\tilde{H}_o^{\mathrm{p}}$ : LATE=0 | $\tilde{H}_o^{\mathrm{p}h}$ : LATE$_h$=0 | $\tilde{H}_o^{\mathrm{p}g}$ : LATE$_g$=0 |
| evaluation | -0.043 | -0.185 | -0.142*** |
| | (0.619) | (0.621) | (0.036) |
| mediation | $\tilde{H}_o^{\mathrm{d}}$ : DLATE=0 | $\tilde{H}_o^{\mathrm{d}h}$ : DLATE$_h$=0 | $\tilde{H}_o^{\mathrm{d}g}$ : DLATE$_g$=0 |
| analysis | -0.070 | -0.152 | -0.081** |
| | (0.723) | (0.722) | (0.040) |
| | $\tilde{H}_o^{\mathrm{i}}$ : ILATE=0 | $\tilde{H}_o^{\mathrm{i}h}$ : ILATE$_h$=0 | $\tilde{H}_o^{\mathrm{i}g}$ : ILATE$_g$=0 |
| | 0.028 | -0.033 | -0.061 |
| | (0.496) | (0.497) | (0.043) |

Note: The figures are the estimated effects and the standard errors (in parentheses). The significance level is based on the $t$-statistic version of $\mathcal{W}$. Significance: ***: 1% level, **: 5% level; *: 10% level. The standard errors are bootstrapped from 1,000 re-samples.

# References

[1] Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, **113**, 231-263.

[2] Aigner, D., Lovell, C.A.K. and Schmidt, P. (1977), "Formulation and Estimation of Stochastic Frontier Production Function Models," *Journal of Econometrics*, **6**, 21-37.

[3] Amsler, C., Prokhorov, A. and Schmidt, P. (2016), "Endogeneity in Stochastic Frontier Models," *Journal of Econometrics*, **190**, 280-288.

[4] Bampasidou, M., Carlos A. Flores, C. A., Flores-Lagunes, A. and Parisian, D. J. (2016), "The Role of Degree Attainment in the Differential Impact of Job Corps on Adolescents and Young Adults," *Research in Labor Economics*, **40**, 113-156.

[5] Bauder, T.A., Waskom, R.M., Pearson, R. (2010), "Best Management Practices For Agricultural Pesticide Use To Protect Water Quality," Colorado State University Extension.

[6] Benavente, J.M., Crespi, G. and Manoli, A. (2007), "Public Support to Firm Level Innovation: An Evaluation of the FONTEC Program," *OVE Working Papers, WP-05/07.*

[7] Caudill, S.B. and Ford, J.M. (1993), "Biases in Frontier Estimation Due to Heteroscedasticity," *Economics Letters*, **41**, 17-20.

[8] Chudnovsky D., Lopez, A., Rossi, M. and Ubfal, D. (2006), "Evaluating A Program of Public Funding of Private Innovation Activities. An Econometric Study of FONTAR in Argentina," *OVE Working Papers WP-16/06.*

[9] Crespo-Cebada, E., Pedraja-Chaparro, F. and Santín, D. (2014), "Does School Ownership Matter? An Unbiased Efficiency Comparison for Regions of Spain," *Journal of Productivity Analysis*, **41**, 153-172.

[10] Donald, S.G., Hsu, Y.-C. and Lieli, R.P. (2014a), "Testing the Unconfoundedness Assumption Via Inverse Probability Weighted Estimators of (L)Att," *Journal of Business & Economic Statistics*, **32**, 395-415.

[11] Donald, S.G., Hsu, Y.-C. and Lieli, R.P. (2014b), “Inverse Probability Weighted Estimation of Local Average Treatment Effects: A Higher Order MSE Expansion,” *Statistics and Probability Letters*, **95**, 132-138.

[12] Duflo, E. and Pande, R. (2007a), “Dams,” *The Quarterly Journal of Economics*, **122**, 601-646.

[13] Duflo, E. and Pande, R. (2007b), “Dams, Poverty, Public Goods and Malaria Incidence in India,” *http://hdl.handle.net/1902.1/IOJHHXOOLZ, Harvard Dataverse*, V4.

[14] Fisher, M.J. and Cook, S.E. (2013), “Water, Food and Poverty in River Basins: Defining the Limits,” Routledge.

[15] Flores, C. A. and Flores-Lagunes, A. (2009) “Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness,” *IZA Discussion Paper No. 4237.*

[16] Flores, C. A. and Flores-Lagunes, A. (2010) “Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects,” *Working Paper.*

[17] Fortin, N., Lemieux, T. and Firpo, S. (2011), “Decomposition Methods in Economics,” *NBER Working Paper No. 16045.*

[18] Frenken, K. and Faures, J. M. (1997), “Irrigation Potential in Africa: A Basin Approach” (Vol. 4). Food & Agriculture Org.

[19] Frölich, M. (2007), “Nonparametric IV Estimation of Local Average Treatment Effects with Covariates,” *Journal of Econometrics*, **139**, 35-75.

[20] Frölich, M. and Huber, M. (2014), "Direct and Indirect Treatment Effects: Causal Chains and Mediation Analysis with Instrumental Variables," *CEMAP Working Paper CWP31/14*.

[21] Frölich, M. and Huber, M. (2017), "Direct and Indirect Treatment Effects: Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society Series B*, forthcoming.

[22] Glass, A.J., Kenjegalieva, K. and Sickles, R.C. (2016), "A Spatial Autoregressive Stochastic Frontier Model for Panel Data with Asymmetric Efficiency Spillovers," *Journal of Econometrics*, **190**, 289-300.

[23] Greene, W. (2005), "Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model," *Journal of Econometrics*, **126**, 269-303.

[24] Griffiths, W.E. and Hajargasht, G. (2016), "Some Models for Stochastic Frontiers with Endogeneity," *Journal of Econometrics*, **190**, 341-8.

[25] Hong, H. and Nekipelov, D. (2010), "Semiparametric Efficiency in Nonlinear LATE Models," *Quantitative Economics*, **1**, 279-304.

[26] Huber, M. (2014), "Identifying Causal Mechanisms (Primarily) Based on Inverse Probability Weighting," *Journal of Applied Econometrics*, **29**, 920-943.

[27] Huber, M., Lechner, M. and Mellace, G., (2016), "The Finite Sample Performance of Estimators for Mediation Analysis Under Sequential Conditional Independence," *Journal of Business and Economic Statistics*, **34**, 139-160.

[28] Imbens, G. W. (2010), "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, **48**, 399-423.

[29] Imbens, G. W. and Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, **62**, 467-475.

[30] Kleit, A. N. and Terrell, D. (2001), "Measuring Potential Efficiency Gains from Deregulation of Electricity Generation: A Bayesian Approach," *The Review of Economics and Statistics*, **83**, 523-530.

[31] Lopez-Acevedo, G. and Tinajero-Bravo, M. (2011), "Evaluating Enterprise Support Programs Using Panel Firm Data," *6th IZA/World Bank Conference: Employment and Development.*

[32] MacKinnon, D. P. (2008), "Introduction to Statistical Mediation Analysis." Taylor & Francis: New York.

[33] McCully, P. (2001), "Silenced Rivers: The Ecology and Politics of Large Dams." Zed Books: London.

[34] Schochet, P. Z., Burghardt, J. and McConnell, S. (2008), "Does Job Corps Work? Impact Findings from the National Job Corps Study," *American Economic Review,* **98**, 1864-1886.

[35] Singh, S. (2002), "Taming the Waters: The Political Economy of Large Dams in India." Oxford University Press: New Delhi.

[36] Wang, H.-J., Chang, C.-C. and Chen, P.-C. (2008), "The Cost Effects of Government-Subsidised Credit: Evidence from Farmers' Credit Unions in Taiwan," *Journal of Agricultural Economics,* **59**, 132-149.

[37] Wooldridge, J. M. (2010), "Econometric Analysis of Cross Section and Panel Data," The MIT Press.

# Supplementary Appendix of

# "A Stochastic Frontier Model with an Endogenous Treatment Status and a Mediator"

**Yi-Ting Chen**

Institute of Economics

Academia Sinica

**Yu-Chin Hsu**

Institute of Economics

Academia Sinica

**Hung-Jen Wang**

Department of Economics

National Taiwan University, and

Institute of Economics

Academia Sinica

This version: March 4, 2017

# 1 Introduction

This appendix is a supplement to the paper "A Stochastic Frontier Model with an Endogenous Treatment Status and a Mediator." It includes seven sections that are not presented in the paper for the sake of maintaining brevity. In Section 2, we discuss the asymptotics of the estimators proposed in the paper and the Wald test for the CLATE hypotheses and the LATE hypotheses, and provide a bootstrap method for statistical inference. In Section 3, we present a Monte Carlo simulation. In Section 4, we derive the formula of $\Psi_{d'}(Z_2, X, \eta)$, which is shown in (5.12), for the bivariate probit model. In Section 5, we provide the proofs of Theorem 3.1 and Theorem 3.2 of the paper. In Section 6, we show that Assumption 3.3(e) holds for the dam example. In Section 7, we report the estimates of auxiliary parameters in the dam example.

# 2 Asymptotics and Bootstrap

In Section 4 of the paper, we demonstrate that the first-step estimators: $\hat{\eta}$, $\hat{\alpha}_f$ and $\hat{\alpha}_{z_1}$ are respectively the MLEs for $\eta$, $\alpha_f$ and $\alpha_{z_1}$. More specifically, let $\mathcal{E}$, $\mathcal{A}_f$ and $\mathcal{A}_{z_1}$ be, respectively, the parameter spaces of $\eta$, $\alpha_f$ and $\alpha_{z_1}$, and $e$, $a_{z_1}$ and $a_f$ be, respectively, arbitrary vectors in $\mathcal{E}$, $\mathcal{A}_f$ and $\mathcal{A}_{z_1}$. These MLEs are defined as:

$$\hat{\eta} = \arg\max_{e\in\mathcal{E}} \frac{1}{n}\sum_{i=1}^{n} \ell_{\eta,i}(e),$$

$$\hat{\alpha}_f = \arg\max_{a_f\in\mathcal{A}_f} \frac{1}{n}\sum_{i=1}^{n} \ell_{f,i}(a_f)$$

and

$$\hat{\alpha}_{z_1} = \arg\max_{a_{z_1}\in\mathcal{A}_{z_1}} \frac{1}{n}\sum_{i=1}^{n} \ell_{z_1,i}(a_{z_1}),$$

based on the log-probability or log-density functions:

$$\ell_{\eta,i}(e) \equiv \log P(M_i, D_i | Z_{1i}, Z_{2i}, X_i, e),$$

$$\ell_{f,i}(a_f) \equiv \log f_{Z_2}(Z_{2i}, a_f)$$

and

$$\ell_{z_1,i}(a_{z_1}) \equiv \log \left( Q_{Z_1}(X_i^\top a_{z_1})^{Z_{1i}} (1 - Q_{Z_1}(X_i^\top a_{z_1}))^{1-Z_{1i}} \right).$$

It is standard to argue for the consistency and the asymptotic normality of these estimators by the ML theory under our assumptions and suitable regularity conditions. Since the ML theory is well-known, we do not repeat the regularity conditions for the consistency and asymptotic normality of the MLE; see, e.g., White (1994) and Newey and McFadden (1994) for the ML theory and conditions. Given the consistency of $\hat{\eta}$, $\hat{\alpha}_f$ and $\hat{\alpha}_{z_1}$, by taking the mean-value expansions of the estimation equations of these estimators under suitable conditions and using the information matrix equality, we can show that $\hat{\eta}$, $\hat{\alpha}_f$ and $\hat{\alpha}_{z_1}$ have the following influence-function representations:

$$\sqrt{n}(\hat{\eta} - \eta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\eta,i}(\eta) + o_p(1), \tag{A.1}$$

$$\sqrt{n}(\hat{\alpha}_f - \alpha_f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{f,i}(\alpha_f) + o_p(1) \tag{A.2}$$

and

$$\sqrt{n}(\hat{\alpha}_{z_1} - \alpha_{z_1}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{z_1,i}(\alpha_{z_1}) + o_p(1), \tag{A.3}$$

where $\psi_{\eta,i}(\eta)$, $\psi_{f,i}(\alpha_f)$ and $\psi_{z_1,i}(\alpha_{z_1})$ are instantaneous transformations of $\boldsymbol{W}_i$ such that

$$\psi_{\eta,i}(\eta) = E[\nabla_e \ell_{\eta,i}(\eta) \nabla_{e^\top} \ell_{\eta,i}(\eta)]^{-1} \nabla_e \ell_{\eta,i}(\eta),$$

$$\psi_{f,i}(\alpha_f) = E[\nabla_{a_f}\ell_{f,i}(\alpha_f)\nabla_{a_f^\top}\ell_{f,i}(\alpha_f)]^{-1}\nabla_{a_f}\ell_{f,i}(\alpha_f)$$

and

$$\psi_{z_1,i}(\alpha_{z_1}) = E[\nabla_{a_{z_1}}\ell_{z_1,i}(\alpha_{z_1})\nabla_{a_{z_1}^\top}\ell_{z_1,i}(\alpha_{z_1})]^{-1}\nabla_{a_{z_1}}\ell_{z_1,i}(\alpha_{z_1}).$$

Note that in (A.2), the influence function $\psi_{f,i}(\alpha_f)$ is defined in the case where $Z_2$ is free of estimation. In the case where $Z_2$ is dependent on a vector of estimators, as in the dam example, (A.2) is still valid but $\psi_{f,i}(\alpha_f)$ needs to be redefined by replacing $\nabla_{a_f}\ell_{f,i}(\alpha_f)$ with its mean-value expansion with respect to the parameter vector underlying $Z_2$. This redefinition of $\psi_{f,i}(\alpha_f)$ would not influence the following discussions. From (A.1), (A.2) and (A.3), we can also write that

$$\sqrt{n}(\hat{\alpha} - \alpha) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\alpha,i}(\alpha) + o_p(1), \tag{A.4}$$

where

$$\psi_{\alpha,i}(\alpha) \equiv (\psi_{\eta,i}(\eta)^\top, \psi_{f,i}(\alpha_f)^\top, \psi_{z_1,i}(\alpha_{z_1})^\top)^\top. \tag{A.5}$$

Similarly, because the estimator $\hat{\beta}_d$ can be interpreted as a second-step moment estimator for $\beta$, it is also standard to argue for its consistency and asymptotic normality under suitable regularity conditions based on the two-stage estimation theory. This estimation theory is also quite standard, so the regularity conditions for the consistency and asymptotic normality of the two-stage estimator are not reported here to maintain brevity; see, e.g., Newey and McFadden (1994) and Wooldridge (2010) for the theory and conditions. Given the consistency of $\hat{\beta}_d$, by taking the mean-value expansions of the estimation equation in (4.3) under suitable conditions and using (A.4), we can show that $\hat{\beta}_d$ has the influence-function representation:

$$\sqrt{n}(\hat{\beta}_d - \beta_d) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\beta_d,i}(\alpha) + o_p(1), \tag{A.6}$$

where

$$\psi_{\beta_d,i}(\beta_d) \equiv -E\left[\nabla_{b^2}\varrho_{d,i}(\alpha,\beta_d)\right]^{-1}\left(\nabla_{b_d}\varrho_{d,i}(\alpha,\beta_d) + E\left[\nabla_{a^\top}\left(\nabla_{b_d}\varrho_{d,i}(\alpha,\beta_d)\right)\right]\psi_{\alpha,i}(\alpha)\right),$$

with $a \equiv (e^\top, a_f^\top, a_{z_1}^\top)^\top$, for $d \in \{0,1\}$.

Denote $\beta \equiv (\beta_0^\top, \beta_1^\top)^\top$, $\hat{\beta} \equiv (\hat{\beta}_0^\top, \hat{\beta}_1^\top)^\top$, $\theta \equiv (\alpha^\top, \beta^\top)^\top$, $\hat{\theta} \equiv (\hat{\alpha}^\top, \hat{\beta}^\top)^\top$ and $\psi_{\beta,i}(\beta) \equiv (\psi_{\beta_0,i}(\beta_0)^\top, \psi_{\beta_1,i}(\beta_1)^\top)^\top$. From (A.4) and (A.6), we can see that

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\theta) + o_p(1), \quad \text{with} \quad \psi_i(\theta) \equiv \begin{bmatrix} \psi_{\alpha,i}(\alpha) \\ \psi_{\beta,i}(\beta) \end{bmatrix}. \tag{A.7}$$

Because $\psi_i(\theta)$ is an instantaneous transformation of $\boldsymbol{W}_i$, $\{\psi_i(\theta)\}_{i=1}^n$ is an independently and identically distributed (IID) sequence if $\{\boldsymbol{W}_i\}_{i=1}^n$ is an IID sequence. Given this IIDness and the condition that the elements of $\psi_i(\theta)$ have finite second moments, we may use a central limit theorem and the Cramër-Wold device to demonstrate that $\hat{\theta}$ has the asymptotic normality:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\text{d}} N(0, \Sigma), \tag{A.8}$$

with the asymptotic covariance matrix $\Sigma \equiv E\left[\psi_i(\theta)\psi_i(\theta)^\top\right]$ because of (A.7).

The result in (A.8) is applicable to establishing a generalized test for the CLATE hypotheses: $H_o^{\text{p}}$, $H_o^{\text{p}h}$, $H_o^{\text{p}g}$,$H_o^d$, $H_o^{\text{d}h}$, $H_o^{\text{p}g}$, $H_o^i$, $H_o^{\text{i}h}$ and $H_o^{\text{i}g}$ discussed in Section 2.2. For ease of exposition, suppose that $\alpha_d$ is known or assumed to be non-zero, so $H_o^i$, $H_o^{\text{i}h}$ and $H_o^{\text{i}g}$ only include the restrictions on $\beta_{01}^h$, $\beta_{00}^h$, $\beta_{01}^g$ and $\beta_{00}^g$. Let $\Sigma_\beta$ be the asymptotic covariance matrix of $n^{1/2}(\hat{\beta} - \beta)$ implied by (A.8), and $\delta$ be a finite-dimensional linear transformation of $\beta$ such that $\delta = S_\delta\beta$ for some $S_\delta$ which is a deterministic matrix of full row rank and is dependent on the choice of $\delta$. These hypotheses are particular examples of the hypothesis: $\delta = 0$ with different $S_\delta$. Denote $\hat{\delta} = S_\delta\hat{\beta}$. According to (A.8), we have the result:

$$\sqrt{n}(\hat{\delta} - \delta) \xrightarrow{\text{d}} N(0, \Sigma_\delta), \tag{A.9}$$

with $\Sigma_\delta = S_\delta \Sigma_\beta S_\delta^\top$. Let $\hat{\Sigma}_\delta$ be a positive-definite matrix which is consistent for $\Sigma_\delta$. Given (A.9), we can establish a Wald test statistic:

$$\mathcal{W} = n\hat{\delta}^\top \hat{\Sigma}_\delta^{-1} \hat{\delta} \tag{A.10}$$

that has the asymptotic null distribution: $\mathcal{W} \xrightarrow{\text{d}} \chi^2(k)$, with $\hat{\Sigma}_\delta$ denoting a consistent estimator for the asymptotic covariance matrix of $n^{1/2}(\hat{\delta} - \delta)$ and $k \equiv \dim(\delta)$. This generalized test is applicable to the CLATE hypotheses by matching $\delta$ with the $\beta$ restrictions of the hypothesis being tested.

By a slight modification, this test is also applicable to examining the LATE hypotheses: $\tilde{H}_o^{\text{p}}$, $\tilde{H}_o^{\text{p}h}$, $\tilde{H}_o^{\text{p}g}$,$\tilde{H}_o^{\text{d}}$, $\tilde{H}_o^{\text{d}h}$, $\tilde{H}_o^{\text{d}g}$, $\tilde{H}_o^{\text{i}}$, $\tilde{H}_o^{\text{i}h}$ and $\tilde{H}_o^{\text{i}g}$ discussed in Section 2.3. In this scenario, we need to replace $\delta$ and $\hat{\delta}$ by $\delta = E[\xi_i(\theta)]$ and $\hat{\delta} = n^{-1}\sum_{i=1}^n \xi_i(\hat{\theta})$, where $\xi_i(\cdot)$ is dependent on the hypothesis being tested. For example, we can set $\delta = \mathcal{H}_{11}(\nu_1^h) - \mathcal{H}_{01}(\nu_0^h)$, $\hat{\delta} = \hat{\mathcal{H}}_{11}(\hat{\nu}_1^h) - \hat{\mathcal{H}}_{01}(\hat{\nu}_0^h)$ and

$$\xi_i(\theta) = \left(\frac{Q_D(\gamma_{z1} + X_i^\top \gamma_x) - Q_D(X_i^\top \gamma_x)}{E[Q_D(\gamma_{z1} + X_i^\top \gamma_x] - E[Q_D(X_i^\top \gamma_x)]}\right)\left(h_1(X_i, \alpha_m, \beta_{11}^h, \beta_{10}^h) - h_1(X_i, \alpha_m, \beta_{01}^h, \beta_{00}^h)\right)$$

when the hypothesis being tested is $\tilde{H}_o^{\text{d}h}$. Under suitable conditions, we may also use the mean-value expansion to show the influence-function representation of $\hat{\delta}$:

$$\sqrt{n}(\hat{\delta} - \delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\delta,i}(\theta) + o_p(1), \tag{A.11}$$

where

$$\psi_{\delta,i}(\theta) \equiv \xi_i(\theta) + E[\nabla_{\theta^\top}\xi_i(\theta)]\psi_i(\theta).$$

Similar to the derivation of (A.8), we may use (A.11) and the fact that $\psi_{\delta,i}(\theta)$ is an instantaneous transformation of $\boldsymbol{W}_i$ to show that, under certain regularity conditions, (A.9) holds for $\Sigma_\delta = E[\psi_{\delta,i}(\theta)^2]$ with $k = 1$. By this modification, the generalized test in (A.10) becomes applicable to testing the LATE hypotheses.

In general, the formula of $\Sigma_\delta$ is quite complicated. Thus, it is useful to consider $\hat{\Sigma}_\delta$ to be a bootstrap estimator, rather than a plug-in estimator, for $\Sigma_\delta$ from the viewpoint of practitioners. In this study, we use the following procedure to compute the bootstrap estimator $\hat{\Sigma}_\delta$:

1. Generate a bootstrap sample: $\{\boldsymbol{W}^*_{i(j)}\}_{i=1}^n$ by randomly drawing $\{\boldsymbol{W}_i\}_{i=1}^n$ with replacement, where $j = 1, 2, \ldots, B$, and $B$ denotes the number of replications.

2. Generate the bootstrap estimator vector: $\hat{\theta}^*_j$ by using $\{\boldsymbol{W}^*_{i(j)}\}_{i=1}^n$ in place of the role of $\{\boldsymbol{W}_i\}_{i=1}^n$ in the computation of $\hat{\theta}$.

3. Generate the bootstrap statistic: $\hat{\delta}^*_j$ by using $\{\boldsymbol{W}^*_{i(j)}\}_{i=1}^n$ and $\hat{\theta}^*_j$ in place of the role of $\{\boldsymbol{W}_i\}_{i=1}^n$ and $\hat{\theta}$ in the computation of $\hat{\delta}$.

4. Generate the sample of bootstrap statistics $\{\hat{\delta}^*_j\}_{j=1}^B$ by repeating the above two steps for $j = 1, 2, \ldots, B$.

5. Compute the bootstrap asymptotic covariance matrix estimator for $\Sigma_\delta$:

$$\hat{\Sigma}_\delta = \frac{1}{B}\sum_{j=1}^{B}\left(\sqrt{n}\Big(\hat{\delta}^*_j - \frac{1}{B}\sum_{j=1}^{B}\hat{\delta}^*_j\Big)\right)\left(\sqrt{n}\Big(\hat{\delta}^*_j - \frac{1}{B}\sum_{j=1}^{B}\hat{\delta}^*_j\Big)\right)^\top .$$

By this bootstrap method, we can implement the CLATE tests and the LATE tests in a simple way.

# 3 Monte Carlo Simulation

Our Monte Carlo simulation is based on the following data generating process (DGP) for $\boldsymbol{W}$:

- $X = (1, X_1, X_2)^\top$, $Z_1 = I(U_{Z_1} \geq 0)$ and $(U_{Z_1}, Z_2, X_1, X_2)^\top \sim N(0, I_4)$; $I_4$ stands for the $4 \times 4$ identity matrix.

- $(\tilde{U}_Y, U_M, U_D)^\top$ is independent of $(U_{Z_1}, Z_2, X_1, X_2)^\top$, and has the distribution:

$$\begin{bmatrix} \tilde{U}_Y \\ U_M \\ U_D \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{ym} & \rho_{yd} \\ \rho_{ym} & 1 & 0 \\ \rho_{yd} & 0 & 1 \end{bmatrix} \right); \tag{A.16}$$

$$\begin{aligned} U_Y &= \sigma_y \left( \tilde{U}_Y - E[\tilde{U}_Y | X, \mathcal{C}] \right) \\ &= \sigma_y \left( \tilde{U}_Y - \rho_{yd} \left( \frac{\phi(\gamma_z + X^\top \gamma_x) - \phi(X^\top \gamma_x)}{\Phi(\gamma_{z1} + X^\top \gamma_x) - \Phi(X^\top \gamma_x)} \right) \right), \end{aligned} \tag{A.17}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the PDF and the CDF of $N(0,1)$. The derivation of $E[\tilde{U}_Y | X, \mathcal{C}]$ is provided at the end of this section.

- $(Y, M, D)$ is generated from the process:

$$\begin{aligned} Y &= \tilde{h}(D, M, X) - \tilde{g}(D, M, X) + U_Y, \\ M &= 1(\alpha_d D + \alpha_{z_2} Z_2 + U_M \geq 0), \\ D &= 1(\gamma_{z1} Z_1 + \gamma_{x_1} X_1 + U_D \geq 0), \end{aligned} \tag{A.18}$$

where

$$\tilde{h}(D, M, X) = h(X, \beta^h_{dj}) = \beta^h_{dj(0)} + \beta^h_{dj(1)} X_1$$

and

$$\tilde{g}(D, M, X) = g(X, \beta^g_{dj}) = \sqrt{\frac{2}{\pi}} \exp(\beta^g_{dj(2)} X_2), \tag{A.19}$$

if $(D, M) = (d, j)$, for $d, j \in \{0, 1\}$.

This DGP is encompassed by the parametric structural model for $(Y, M, D)$ in Assumptions 3.2. Given (A.16) and (A.18), $M$ and $D$ are endogenous when $\rho_{ym}$ and $\rho_{yd}$ are non-zero; meanwhile, the $U_Y$ in (A.17) amounts to $U_Y = \sigma_y(\tilde{U}_Y - E[\tilde{U}_Y|X, \mathcal{C}])$ and hence satisfies the condition: $E[U_Y|X, \mathcal{C}] = 0$ in Assumptions 3.2($b$). The parameter $\sigma_y$ in (A.17) is introduced to investigate how the degrees of data noise may affect the performance of our method. The functional form of $g(X, \beta^g_{dj})$ in (A.19) is non-negative. It is motivated by a popular specification for the production inefficiency term in the traditional context (2.1), in which $u|X$ has a conditional half-normal distribution so that its conditional mean is $g(X, \beta^g) = \sqrt{2/\pi}\exp(\beta^{g\top}X)$.

In the simulation, we consider the following parameter settings for this DGP:

- $\rho_{ym} = \rho_{yd} = \rho = 0.1$ and 0.7;
- $\sigma_y$ =0.1, 0.5, and 1;
- $(\alpha_d, \alpha_{z_2}, \gamma_{z_1}, \gamma_{x_1}) = (0.1, 1, 1, 0.1)$ and

$$(\beta^h_{dj(0)}, \beta^h_{dj(1)}, \beta^g_{dj(2)}) = \begin{cases} (0.3, 0.3, 0.15), & \text{if } (d, j) = (1, 1), \\ (0.6, 0.6, 0.30), & \text{if } (d, j) = (1, 0), \\ (0.9, 0.9, 0.45), & \text{if } (d, j) = (0, 1), \\ (1.2, 1.2, 0.60), & \text{if } (d, j) = (0, 0). \end{cases}$$

The correlation coefficient $\rho$ measures the strength of endogeneity, and the standard deviation $\sigma_y$ measures the degrees of data noise. The weight function $m_{d'}(\cdot)$ in (3.10) is computed numerically using the Gauss-Hermite quadrature with 100 quadrature points.

We let the model for $(Y, M, D)$ be correctly specified for (A.18). For simplicity, we focus on the sub-model for $Y$, and estimate $\beta_d = (\beta^h_{d1(0)}, \beta^h_{d1(1)}, \beta^h_{d0(0)}, \beta^h_{d0(1)}, \beta^g_{d1(2)}, \beta^g_{d0(2)})^\top$, for $d \in \{0, 1\}$, by the WNLSE in (4.4) with the replacement of $\hat{\alpha}$ by $\alpha$. The sample size is $n$ = 250, 1,000 or 10,000, and the number of replications is 1,000. This simulation design allows us to investigate the estimator's performance in samples of different sample sizes and degrees of endogeneity and data noise by using different $(n, \rho, \sigma_y)$. Our empirical example would most likely resemble the cases with $n = 250$ and $\sigma_y = 0.1$. The actual sample size of the empirical example is 266, and by choosing $\sigma_y = 0.1$, we are able to closely match the $\mathrm{R}^2$ statistics from an OLS regression of $Y$ on $X$. We report the sample mean, bias, and the MSE of the estimates and they are calculated from the 1,000 replications. The results are reported in Table A.1 to Table A.3.

Table A.1 presents the simulation results with $\sigma_y = 0.1$, with the upper panel reporting cases of $\rho = 0.1$ (weak endogeneity) and the lower panel $\rho = 0.7$ (strong endogeneity). Given the small value of $\sigma_y$, the models are in general easy to estimate, and the bias and the MSE are small even for small samples ($n = 250$). The largest MSE in this case is 0.038. We notice that $\hat{\beta}^g_{dj}$ tends to have larger bias compared to $\hat{\beta}^h_{dj}$, which may be due to the non-linearity of the $g(\cdot)$ function. On the other hand, the size of the MSE appears to come mainly from the variance of the estimate, instead of the bias. This is indicated by the fact that the MSE equals the sum of the variance and the square of the bias.

Different values of $\rho$ do not appear to exert strong impacts on the estimation, as can be seen by comparing the upper and the lower panels of the table. Although the MSE appears to be somewhat larger with a larger degree of endogeneity of the model ($\rho = 0.7$), the margin of difference is small.

Tables A.2 and A.3 present the simulation results with $\sigma_y = 0.5$ and 1, respectively. Larger values of $\sigma_y$ make the data more noisy, which has adverse effects on the non-linear model's estimation. For small samples ($n = 250$), the bias and the MSE from the cases where $\sigma_y = 1$ are about 10 to 15 times larger than those in the cases of $\sigma_y = 0.1$. A back-of-the-envelope calculation indicates that the increase in the MSE mainly comes from increases in the variance of the estimate, which is not surprising. The increases in bias and the MSE, nevertheless, quickly subside as the sample size becomes bigger. With $n = 10,000$, both the bias and the MSE are numerically not much different in the cases where $\sigma_y = 0.1$ and $\sigma_y = 1$.

In Tables A.2 and A.3, the bias and the MSE are reduced as the sample size increases, which is evidence of the estimator's consistency. Although there are more cases in Table A.1 where the reductions in the bias and the MSE are not obvious, we note that the two statistics are quite small in this table to begin with, and sampling errors may be important in explaining the differences in these cases.

Table A.1: Simulation Results for $\sigma_y = 0.1$

| $\rho = 0.1$ | $n = 250$ | | | $n = 1,000$ | | | $n = 10,000$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma_y = 0.1$ | mean | bias | MSE | mean | bias | MSE | mean | bias | MSE |
| $\hat{\beta}^h_{11(0)}$ | 0.299 | -0.001 | 0.022 | 0.300 | -0.0004 | 0.002 | 0.300 | 0.0001 | 0.0001 |
| $\hat{\beta}^h_{11(1)}$ | 0.306 | 0.006 | 0.038 | 0.298 | -0.002 | 0.004 | 0.300 | 0.0002 | 0.0003 |
| $\hat{\beta}^g_{11(2)}$ | 0.149 | -0.001 | 0.012 | 0.149 | -0.001 | 0.002 | 0.150 | -0.0003 | 0.0001 |
| $\hat{\beta}^h_{10(0)}$ | 0.599 | -0.001 | 0.020 | 0.600 | -0.0003 | 0.002 | 0.600 | 0.0001 | 0.0001 |
| $\hat{\beta}^h_{10(1)}$ | 0.606 | 0.006 | 0.034 | 0.599 | -0.001 | 0.004 | 0.600 | 0.0001 | 0.0002 |
| $\hat{\beta}^g_{10(2)}$ | 0.275 | -0.025 | 0.015 | 0.296 | -0.004 | 0.002 | 0.300 | -0.0005 | 0.0001 |
| $\hat{\beta}^h_{01(0)}$ | 0.901 | 0.001 | 0.007 | 0.895 | -0.005 | 0.001 | 0.896 | -0.004 | 0.0001 |
| $\hat{\beta}^h_{01(1)}$ | 0.898 | -0.002 | 0.025 | 0.895 | -0.005 | 0.001 | 0.896 | -0.004 | 0.0001 |
| $\hat{\beta}^g_{01(2)}$ | 0.452 | 0.002 | 0.010 | 0.461 | 0.011 | 0.003 | 0.451 | 0.001 | 0.0003 |
| $\hat{\beta}^h_{00(0)}$ | 1.201 | 0.001 | 0.007 | 1.195 | -0.005 | 0.001 | 1.196 | -0.004 | 0.0001 |
| $\hat{\beta}^h_{00(1)}$ | 1.198 | -0.002 | 0.025 | 1.195 | -0.005 | 0.001 | 1.196 | -0.004 | 0.0001 |
| $\hat{\beta}^g_{00(2)}$ | 0.555 | -0.045 | 0.019 | 0.579 | -0.021 | 0.005 | 0.596 | -0.004 | 0.0003 |
| $\rho = 0.7$ | $n = 250$ | | | $n = 1,000$ | | | $n = 10,000$ | | |
| $\sigma_y = 0.1$ | mean | bias | MSE | mean | bias | MSE | mean | bias | MSE |
| $\hat{\beta}^h_{11(0)}$ | 0.300 | 0 | 0.024 | 0.296 | -0.004 | 0.002 | 0.299 | -0.001 | 0.0001 |
| $\hat{\beta}^h_{11(1)}$ | 0.305 | 0.005 | 0.044 | 0.299 | -0.001 | 0.003 | 0.299 | -0.001 | 0.0003 |
| $\hat{\beta}^g_{11(2)}$ | 0.155 | 0.005 | 0.014 | 0.152 | 0.002 | 0.002 | 0.150 | 0.0002 | 0.0001 |
| $\hat{\beta}^h_{10(0)}$ | 0.601 | 0.001 | 0.022 | 0.596 | -0.004 | 0.002 | 0.599 | -0.001 | 0.0001 |
| $\hat{\beta}^h_{10(1)}$ | 0.605 | 0.005 | 0.040 | 0.599 | -0.001 | 0.003 | 0.600 | -0.001 | 0.0002 |
| $\hat{\beta}^g_{10(2)}$ | 0.282 | -0.018 | 0.015 | 0.299 | -0.001 | 0.002 | 0.300 | 0.0001 | 0.0001 |
| $\hat{\beta}^h_{01(0)}$ | 0.897 | -0.003 | 0.008 | 0.897 | -0.003 | 0.001 | 0.896 | -0.004 | 0.0001 |
| $\hat{\beta}^h_{01(1)}$ | 0.894 | -0.006 | 0.020 | 0.894 | -0.006 | 0.001 | 0.895 | -0.005 | 0.0001 |
| $\hat{\beta}^g_{01(2)}$ | 0.458 | 0.008 | 0.011 | 0.462 | 0.012 | 0.003 | 0.452 | 0.002 | 0.0003 |
| $\hat{\beta}^h_{00(0)}$ | 1.197 | -0.003 | 0.008 | 1.197 | -0.003 | 0.001 | 1.196 | -0.004 | 0.0001 |
| $\hat{\beta}^h_{00(1)}$ | 1.194 | -0.006 | 0.020 | 1.193 | -0.007 | 0.001 | 1.195 | -0.005 | 0.0001 |
| $\hat{\beta}^g_{00(2)}$ | 0.558 | -0.042 | 0.019 | 0.581 | -0.019 | 0.005 | 0.596 | -0.004 | 0.0003 |

True values are: $(\beta^h_{11(0)}, \beta^h_{11(1)}, \beta^g_{11(2)}) = (0.3, 0.3, 0.15)$, $(\beta^h_{10(0)}, \beta^h_{10(1)}, \beta^g_{10(2)}) = (0.6, 0.6, 0.30)$, $(\beta^h_{01(0)}, \beta^h_{01(1)}, \beta^g_{01(2)}) = (0.9, 0.9, 0.45)$, $(\beta^h_{00(0)}, \beta^h_{00(1)}, \beta^g_{00(2)}) = (1.2, 1.2, 0.60)$.

Table A.2: Simulation Results for $\sigma_y = 0.5$

| $\rho = 0.1$ | $n = 250$ | | | $n = 1,000$ | | | $n = 10,000$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma_y = 0.5$ | mean | bias | MSE | mean | bias | MSE | mean | bias | MSE |
| $\hat{\beta}^h_{11(0)}$ | 0.317 | 0.017 | 0.111 | 0.296 | -0.004 | 0.010 | 0.300 | 0.001 | 0.001 |
| $\hat{\beta}^h_{11(1)}$ | 0.319 | 0.019 | 0.155 | 0.298 | -0.002 | 0.017 | 0.300 | -0.001 | 0.001 |
| $\hat{\beta}^g_{11(2)}$ | 0.152 | 0.002 | 0.043 | 0.151 | 0.001 | 0.014 | 0.149 | -0.001 | 0.001 |
| $\hat{\beta}^h_{10(0)}$ | 0.616 | 0.016 | 0.098 | 0.596 | -0.004 | 0.009 | 0.600 | -0.001 | 0.001 |
| $\hat{\beta}^h_{10(1)}$ | 0.617 | 0.017 | 0.140 | 0.598 | -0.002 | 0.015 | 0.600 | -0.001 | 0.001 |
| $\hat{\beta}^g_{10(2)}$ | 0.226 | -0.074 | 0.051 | 0.281 | -0.019 | 0.014 | 0.298 | -0.002 | 0.001 |
| $\hat{\beta}^h_{01(0)}$ | 0.906 | 0.006 | 0.030 | 0.897 | -0.003 | 0.004 | 0.897 | -0.003 | 0.0004 |
| $\hat{\beta}^h_{01(1)}$ | 0.913 | 0.013 | 0.078 | 0.892 | -0.008 | 0.005 | 0.896 | -0.004 | 0.0004 |
| $\hat{\beta}^g_{01(2)}$ | 0.463 | 0.013 | 0.030 | 0.469 | 0.019 | 0.006 | 0.452 | 0.002 | 0.0004 |
| $\hat{\beta}^h_{00(0)}$ | 1.208 | 0.008 | 0.031 | 1.197 | -0.003 | 0.004 | 1.197 | -0.003 | 0.0004 |
| $\hat{\beta}^h_{00(1)}$ | 1.215 | 0.015 | 0.076 | 1.191 | -0.009 | 0.005 | 1.196 | -0.004 | 0.0004 |
| $\hat{\beta}^g_{00(2)}$ | 0.529 | -0.071 | 0.042 | 0.569 | -0.031 | 0.009 | 0.595 | -0.005 | 0.001 |

| $\rho = 0.7$ | $n = 250$ | | | $n = 1,000$ | | | $n = 10,000$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma_y = 0.5$ | mean | bias | MSE | mean | bias | MSE | mean | bias | MSE |
| $\hat{\beta}^h_{11(0)}$ | 0.298 | -0.002 | 0.157 | 0.275 | -0.025 | 0.012 | 0.297 | -0.003 | 0.001 |
| $\hat{\beta}^h_{11(1)}$ | 0.329 | 0.029 | 0.255 | 0.291 | -0.009 | 0.015 | 0.299 | -0.001 | 0.001 |
| $\hat{\beta}^g_{11(2)}$ | 0.146 | -0.004 | 0.044 | 0.155 | 0.005 | 0.016 | 0.151 | 0.001 | 0.001 |
| $\hat{\beta}^h_{10(0)}$ | 0.599 | -0.001 | 0.142 | 0.577 | -0.023 | 0.011 | 0.598 | -0.002 | 0.001 |
| $\hat{\beta}^h_{10(1)}$ | 0.629 | 0.029 | 0.235 | 0.592 | -0.008 | 0.014 | 0.599 | -0.001 | 0.001 |
| $\hat{\beta}^g_{10(2)}$ | 0.217 | -0.083 | 0.056 | 0.278 | -0.022 | 0.016 | 0.300 | -0.0003 | 0.001 |
| $\hat{\beta}^h_{01(0)}$ | 0.920 | 0.020 | 0.030 | 0.902 | 0.002 | 0.004 | 0.897 | -0.003 | 0.0003 |
| $\hat{\beta}^h_{01(1)}$ | 0.890 | -0.010 | 0.057 | 0.888 | -0.012 | 0.006 | 0.895 | -0.005 | 0.0003 |
| $\hat{\beta}^g_{01(2)}$ | 0.474 | 0.024 | 0.022 | 0.468 | 0.018 | 0.006 | 0.452 | 0.002 | 0.0004 |
| $\hat{\beta}^h_{00(0)}$ | 1.222 | 0.022 | 0.030 | 1.202 | 0.002 | 0.004 | 1.197 | -0.003 | 0.0003 |
| $\hat{\beta}^h_{00(1)}$ | 1.191 | -0.009 | 0.057 | 1.190 | -0.010 | 0.006 | 1.195 | -0.005 | 0.0003 |
| $\hat{\beta}^g_{00(2)}$ | 0.540 | -0.060 | 0.036 | 0.565 | -0.035 | 0.010 | 0.595 | -0.005 | 0.001 |

True values are: $(\beta^h_{11(0)}, \beta^h_{11(1)}, \beta^g_{11(2)}) = (0.3, 0.3, 0.15)$, $(\beta^h_{10(0)}, \beta^h_{10(1)}, \beta^g_{10(2)}) = (0.6, 0.6, 0.30)$, $(\beta^h_{01(0)}, \beta^h_{01(1)}, \beta^g_{01(2)}) = (0.9, 0.9, 0.45)$, $(\beta^h_{00(0)}, \beta^h_{00(1)}, \beta^g_{00(2)}) = (1.2, 1.2, 0.60)$.

Table A.3: Simulation Results for $\sigma_y = 1$

| $\rho = 0.1$ | $n$ =250 | | | $n$ =1,000 | | | $n$ =10,000 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma_y = 1$ | mean | bias | MSE | mean | bias | MSE | mean | bias | MSE |
| $\hat{\beta}^h_{11(0)}$ | 0.357 | 0.057 | 0.349 | 0.296 | -0.004 | 0.036 | 0.301 | 0.001 | 0.003 |
| $\hat{\beta}^h_{11(1)}$ | 0.341 | 0.041 | 0.539 | 0.299 | -0.001 | 0.058 | 0.299 | -0.001 | 0.003 |
| $\hat{\beta}^g_{11(2)}$ | 0.132 | -0.018 | 0.093 | 0.152 | 0.002 | 0.038 | 0.149 | -0.001 | 0.003 |
| $\hat{\beta}^h_{10(0)}$ | 0.654 | 0.054 | 0.312 | 0.597 | -0.003 | 0.032 | 0.601 | 0.001 | 0.003 |
| $\hat{\beta}^h_{10(1)}$ | 0.639 | 0.039 | 0.484 | 0.599 | -0.001 | 0.052 | 0.599 | -0.001 | 0.002 |
| $\hat{\beta}^g_{10(2)}$ | 0.161 | -0.139 | 0.118 | 0.245 | -0.055 | 0.042 | 0.295 | -0.005 | 0.003 |
| $\hat{\beta}^h_{01(0)}$ | 0.916 | 0.016 | 0.104 | 0.899 | -0.001 | 0.016 | 0.897 | -0.003 | 0.001 |
| $\hat{\beta}^h_{01(1)}$ | 0.932 | 0.032 | 0.385 | 0.888 | -0.012 | 0.015 | 0.896 | -0.004 | 0.001 |
| $\hat{\beta}^g_{01(2)}$ | 0.451 | 0.001 | 0.078 | 0.482 | 0.032 | 0.015 | 0.456 | 0.006 | 0.001 |
| $\hat{\beta}^h_{00(0)}$ | 1.213 | 0.013 | 0.104 | 1.198 | -0.002 | 0.016 | 1.197 | -0.003 | 0.001 |
| $\hat{\beta}^h_{00(1)}$ | 1.234 | 0.034 | 0.385 | 1.188 | -0.012 | 0.015 | 1.196 | -0.004 | 0.001 |
| $\hat{\beta}^g_{00(2)}$ | 0.477 | -0.123 | 0.098 | 0.541 | -0.059 | 0.024 | 0.591 | -0.009 | 0.001 |
| $\rho = 0.7$ | $n$ =250 | | | $n$ =1,000 | | | $n$ =10,000 | | |
| $\sigma_y = 1$ | mean | bias | MSE | mean | bias | MSE | mean | bias | MSE |
| $\hat{\beta}^h_{11(0)}$ | 0.299 | -0.001 | 0.571 | 0.257 | -0.043 | 0.045 | 0.296 | -0.004 | 0.003 |
| $\hat{\beta}^h_{11(1)}$ | 0.367 | 0.067 | 0.945 | 0.283 | -0.017 | 0.053 | 0.298 | -0.002 | 0.003 |
| $\hat{\beta}^g_{11(2)}$ | 0.118 | -0.032 | 0.095 | 0.144 | -0.006 | 0.040 | 0.152 | 0.002 | 0.004 |
| $\hat{\beta}^h_{10(0)}$ | 0.600 | -0.0004 | 0.511 | 0.559 | -0.041 | 0.040 | 0.596 | -0.004 | 0.003 |
| $\hat{\beta}^h_{10(1)}$ | 0.659 | 0.059 | 0.855 | 0.584 | -0.016 | 0.048 | 0.598 | -0.002 | 0.003 |
| $\hat{\beta}^g_{10(2)}$ | 0.149 | -0.151 | 0.120 | 0.225 | -0.075 | 0.048 | 0.297 | -0.003 | 0.004 |
| $\hat{\beta}^h_{01(0)}$ | 0.955 | 0.055 | 0.131 | 0.910 | 0.010 | 0.012 | 0.897 | -0.003 | 0.001 |
| $\hat{\beta}^h_{01(1)}$ | 0.890 | -0.010 | 0.220 | 0.883 | -0.017 | 0.014 | 0.896 | -0.004 | 0.001 |
| $\hat{\beta}^g_{01(2)}$ | 0.468 | 0.018 | 0.063 | 0.480 | 0.030 | 0.013 | 0.455 | 0.005 | 0.001 |
| $\hat{\beta}^h_{00(0)}$ | 1.257 | 0.057 | 0.133 | 1.210 | 0.010 | 0.012 | 1.197 | -0.003 | 0.001 |
| $\hat{\beta}^h_{00(1)}$ | 1.186 | -0.014 | 0.214 | 1.183 | -0.017 | 0.014 | 1.196 | -0.004 | 0.001 |
| $\hat{\beta}^g_{00(2)}$ | 0.501 | -0.099 | 0.081 | 0.538 | -0.062 | 0.024 | 0.591 | -0.009 | 0.001 |

True values are: $(\beta^h_{11(0)}, \beta^h_{11(1)}, \beta^g_{11(2)}) = (0.3, 0.3, 0.15)$, $(\beta^h_{10(0)}, \beta^h_{10(1)}, \beta^g_{10(2)}) = (0.6, 0.6, 0.30)$, $(\beta^h_{01(0)}, \beta^h_{01(1)}, \beta^g_{01(2)}) = (0.9, 0.9, 0.45)$, $(\beta^h_{00(0)}, \beta^h_{00(1)}, \beta^g_{00(2)}) = (1.2, 1.2, 0.60)$.

## Derivation of $E[\tilde{U}_Y|X, \mathcal{C}]$ in (A.17)

To derive the conditional mean $E[\tilde{U}_Y|X, \mathcal{C}]$, note that according to (2.8) and (3.3), conditional on $X$, the event of being a complier is equivalent to:

$$\begin{aligned} \mathcal{C} &= \{D(1) = 1, D(0) = 0\} \\ &= \{U_D \geq -(\gamma_{z1} + X^\top \gamma_x) \text{ and } U_D \leq -X^\top \gamma_x\} \\ &= \{-(\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x\}. \end{aligned} \tag{A.18}$$

Therefore,

$$\begin{aligned} E[\tilde{U}_Y|X, \mathcal{C}] =& E[\tilde{U}_Y| - (\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x] \\ =& E[\rho_{yd} U_D + \epsilon| - (\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x] \\ =& E[\rho_{yd} U_D| - (\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x] \\ =& \rho_{yd} \frac{E[U_D \cdot 1(-(\gamma_{z1} + X^\top \gamma_x) \leq U_D)] - E[U_D \cdot 1(-X^\top \gamma_x \leq U_D)]}{P(-(\gamma_{z1} + X^\top \gamma_x) \leq U_D) - P(-X^\top \gamma_x \leq U_D)} \\ =& \rho_{yd} \left( \frac{\phi(\gamma_{z1} + X^\top \gamma_x) - \phi(X^\top \gamma_x)}{1 - \Phi(-(\gamma_{z1} + X^\top \gamma_x)) - (1 - \Phi(-X^\top \gamma_x))} \right) \\ =& \rho_{yd} \left( \frac{\phi(\gamma_{z1} + X^\top \gamma_x) - \phi(X^\top \gamma_x)}{\Phi(\gamma_{z1} + X^\top \gamma_x) - \Phi(X^\top \gamma_x)} \right), \end{aligned}$$

where the second equality follows from the fact that, given (A.16), $\tilde{U}_Y$ can be rewritten as $\rho_{yd} U_D + \epsilon$ where $\epsilon$ is normally distributed with mean 0 and is independent of $U_D$. The third equality holds because the $\epsilon$ is independent of $U_D$ and its mean is 0. The fourth equality follows by definition. The fifth equality holds by the fact that $E[U_D \cdot 1(-a \leq U_D)] = \phi(a)$ and $P(U_D \geq -a) = P(U_D \leq a)$ by the symmetry of $U_D$. Then the last equality follows. See also Cameron and Trivedi (2005, Proposition 6.1) for a related discussion. □

# 4 Derivation of (5.12)

To derive (5.12), note that (3.8) can be represented as:

$$\begin{aligned}&\Psi_{d'}(z_2, X, \eta)\\ &= E\left[E\left[1(U_M \geq -(\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x))|U_D\right]| - (\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x)\right].\end{aligned} \tag{A.12}$$

It is known that if

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),$$

then

$$x_1|x_2 \sim N\left(\mu_{1|2}, \Sigma_{11|2}\right), \tag{A.13}$$

where $\mu_{1|2} := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ and $\Sigma_{11|2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$; moreover, if $x \sim N(0,1)$, then the conditional PDF of $x|x > -u$ is of the inverse Mills ratio:

$$f(v|x > -u) = \frac{\phi(v)}{\Phi(u)}, \tag{A.14}$$

for some $v > -u$. See, e.g., Wooldridge (2010, p.594-595) for the use of these two properties of normality in exploring the bivariate probit model. From (5.11) and (A.13), we can obtain the result:

$$U_M|U_D, Z_1, Z_2, X \sim N(\rho_{md}U_D, (1 - \rho_{md}^2))$$

and use this result to define a standard normal random variable:

$$U_M^* := \left.\frac{U_M - \rho_{md}U_D}{\sqrt{1 - \rho_{md}^2}}\right| U_D, Z_1, Z_2, X, \eta \sim N(0,1), \tag{A.15}$$

From (A.15), we can further obtain the result:

$$
\begin{aligned}
&E\left[1(U_M \geq -(\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x))|U_D\right] \\
&= E\left[1(U_M^* \geq -\frac{(\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x + \rho_{md} U_D)}{\sqrt{1-\rho_{md}^2}}) \middle| U_D\right] \\
&= \Phi\left(\frac{\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x + \rho_{md} U_D}{\sqrt{1-\rho_{md}^2}}\right),
\end{aligned}
$$

and use this result to write (A.12) as:

$$
\begin{aligned}
&\Psi_{d'}(z_2, X, \eta) \\
&= E\left[\Phi\left(\frac{\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x + \rho_{md} U_D}{\sqrt{1-\rho_{md}^2}}\right) \middle| -(\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x)\right] \\
&= \frac{1}{\Phi(\gamma_{z1} + X^\top \gamma_x) - \Phi(X^\top \gamma_x)} \int_{X^\top \gamma_x}^{\gamma_{z1} + X^\top \gamma_x} \Phi\left(\frac{\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x + \rho_{md} u}{\sqrt{1-\rho_{md}^2}}\right) \phi(u)\mathrm{d}u,
\end{aligned}
$$

where the last equality is due to a probability of the truncated normal distribution like (A.14). This proves (5.12). □

# 5 Proof of Theorems

## Proof of Theorem 3.1

According to (3.2) and (3.8), we can write that

$$
\begin{aligned}
&E[M(d')|Z_2, X, \mathcal{C}] \\
&= E[1(U_M \geq -(\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x))|Z_2, X, -(\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x] \\
&= P(U_M \geq -(\alpha_d d' + \alpha_z Z_2 + X^\top \alpha_x)| -(\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x) \\
&= \Psi_{d'}(Z_2, X, \alpha_q, \gamma),
\end{aligned}
\tag{A.19}
$$

where the second equality is due to a requirement of Assumption 3.4($c$): $U_M \perp (Z_2, X)$. Let $f_{Z_2|X,\mathcal{C}}(\cdot)$ be the conditional PDF of $Z_2|X, \mathcal{C}$. By the law of total probability, we have

$$\begin{aligned} E[M(d')|X,\mathcal{C}] &= \int_R E[M(d')|Z_2 = z_2, X, \mathcal{C}] f_{Z_2|X,\mathcal{C}}(z_2)\mathrm{d}z_2, \\ &= \int_R \Psi_{d'}(z_2, X, \alpha_q, \gamma) f_{Z_2|X,\mathcal{C}}(z_2)\mathrm{d}z_2, \\ &= \int_R \Psi_{d'}(z_2, X, \alpha_q, \gamma) f_{Z_2}(z_2, \alpha_f)\mathrm{d}z_2, \end{aligned} \tag{A.20}$$

where the second equality is obtained by evaluating (A.19) at $Z_2 = z_2$; the last equality is also due to Assumption 3.4($b$): $Z_2 \perp (U_D, X)$. The result in (3.10) is a combination of (2.16) and (A.20). □

## Proof of Theorem 3.2

Note that the model that Frölich and Huber (2014) consider is

$$\begin{aligned} &Y = \kappa(D, M, X, U_Y), \\ &M = 1(\zeta(D, Z_2, X, U_M) \geq 0), \\ &D = 1(\chi(Z_1, X, U_D) \geq 0), \end{aligned}$$

where $\kappa(\cdot)$, $\zeta(\cdot)$ and $\chi(\cdot)$ are measurable functions. Theorem 4 of Frölich and Huber (2014) is proved to be true under the following assumption:

(i) IV independence:

$(Z_1, Z_2) \perp (U_Y, U_M)|T, X$,

$Z_1 \perp (U_Y, U_M, T)|Z_2, X$;

$T$ is the subpopulation type: always takers, compliers, defiers or never takers.

(ii) Conditional independence of IVs:

$Z_1 \perp Z_2|X$.

(iii) Monotonicity: $P(D(1) \geq D(0)|X) = 1$;

Existence of compliers: $E[D(1)] > E[D(0)]$.

(iv) Monotonicity of $M$ in $Z_2$ and $U_M$:

$U_M$ is a continuous random variable with a distribution function $F_{U_M|X=x,\mathcal{C}}(v)$ which is strictly increasing in the support of $U_M$ for almost all $x$.

$\zeta(D, Z_2, X, U_M)$ is (normalized to be) strictly monotonic in $Z_2$ and in $U_M$.

(v) Common support of $M$:

$0 < E[Z_1|M, U_M, X, \mathcal{C}] < 1$.

We first claim that our Assumptions 3.2 and 3.3 are sufficient for these conditions. First, the types $T$ are determined by $U_D$ conditional on $X$. Therefore, our Assumption 3.3($a$) is sufficient for (i). Assumption 3.3($b$) is identical to (ii). As mentioned before, the sign restriction, $\gamma_{z1} > 0$, in Assumption 3.3($c$) combined with the model $D = 1(\gamma_{z1}Z_1 + X^\top\gamma_x + U_D \geq 0)$ imply (3.4) and (3.5). Therefore, (iii) is satisfied. Our model for $M$ is $M = 1(\alpha_d D + \alpha_{z_2}Z_2 + X^\top\alpha_x + U_M \geq 0)$ so $\zeta(D, Z_2, X, U_M) = \alpha_d D + \alpha_{z_2}Z_2 + X^\top\alpha_x + U_M$. The sign condition, $\alpha_{z_2} > 0$, and the model implies that $\alpha_d D + \alpha_{z_2}Z_2 + X^\top\alpha_x + U_M$ is strictly increasing in $Z_2$ and $U_M$. Therefore, Assumption 3.3($d$) combined with Assumption 3.2 is sufficient for (iv). Assumption 3.3($e$) is identical to (v). Then we use Theorem 4 of Frölich and Huber (2014), and with this linear index model of $M$ we can show that, for $d, d' \in \{0, 1\}$,

$$\begin{aligned}&E\Big[\Big(Y(d, M(d')) - h_{d'}(X, \alpha_m, b^h_{d1}, b^h_{d0}) + g_{d'}(X, \alpha_m, b^g_{d1}, b^g_{d0})\Big)^2\Big|\mathcal{C}\Big]\\ =&E\Big[\tilde{w}(d, d')\Big(Y - h_{d'}(X, \alpha_m, b^h_{d1}, b^h_{d0}) + g_{d'}(X, \alpha_m, b^g_{d1}, b^g_{d0})\Big)^2\Big]\Big/\Delta,\end{aligned}$$

where

$$\tilde{w}(d,d') \equiv \begin{cases} D\frac{(Z_1 - E[Z_1|X])}{(E[Z_1|X])(1-E[Z_1|X])}, & (d,d') = (1,1), \\ D\frac{(Z_1 - E[Z_1|X])}{(E[Z_1|X])(1-E[Z_1|X])}\frac{f_{Z_2|X,\mathcal{C}}(Z_2 + \alpha_d/\alpha_{z_2})}{f_{Z_2|X,\mathcal{C}}(Z_2)}, & (d,d') = (1,0), \\ (D-1)\frac{(Z_1 - E[Z_1|X])}{(E[Z_1|X])(1-E[Z_1|X])}\frac{f_{Z_2|X,\mathcal{C}}(Z_2 - \alpha_d/\alpha_{z_2})}{f_{Z_2|X,\mathcal{C}}(Z_2)}, & (d,d') = (0,1), \\ (D-1)\frac{(Z_1 - E[Z_1|X])}{(E[Z_1|X])(1-E[Z_1|X])}, & (d,d') = (0,0). \end{cases}$$

For the implication of a single-index model of $M$ on the weight functions $\tilde{w}(1,0)$ and $\tilde{w}(0,1)$, see the end of Section 3.3 of Frölich and Huber (2014). Assumption 3.4($a$) implies that $E[Z_1|X] = Q_{Z_1}(X^\top \alpha_{z_1})$ and Assumption 3.4($b$) implies that $f_{Z_2|X,\mathcal{C}}(Z_2) = f_{Z_2}(Z_2, \alpha_f)$. Then under our assumptions, we have $\tilde{w}(d,d') = w(d,d',\alpha_w)$ for $d,d' \in \{0,1\}$. This completes our proof. □

# 6 Assumption 3.3(e) in the Dam Example

Note that $M$ takes on 0 and 1, so Assumption 3.3($e$) is equivalent to the following

$$P(Z_1 = z, M = m | U_M, X, \mathcal{C}) > 0 \quad \text{for all } z, m \in \{0,1\}. \tag{A.21}$$

Recall that, in the dam example, we have

$$\begin{aligned} M &= 1(\alpha_d D + \alpha_{z_2} Z_2 + X\alpha_x + U_M \geq 0), \\ D &= 1(\gamma_{z_1} Z_1 + X\gamma_x + U_D \geq 0), \\ Z_1 &= 1(X\alpha_{z_1} + U_{Z_1}), \end{aligned}$$

where

$$\begin{bmatrix} U_{Z_1} \\ U_D \\ U_M \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho_{md} \\ 0 & \rho_{md} & 1 \end{bmatrix} \right).$$

For a complier, it has to be the case where $-(\gamma_{z1} + X^\top \gamma_x) \leq U_D \leq -X^\top \gamma_x$. Because $U_D \sim N(0,1)$, we have $P(\text{complier}|X) > 0$ almost surely in $X$. To show (A.21), note that because $U_{Z_1}$ is independent of $U_M$ and $U_D$, the specification in (5.14) means that $P(Z_1 = 1|U_M, X, \mathcal{C}) = \Phi(X^\top \alpha_{z_1})$. This conditional probability is strictly greater than 0 and strictly less than 1. Moreover,

$$\begin{aligned} P(M = 1|Z_1, U_M, X, C) &= P(\alpha_d D + \alpha_{z_2} Z_2 + X\alpha_x + U_M \geq 0|Z_1, U_M, X, \mathcal{C}) \\ &= P(\alpha_d D + \alpha_{z_2} Z_2 + X\alpha_x + U_M \geq 0|D, U_M, X, \mathcal{C}), \end{aligned}$$

where the second equality holds because for compliers, $D = Z_1$. Note that $Z_2$ is independent of everything in our empirical study and it has full support of the real line, so we have $< P(M = 1|Z_1, U_M, X, C) < 1$ too. This implies that $P(Z_1 = z, M = m|U_M, X, \mathcal{C}) = P(M = m|Z_1, U_M, X, C) \cdot P(Z_1 = z|U_M, X, \mathcal{C}) > 0$ for all $z, m \in \{0, 1\}$ which shows (A.21). □

# 7 Auxiliary Estimates in the Dam Example

In this section, we provide estimation results of the auxiliary parameters in the submodels of the paper. All the reported standard errors of the estimates are from the econometric software's (Stata) output, i.e., they are not bootstrapped standard errors.

| Model | Parameters | Results |
|---|---|---|
| The model of $P(M, D \mid Z_1, Z_2, X,\eta)$; (5.11) of the paper. | $\eta = (\alpha_d, \alpha_{z_2}, \alpha_{\boldsymbol{x}}^\top, \gamma_{z1}, \gamma_{\boldsymbol{x}}^\top, \rho_{md})^\top$, | Table A.4 |
| The model of the instrument propensity score; (5.14) of the paper. | $\alpha_{z1(\boldsymbol{x})}^\top$ | Table A.5 |
| The distribution of $Z_2$; (5.15) of the paper. | $\alpha_f = (\alpha_{f(m1)},\ \alpha_{f(s1)}, \alpha_{f(m2)},\ \alpha_{f(s2)},\ \alpha_{f(p)})^\top$ | Table A.6 |

Table A.4: The Model of M and D

| | Model of M | | | Model of D | |
|---|---|---|---|---|---|
| | coef. | std.err. | | coef. | std.err. |
| $\alpha_0$ | -6.668*** | 2.377 | $\gamma_0$ | 0.052 | 2.244 |
| $\alpha_d$ | -1.040** | 0.434 | | | |
| $\alpha_{z2}$ | 0.224*** | 0.064 | $\gamma_{z1}$ | 0.384* | 0.198 |
| $\alpha_r$ | -0.030 | 1.272 | $\gamma_r$ | -3.432*** | 1.296 |
| $\alpha_e$ | -0.239 | 0.189 | $\gamma_e$ | 0.064 | 0.193 |
| $\alpha_p$ | -0.976 | 1.735 | $\gamma_p$ | 11.651*** | 2.695 |
| $\alpha_f$ | 0.762*** | 0.138 | $\gamma_f$ | -0.172** | 0.087 |
| $\alpha_L$ | -0.594*** | 0.221 | $\gamma_L$ | 0.276 | 0.215 |
| $\alpha_{c1}$ | 0.028* | 0.017 | $\gamma_{c1}$ | -0.009 | 0.014 |
| $\alpha_{c2}$ | -0.057 | 0.047 | $\gamma_{c2}$ | 0.185*** | 0.046 |
| $\alpha_{r2}$ | -0.710 | 1.605 | $\gamma_{r2}$ | -2.115 | 1.946 |
| $\rho_{md}$ | 0.582** | 0.242 | | | |

Note 1: Significance: ***: 1% level, **: 5% level; *: 10% level.

Note 2: Subscripts of the coefficients indicate the associated variables: intercept (0), $D$ ($d$), $Z_1$ ($z1$), $Z_2$ ($z2$), $rain$ ($r$), $elevation$ ($e$), *pre_dam* ($p$), *pre_fert* ($f$), *pre_land* ($L$), $pc_1$ ($c1$), $pc_2$ ($c2$), and $rain2$ ($r2$).

Table A. 5: The Propensity Score Model of $Z1$

| dep.var.: $Z1$ | coef. | std.err. |
|---|---|---|
| $\alpha_{z1(0)}$ | -2.985 | 1.905 |
| $\alpha_{z1(r)}$ | -1.844* | 1.006 |
| $\alpha_{z1(e)}$ | 0.285* | 0.168 |
| $\alpha_{z1(p)}$ | 3.029** | 1.313 |
| $\alpha_{z1(f)}$ | 0.166** | 0.078 |
| $\alpha_{z1(L)}$ | -0.217 | 0.178 |
| $\alpha_{z1(c1)}$ | -0.003 | 0.012 |
| $\alpha_{z1(c2)}$ | -0.039 | 0.035 |
| $\alpha_{z1(r2)}$ | -1.778 | 1.575 |

Note 1: Significance: ***: 1% level, **: 5% level; *: 10% level.

Note 2: For the meaning of coefficient subscripts, see Note 2 of Table A.4.

Table A. 6: The Mixture Normal Distribution of $Z2$

| dep.var.: $Z2$ | coef. | std.err. |
|---|---|---|
| $\alpha_{f(m1)}$ | -1.369*** | 0.291 |
| $\alpha_{f(s1)}$ | 0.996*** | 0.155 |
| $\alpha_{f(m2)}$ | 1.110*** | 0.244 |
| $\alpha_{f(s2)}$ | 1.007*** | 0.133 |
| $\alpha_{f(p)}$ | 0.448*** | 0.104 |

Note 1: Significance: ***: 1% level, **: 5% level; *: 10% level.

Note 2: The subscripts $f(m1)$ and $f(m2)$ indicates the mean of the two normal distributions, and $f(s1)$ and $f(s2)$ are the standard deviations of the distributions. The subscript $f(p)$ indicates the weight on the first distribution.

# References

[1] Cameron, A. C. and Trivedi, P. K. (2005), "Microeconometrics: Methods and Applications," Cambridge University Press.

[2] Frölich, M. and Huber, M. (2014), "Direct and Indirect Treatment Effects: Causal Chains and Mediation Analysis with Instrumental Variables," *CEMAP Working Paper CWP31/14*.

[3] Newey, W. K. and McFadden, D. L. (1994), "Large Sample Estimation and Hypothesis Testing," *In R. F. Engle and D. L. McFadden (Eds.), Handbook of*

*Econometrics, Volume 4*, 2111-2245. Amsterdam: Elsevier.

[4] White, H. (1994), "Estimation, Inference, and Specification Analysis," Cambridge: Cambridge University Press.

[5] Wooldridge, J. M. (2010), "Econometric Analysis of Cross Section and Panel Data," The MIT Press.