

# Constructing an Instrument as a Function of Covariates

Moses Stewart  
Harvard College\*

June 2025

## Abstract

Researchers often use instrumental variables (IV) models to investigate the causal relationship between an endogenous variable and an outcome while controlling for covariates. When an exogenous variable is unavailable to serve as the instrument for an endogenous treatment, a recurring empirical practice is to construct one from a nonlinear transformation of the covariates. We investigate how reliable these estimates are under mild forms of misspecification. Our main result shows that for instruments constructed from covariates, the IV estimand can be arbitrarily biased under mild forms of misspecification, even when imposing constant linear treatment effects. We perform a semi-synthetic exercise by calibrating data to alternative models proposed in the literature and estimating the average treatment effect. Our results show that IV specifications that use instruments constructed from covariates are non-robust to nonlinearity in the true structural function.

## 1 Introduction

Instrumental variable (IV) strategies are widely used to investigate the causal effect of an endogenous variable on an outcome of interest. Since the work of [Imbens and Angrist \(1994\)](#), these estimators are often interpreted as estimating a local average treatment effect (LATE) or, under homogeneity conditions, the population treatment effect of the endogenous variable of interest.

However, estimating this causal effect requires finding valid instruments that satisfy the exclusion restriction, which can be challenging in many applications. One response to this challenge is to construct an instrumental variable that is a nonlinear function of covariates.<sup>1</sup> The leading case is to use the “product instrument,” the interaction of two covariates. A sample of recent articles published in the *American Economic Review* that use this approach are [Munshi and Rosenzweig \(2016\)](#), [Bettinger et al. \(2017\)](#), [Aladangady \(2017\)](#), and [Rogall \(2021\)](#). Other examples include [Shao \(2014\)](#), [Bharadwaj \(2015\)](#), [Helmers et al. \(2017\)](#), [Crawford and Simpson \(2020\)](#), [Cagé et al. \(2022\)](#), [Liu et al. \(2023\)](#), and [Ghose et al. \(2024\)](#).

In this paper, we study the causal interpretability of these IV estimators that use a nonlinear function of covariates as the instrumental variable, under misspecification. A researcher may be interested in the effect of a variable  $D$  on some outcome  $Y$ , where  $D$  may be endogenous to unobserved factors affecting  $Y$ . The researcher’s model specifies  $Y$  as a linear function of  $D$  and some covariates  $X$ , with the causal relationships governed by a parameter  $\tilde{\theta}$ . They do not have access to an excluded variable  $Z$  that is correlated with the endogenous variable  $D$  but not the unobserved factors affecting the relationship

---

\*Advised by Rahul Singh. Thanks to Jesse Shapiro for his guidance throughout the research process and Isaiah Andrews for his comments.

<sup>1</sup>Here we use covariate to mean “included exogenous variable,” and reserve “excluded variable” to refer to excluded exogenous variables.

between  $D$  and  $Y$  (conditional upon  $X$ ). Instead, a recurring practice is for the researcher to construct an instrumental variable  $\tilde{f}(X) = X_1 \cdot X_2$  that is the interaction of covariates, which they argue satisfy the IV assumptions. The researcher then uses IV estimators to evaluate  $\tilde{\theta}$ .

For example, [Aladangady \(2017\)](#) uses a linear model to examine the effect of log house price fluctuations  $D$  on log consumer spending  $Y$  in the United States while trying to control for potential unobserved confounding caused by preferences or expectations that change over time. They would like to use a measurement of local land availability or zoning regulations  $Z$  that is uncorrelated with preference shifts but varies geographically. However, they do not have access to detailed zoning data across regions. Therefore, they use the interaction between long-term real interest rates  $X_1$  and housing supply elasticity measures  $X_2$  as an instrument for house price fluctuations and estimate the two-stage least squares (2SLS) model,

$$Y = \tilde{\pi}_0 + \tilde{\theta}D + \tilde{\pi}_1X_1 + \tilde{\pi}_2X_2 + \tilde{\epsilon}_i, \quad D = \tilde{\tau}_0 + \tilde{\delta}\tilde{f}(X) + \tilde{\tau}_1X_1 + \tilde{\tau}_2X_2 + \tilde{v}_i,$$

for  $\tilde{f}(X) = X_1 \cdot X_2$ . By including long-term real interest rates  $X_1$  and housing supply elasticity measures  $X_2$  in the specification for consumer spending  $Y$ , they hope to control for the *direct effect* of interest rates and housing elasticity on spending trends that may be related to housing prices. This relaxes the exogeneity assumption required on the instrument  $\tilde{f}(X)$  because even if long-term interest rates or housing elasticity are correlated with unobserved determinants of spending, as long as their interaction is uncorrelated with these unobserved determinants, the model assumptions required for IV are still satisfied.

We show that for IV estimators that use a function of covariates as an instrument, even while imposing homogeneous treatment effects, the researcher’s model is uninterpretable as a causal effect under mild forms of misspecification. As the main contribution we show that for any level of bias and for any IV estimand which relies on an instrument that is a transformation of covariates  $f(X, Z) = \tilde{f}(X)$ , there is a continuous data generating process (DGP) with constant, linear treatment effects under which we can attain that level of bias.

Quantitatively, we investigate the sensitivity of linear IV estimates that use a function of covariates  $\tilde{f}(X)$  as an instrument to a reasonable set of nonlinear DGPs. For a collection of influential papers that use linear IV with product instruments, we perform a semi-synthetic exercise by estimating the treatment effect of the endogenous variable  $D$  under nonlinear structural functions proposed in the literature. For three out of four pairs of papers we examine, we find that a 95% confidence interval for the treatment effect  $\hat{\theta}$  rejects the ground truth in our simulations. These findings emphasize the sensitivity of IV estimands that use product instruments to the true relationship between the outcome  $Y$  and covariates  $X$  in the true structural function. This paper aims to guide the causal interpretation of these estimands.

A large literature following [Imbens and Angrist \(1994\)](#) and [Angrist et al. \(1996\)](#) studies the interpretation of IV estimators under misspecification. Our work is closest to [Blandhol et al. \(2022\)](#), which investigates the LATE interpretation of IV estimators that include exogenous covariates  $X$ . Unlike [Blandhol et al. \(2022\)](#), we allow for continuous outcomes and continuous treatment effects. [Blandhol et al. \(2022\)](#) shows that IV estimators with instruments that do not satisfy a condition called “saturated covariates” cannot be interpreted as LATE. As we show in section 2, the estimators we study are undefined under saturated covariates since there is no variation in the instrument  $\tilde{f}(X)$ . Appendix C expands on

the relationship between our work and [Blandhol et al. \(2022\)](#) and generalizes some of the results in the latter.

Our work is also closely related to [Andrews et al. \(2023\)](#), which studies the causal consistency of nonlinear IV estimators under misspecification. They show that IV estimators with instruments that do not satisfy a condition known as “strong exclusion” cannot approximately describe the effects of the endogenous variable  $D$ . The IV specifications we study do not have access to excluded variables, and therefore do not satisfy strong exclusion. We differ from [Andrews et al. \(2023\)](#) by showing a more significant failure of the causal interpretability of IV estimators using product instruments even when imposing homogeneous, linear treatment effects, meaning the analyst’s model satisfies causally correct specification.

Researchers often justify their models using economic reasoning, which can lead them to believe their model is correctly specified. This motivates [Gao and Wang \(2023\)](#), who introduce a causal estimator that does not depend on excluded covariates but is sensitive to the functional form chosen by the analyst. If the researcher observes the true structural function  $Y(D, X, \epsilon)$ , or if the model proposed by the researcher  $\tilde{Y}(D, X, \tilde{\epsilon}_i)$  matches the true DGP, then the estimator from [Gao and Wang \(2023\)](#) and the linear IV specifications we study both remain unbiased. Our work focuses on cases where the researcher does not know the true structural function, so their model may be mildly misspecified.

The remainder of the paper proceeds as follows. Section 2 introduces the assumptions of our model, which includes the IV conditions assumed by the analyst and the true structural function. Section 3 establishes our main result and shows that IV estimators that rely on instruments constructed from covariates can be arbitrarily biased. Finally, Section 4 shows our quantitative results and illustrates the wide range of estimates produced by IV estimators that rely on product instruments across plausible DGPs. We reserve our theoretical proofs for the appendix, with the main text focusing on the key results.

## 2 Structural model

A analyst observes variables  $(Y_i, D_i, X_i, Z_i)$  for units  $i \in \{1, \dots, n\}$ . All variables are finite-dimensional, and  $Y_i \in \mathbb{R}$ . To illustrate the possibility of misspecification, we introduce a *true model* that is consistent with the data generating process (DGP) and summarizes the causal relationships between the observed variables, and the *analyst’s model* which may rule out the true DGP.

### 2.1 True model

Under the true model, which is consistent with the true data-generating process, the observed outcome satisfies  $Y_i = Y(D_i, X_i, \epsilon_i)$  for a potential outcome function  $d \mapsto Y(d, X_i, \epsilon_i) \in \mathcal{Y} \subseteq \mathbb{R}$ , covariates  $X_i = (X_{i,1}, \dots, X_{i,J})^\top \in \mathcal{X} \subseteq \mathbb{R}^{\dim(X) \times 1}$ , and unobserved heterogeneity  $\epsilon_i$ . The observed endogenous variable  $D_i$  satisfies  $D_i = D(X_i, Z_i, \nu_i)$  for  $D(X_i, Z_i, \nu_i) \in \mathcal{D} \subseteq \mathbb{R}$ , a potential endogenous variable function, unobserved heterogeneity  $\nu_i$ , and an excluded exogenous variable  $Z_i \in \mathcal{Z} \subseteq \mathbb{R}^{\dim(Z) \times 1}$ . We impose one additional assumption on the true model in line with instrumental variables (IV) models.

**Assumption 1** (Excludability). *The potential outcome  $Y(\cdot)$  is independent of the excluded exogenous variable  $Z$ ,*

$$Y_i = Y(D_i, X_i, Z_i, \epsilon_i) = Y(D_i, X_i, \epsilon_i)$$

Under Assumption 1, we can write  $Y_i = Y(D_i, X_i, \epsilon_i)$ .

*Definition 1* (Linear IV Estimation). Let  $f(X, Z)$  be a choice of instrumental variable for  $f : \mathcal{X} \times \mathcal{Z} \mapsto \mathbb{R}$ . Then, the two-stage least squares (2SLS) IV estimand  $\theta_{IV}$  is given by,

$$\theta_{IV} = \frac{\mathbb{E}\{Y f_{\perp}\}}{\mathbb{E}\{D f_{\perp}\}} = \frac{\text{Cov}(Y, f_{\perp})}{\text{Cov}(D, f_{\perp})},$$

where  $f_{\perp}$  are the residuals from a projection of the instrumental variable  $f(X, Z)$  onto  $X$ ,

$$f_{\perp} = f(X, Z) - X^{\top} \mathbb{E}\{X X^{\top}\}^{-1} \mathbb{E}\{X f(X, Z)\}.$$

Definition 1 establishes the linear IV estimand to estimate the causal effect of the treatment  $D$  on the outcome  $Y(D_i, X_i, \epsilon_i)$  under a linear model. This is the estimand computed by linear method of moments and two-stage least squares estimators for an instrumental variable  $f(X, Z)$  when there is a single endogenous treatment  $D$ .

## 2.2 Analyst's model

The analyst's model is a special case of the true model. Under the analyst's model, the causal effects of the endogenous variable and included variables on the outcome are governed by a parameter  $\tilde{\theta} \in \mathbb{R}$ . Here, we focus on the case where the analyst proposes and estimates a linear model with homogeneous treatment effects,

$$Y_i = \tilde{Y}(D_i, X_i, \tilde{\epsilon}_i) = \tilde{\alpha} + \tilde{\theta} D_i + \tilde{\pi}^{\top} X_i + \tilde{\epsilon}_i. \quad (1)$$

The analyst would like to estimate the causal effect of  $D$  on the outcome  $Y$  in Equation 1, but is unwilling to assume unconfoundedness ( $D \perp \tilde{\epsilon} \mid X$ ). Therefore, to use IV estimation, the analyst constructs an instrument  $f(X, Z)$  under two assumptions sufficient to estimate  $\tilde{\theta}$ .

**Assumption 2** (Exogeneity). *The unobserved heterogeneity ( $\tilde{\epsilon}_i$ ) in  $Y_i$  is uncorrelated ( $\perp$ ), or orthogonal, to the instrumental variable  $f(X_i, Z_i)$  and covariates  $X_i$ ,*

$$f(X, Z), X \perp \tilde{\epsilon}.$$

Assumption 2 says that the instrument  $f(X, Z)$  and included exogenous covariates  $X$  must be uncorrelated ( $\perp$ ) with the unobserved heterogeneity  $\tilde{\epsilon}$  in the outcome under their proposed model.

**Assumption 3** (Relevance). *The instrument  $f(X, Z)$  is a random variable such that  $\mathbb{E}\{D \mid f(X, Z) = w\}$  is a non-trivial function of  $w$ , so  $\text{Cov}(D, f(X, Z)) \neq 0$ .*

Assumption 3 is necessary for  $\theta_{IV}$  in Definition 1 to be defined, so the denominator is not equal to zero. Otherwise, the treatment effect  $\tilde{\theta}$  is unidentified using linear IV estimation.

In this paper, we focus on the case where the analyst does not have access to an excluded exogenous variable  $Z$  to create an instrumental variable  $f(X, Z)$  that is a nontrivial function of  $Z$ . Therefore, the analyst chooses  $f(X, Z)$  to be a nonlinear function of only covariates  $f(X, Z) = \tilde{f}(X) \in \mathbb{R}$  to use IV estimation.

**Lemma 1** (Unsaturated Covariates). *For any function  $f(X, Z) = \tilde{f}(X)$  only of covariates,  $\text{Var}(f_{\perp}) \neq 0$  if and only if  $\tilde{f}(X)$  is not linear in  $X$  for  $f_{\perp} = \tilde{f}(X) - X^{\top} \mathbb{E}\{XX^{\top}\}^{-1} \mathbb{E}\{Xf(X, Z)\}$ .*

Lemma 1 makes it clear why the analyst’s instrumental variable  $\tilde{f}(X)$  must be nonlinear in  $X$  to use the IV estimator from Definition 1. If  $\tilde{f}(X)$  is linear in each element of  $X$ , then the linear IV estimand  $\theta_{IV}$  becomes unidentified because the denominator  $\text{Cov}(D, f_{\perp}) = \text{Cov}(D, 0) = 0$ .

**Lemma 2** (Trivial Exogeneity). *Suppose the analyst believes a stronger form of conditional exogeneity,  $\mathbb{E}\{\tilde{\epsilon} | X\} = 0$ . Then, any function of covariates  $\tilde{f}(X)$  satisfies exogeneity,*

$$\tilde{f}(X) \perp \tilde{\epsilon}.$$

Lemma 2 demonstrates that an instrument based solely on included exogenous covariates automatically satisfies Assumption 2 (exogeneity) for IV estimation when the analyst assumes a stronger exogeneity condition for these covariates. Some researchers using instruments of the form  $f(X, Z) = \tilde{f}(X)$  argue that such instruments “require weaker identifying assumptions than using  $[X_1$  or  $X_2]$  alone” [Bettinger et al. \(2017\)](#). By including  $X_i$  in the second-stage regression, they “control for the direct effect of  $[X_i]$  that may be related to  $[Y_i]$ ” [Aladangady \(2017\)](#). If the covariates  $X$  included in the second-stage equation  $\tilde{Y}(D_i, X_i, \tilde{\epsilon}_i)$  satisfy conditional exogeneity ( $\mathbb{E}[\tilde{\epsilon} | X] = 0$ ), then a nonlinear function of  $X$  can be used as an instrument without needing to independently verify its exogeneity under Assumption 2.

The following proposition shows that this estimand corresponds to the causal effect  $\tilde{\theta}$  of  $D_i$  on  $Y(D_i, X_i, \epsilon_i)$  under the analyst’s model if the true data generating process follows the model assumed by the analyst.

**Claim 1** (IV Validity). *Assume the true data generating process  $Y(D_i, X_i, \epsilon_i)$  follows the analyst’s constant linear effects model in Equation 1 and Assumption 2,*

$$Y_i = Y(D_i, X_i, \epsilon_i) = \tilde{Y}(D_i, X_i, \tilde{\epsilon}_i).$$

*and the analyst uses an instrument  $f(X, Z) = \tilde{f}(X)$  that is a nonlinear function of covariates. Then, under Assumption 3 (relevance), the IV estimand is equal to the causal effect of  $D_i$  on  $Y(D_i, X_i, \epsilon_i)$ ,*

$$\theta_{IV} = \frac{\text{Cov}(Y, f_{\perp})}{\text{Cov}(D, f_{\perp})} = \tilde{\theta},$$

for  $f_{\perp} = \tilde{f}(X) - X^{\top} \mathbb{E}\{XX^{\top}\}^{-1} \mathbb{E}\{Xf(X, Z)\}$ .

Claim 1 shows that if the analyst’s model is correctly specified and their constructed instrument  $f(X, Z) = \tilde{f}(X)$  satisfies Assumption 3 (relevance), IV estimators  $\theta_{IV}$ <sup>2</sup> are consistent for the treatment

<sup>2</sup>Note that  $\theta_{IV}$  is implicitly indexed by  $\tilde{f}(X)$

effect. [Gao and Wang \(2023\)](#) show that even for nonlinear models, if the analyst’s model matches the true DGP, the causal effect  $\theta$  of  $D$  on  $Y$  can be recovered locally. We next move to analyze what happens when the analyst’s model is not correctly specified.

### 3 Main result: Non-interpretability

When the analyst estimates a linear model like Equation 1, any nonlinear effect of the covariates  $X$  on the outcome  $Y$  is captured in the unobserved heterogeneity  $\tilde{\epsilon}$ . This means that IV estimators that rely on moment conditions interacting  $\tilde{\epsilon}$  and the covariates  $X$  will be sensitive to nonlinearity in the true structural function  $Y(D_i, X_i, \epsilon_i)$ .

In Theorem 1, we illustrate this sensitivity using a toy case where the true DGP imposes linear, homogeneous treatment effects. We show that if the analyst uses an instrumental variable  $f(X, Z) = \tilde{f}(X)$  that is only a function of covariates, the bias of the IV estimand  $\theta_{IV}$  can be arbitrarily bad.

**Theorem 1** (Model bias). *Assume the analyst chooses an instrument  $f(X_i, Z_i) = \tilde{f}(X_i)$  that is a nonlinear function of covariates.<sup>3</sup> If Assumption 3 holds, then, for any true effect  $\theta \in \mathbb{R}$ , and any level of potential bias  $\rho \in \mathbb{R}$ , there exists some continuous function  $h : \mathcal{X} \mapsto \mathbb{R}$  such that if the true DGP satisfies,*

$$Y_i = Y(D_i, X_i, \epsilon_i) = \alpha + \theta D_i + h(X_i) + \epsilon_i, \quad (\epsilon_i \perp\!\!\!\perp D_i, X_i).$$

*Then, the bias of the linear IV estimand for the treatment effect is given by  $\theta - \theta_{IV} = \rho$ .*

The proof of Theorem 1 is constructive – we propose an adversarial  $h(\cdot)$  which has a nonlinear term proportional to  $\rho$ . As we raise  $\rho$ , increasing the nonlinearity of the function  $h(\cdot)$ , the bias of the IV estimand  $\theta_{IV}$  for the true treatment effect  $\theta$  increases. This demonstrates the close relationship between the nonlinearity of the true structural function  $Y(D_i, X_i, \epsilon_i)$  in  $X_i$  and the bias of the linear IV estimand.

*Remark 1* (Causal consistency). The adversarial bias result in Theorem 1 is closely related to Proposition 3 of [Andrews et al. \(2023\)](#), which establishes that strong exclusion of the instrument  $f(X, Z)$  is both necessary and sufficient for the IV estimand  $\theta_{IV}$  to be approximately causally consistent. In the class of DGPs considered in Theorem 1, the analyst’s model  $\tilde{Y}(D_i, X_i, \tilde{\epsilon}_i)$  is “causally correctly specified,” in the sense that the structural parameter  $\theta$  can be described by the estimate  $\hat{\theta}_{IV}$  they report. However, because the analyst uses only covariates  $(X_1, X_2)$  and not the excluded variable  $Z$  in their instrument  $f(X, Z)$ , the causal interpretation of  $\hat{\theta}_{IV}$  can be arbitrarily misleading.

**Corollary 1** (Product instrument). *Let Assumption 3 hold, and assume the analyst uses an instrument  $f(X, Z) = X_1 \cdot X_2$  that is a product of covariates  $(X_1, X_2)$  to estimate the IV model in Equation 1. If the true DGP satisfies,*

$$Y_i = Y(D_i, X_i, \epsilon_i) = \alpha + \theta D_i + \pi_1 X_{i,1} + \pi_2 X_{i,2} + \rho X_{i,1} X_{i,2} + \epsilon_i, \quad (\epsilon_i \perp\!\!\!\perp D_i, X_i).$$

---

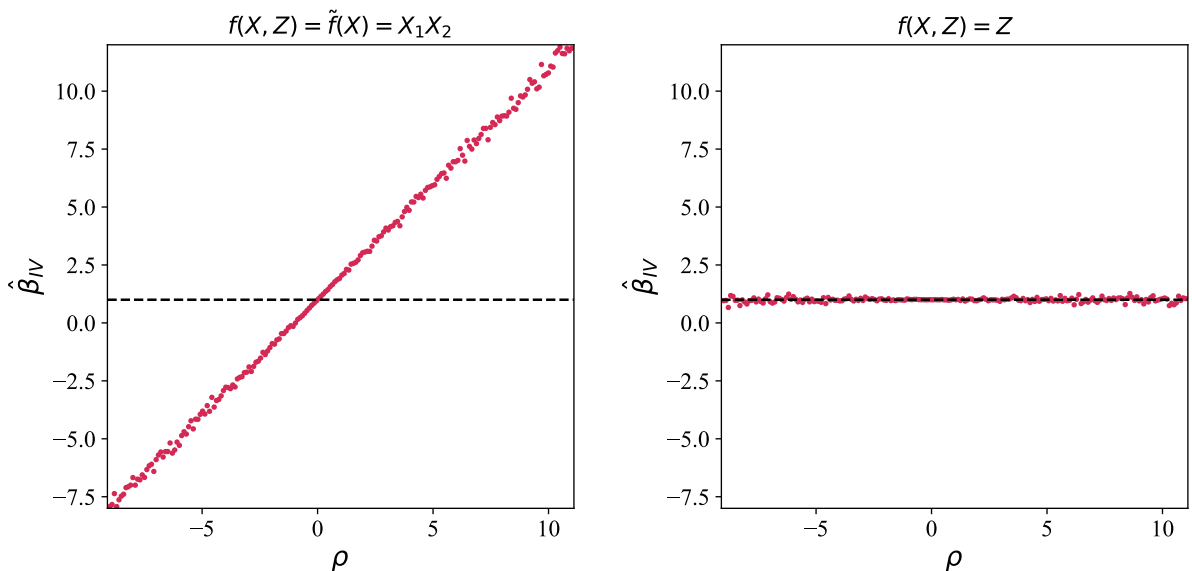
<sup>3</sup>Recall from Lemma 2, that this is necessary for  $\theta_{IV}$  to be well-defined.

Then, the bias of  $\theta_{IV}$  for the treatment effect is given by,

$$\theta - \theta_{IV} = \rho \times \frac{\mathbb{E}\{D(X_1 X_2 - X^\top \mathbb{E}\{XX^\top\})\mathbb{E}\{X(X_1 \cdot X_2)\}\}}{\text{Var}(X_1 X_2 - X^\top \mathbb{E}\{XX^\top\})\mathbb{E}\{X(X_1 \cdot X_2)\}}$$

For the product instrument  $\tilde{f}(X) = X_1 \cdot X_2$  used in the papers cited in the introduction, Corollary 1 characterizes the bias of the IV estimand for a single linear DGP. It shows that the bias is proportional to the coefficient on the interaction term  $X_1 \cdot X_2$ . As the structural function includes increasingly nonlinear interactions, the bias in the IV estimand  $\theta_{IV}$  grows accordingly. This result highlights that sizable violations of the exclusion restriction—leading to substantial bias—can arise when the interaction of excluded covariates strongly affects the outcome. For instance, if  $X_1$  is a binary variable such as gender, then a large coefficient on  $X_1 \cdot X_2$  in the structural function  $Y(D_i, X_i, \epsilon_i)$  suggests that the relationship between  $X_2$  and the outcome differs markedly between men and women. To visualize this connection, Figure 1 plots IV estimates  $\hat{\beta}_{IV}$  against data generated from the model in Corollary 1, illustrating how this misspecification undermines the interpretability of the IV estimand.

Figure 1: Simulation of IV estimates for the DGP  $Y_i = \alpha + \theta D_i + \pi_1 X_{i,1} + \pi_2 X_{i,2} + \rho X_{i,1} X_{i,2} + \epsilon_i$ .



(a) Instrumental variable that is nonlinear function of included variables  $(X_1, X_2)$  (b) Instrumental variable that is function of excluded variable  $Z$

Notes: Each dot represents an estimate of  $\theta_{IV}$  from Definition 1 using 2SLS with 1,000 generated observations. The dashed line indicates the true effect  $\theta = 1$  of  $D$  on  $Y$  in the simulations.

Examining panel (a) of Figure 1 illustrates the sensitivity of these IV estimates to misspecification in the true relationship between  $X$  and  $Y$ . As Corollary 1 implies, we see that by increasing the coefficient on the interaction term  $X_1 \cdot X_2$  in the true structural equation  $Y(D_i, X_i, \epsilon_i)$ , we can inflate the bias of the IV estimate  $\hat{\theta}$  while holding the true treatment effect  $\theta$  constant. In contrast, the consistency of the IV estimates in panel (b) underscores the robustness of the IV estimator against misspecification in the true relationship between  $X$  and  $Y$  for partially linear DGPs when using a product instrument.

*Remark 2 (OVB).* The bias from Theorem 1 can be interpreted as a form of omitted variable bias

(OVB), resulting from the exclusion of the nonlinear effect of the covariates  $X$  on the outcome  $Y$  from the analyst’s model.

*Remark 3* (Rich models). Theorem 1 shows that the linear IV estimand can have arbitrary bias even for simple DGPs with homogeneous linear effects. The proof further implies that estimates can also be biased for richer structural functions  $Y(X_i, D_i, \epsilon_i)$  that are nonlinear in  $X$ .

Remark 2 gives an alternative interpretation of the bias for IV estimators. Remark 3 underlines the generality of Theorem 1: for IV estimands that rely on instruments of covariates  $\tilde{f}(X)$ , if the analyst’s model does not capture how the covariates  $X$  enter the true structural function  $Y(D_i, X_i, \epsilon_i)$ , then the IV estimand can be very biased.

We have shown that the practice of using instruments constructed from nonlinear functions of covariates  $f(X, Z) = \tilde{f}(X)$  can misbehave in interesting ways theoretically. In Section 4, we demonstrate that this behavior is important in empirical work.

## 4 Application

Researchers often have strong economic reasoning to justify their proposed model and may believe that their model matches the true DGP. In these cases, Claim 1 (IV validity) shows that the IV estimand  $\theta_{IV}$  gives the average causal effect of  $D$  on  $Y(D_i, X_i, \epsilon_i)$ .

However, it is also often the case in the literature that non-nested models are used to estimate the same treatment effect. In these instances, every model cannot match the true DGP, and therefore some of these models must be misspecified.

For example, recall from Section 1 that Aladangady (2017) uses IV estimation to investigate the effect of log changes in house price  $D$  on log consumer expenditure  $Y$  in the United States using a model of the form,

$$Y_i = \tilde{Y}(D_i, X_i, \tilde{\epsilon}_i) = \tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i,$$

for  $X \in \mathbb{R}^2$ . Real interest rates  $X_1$  and housing supply elasticity measures  $X_2$  as a function of future expectations  $X_3$  are included as covariates.

Similarly, Disney et al. (2010) estimates the effect of log house price growth  $D$  on log consumption  $Y$  in the United Kingdom, but includes interest rates  $X_1$  and future expectations  $X_3$  in their life-cycle model for expenditure. They use ordinary least-squares to estimate the model,

$$Y_i = \tilde{Y}(D_i, X_i, \tilde{\epsilon}_i) = \tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top \log(X_i) + \tilde{\epsilon}_i,$$

for  $X \in \mathbb{R}^7$ ,<sup>4</sup> where we write  $\log(X)$  to be  $\log(X_i) = (\log(X_1), \log(X_3), \dots)$ .

Both authors justify their models with economic intuition. Still, they do not agree on the functional form of consumer expenditure, or the way the covariates  $X$  enter the model  $\tilde{Y}(D_i, X_i, \epsilon_i)$ . Theorem 1 (model bias) shows that if the author’s model *does not* match the true structural function, IV estimators that use an instrument constructed from covariates can have a high level of bias. Therefore, it is important

---

<sup>4</sup>The authors also include gross household income, passive savings, estimated unanticipated house growth, and household characteristics as covariates.

to understand the sensitivity of these estimates to how the covariates  $X$  affect the outcome  $Y$  in cases when there are several reasonable choices to model the DGP.

Table 1 gives additional examples of competing models used in the literature. For each of these examples, we compare: on one hand, (i) a linear IV model that relies on the product instrument  $\tilde{f}(X) = X_1 \cdot X_2$ ; and on the other hand (ii) a richer nonlinear model proposed in that literature to estimate the same treatment effect. For each pair of papers, the linear IV model is written on the top row and the non-nested alternative model is given on the bottom row.

Table 1: Alternative choices for  $\tilde{Y}(D_i, X_i, \tilde{\epsilon}_i)$

Paper	Model	Outcome $Y$	Endo. Variable $D$
<a href="#">Aladangady (2017)</a>	$Y_i = \tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i$	Log consumption	Log house wealth
<a href="#">Disney et al. (2010)</a>	$Y_i = \tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top \log(X_i) + \tilde{\epsilon}_i$	Log consumption	Log house wealth
<a href="#">Munshi and Rosenzweig (2016)</a>	$Y_i = \tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i$	Migration indicator	Level of wealth
<a href="#">Abramitzky (2008)</a>	$Y_i = \Phi(\tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i) + \tilde{\epsilon}_i$	Migration indicator	Level of wealth
<a href="#">Helmets et al. (2017)</a>	$Y_i = \tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i$	Patent count	Network size
<a href="#">Chang et al. (2006)</a>	$Y = \exp\{\tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i\} + \tilde{\epsilon}_i$	Patent count	Network size
<a href="#">Liu et al. (2023)</a>	$Y_i = \tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i$	Indicator of child	Value of home
<a href="#">Clark and Ferrer (2019)</a>	$Y_i = \frac{\exp\{\tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i\}}{1 + \exp\{\tilde{\alpha}_0 + \tilde{\alpha}_1 D_i + \tilde{\pi}^\top X_i\}} + \tilde{\epsilon}_i$	Indicator of child	Value of home

Notes: Of the papers cited in the introduction that rely on IV estimation using the product instrument  $\tilde{f}(X) = X_1 \cdot X_2$ , Table 1 gives an overview of the alternative models used among papers cited in their literature review.

Table 1 shows that cases often arise when there is no consensus about the functional form of the true structural function  $Y(D_i, X_i, \epsilon_i)$ . Therefore, it is likely that the true DGP is nonlinear in the covariates  $X$ .

To assess the sensitivity of linear IV models that use product instruments, Table 2 performs a semi-synthetic exercise for each pair of papers in Table 1. Specifically, for each pair of papers, we interpret the richer, nonlinear specification as the true structural equation and calibrate the DGP to match that model. We restrict one degree of freedom for the parameters  $(\alpha^\top, \pi^\top)$  by forcing  $\theta = \mathbb{E}\left\{\frac{\partial}{\partial D_i} Y(D_i, X_i, \epsilon_i)\right\} = 1$ . We then use 2SLS to estimate  $\hat{\theta}$  in the linear model IV model from Equation 1.

For example, using the structural model offered by [Disney et al. \(2010\)](#) we construct synthetic draws of  $(Y_i, D_i, X_i)$  such that  $\theta = \alpha_1 = 1$ . Then, we fit the linear IV model used by [Aladangady \(2017\)](#) and

compare how the IV estimate  $\hat{\theta}$  compares to  $\theta$  in the synthetic structural model. Full details are given in Appendix B.

## 4.1 Results

Table 2 shows estimates of  $\hat{\theta}$  from Equation 1 for the estimand  $\theta_{IV}$  in Definition 1 using linear IV estimation.

Table 2: Estimates of  $\hat{\theta}$  from Equation 1 using 2SLS

Paper	Model: $\tilde{Y}(D_i, X_i, \tilde{\epsilon}_i) =$	DGP: $Y(D_i, X_i, \epsilon_i) =$	$\hat{\theta}$
<a href="#">Aladangady (2017)</a>	$\tilde{\alpha} + \tilde{\theta}D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i$	$\alpha_0 + \alpha_1 D_i + \pi^\top \log(X_i) + \epsilon_i$	-1.16 (-2.29, -0.02)
<a href="#">Munshi and Rosenzweig (2016)</a>	$\tilde{\alpha} + \tilde{\theta}D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i$	$\Phi(\alpha_0 + \alpha_1 D_i + \pi^\top X_i) + \epsilon_i$	-2.42 (-3.77, -1.06)
<a href="#">Helmerts et al. (2017)</a>	$\tilde{\alpha} + \tilde{\theta}D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i$	$\exp\{\alpha_0 + \alpha_1 D_i + \pi^\top X_i\} + \epsilon_i$	0.24 (0.18, 0.30)
<a href="#">Liu et al. (2023)</a>	$\tilde{\alpha} + \tilde{\theta}D_i + \tilde{\pi}^\top X_i + \tilde{\epsilon}_i$	$\frac{\exp\{\alpha_0 + \alpha_1 D_i + \pi^\top X_i\}}{1 + \exp\{\alpha_0 + \alpha_1 D_i + \pi^\top X_i\}} + \epsilon_i$	0.35 (0.30, 0.41)

Notes: Each row estimates  $\hat{\theta}$  from Equation 1 using the IV model from the corresponding paper. Data is generated following the richer, nonlinear model, i.e. the structural function listed under DGP.  $\hat{\theta}$  reports the average estimate across 1000 simulations for the estimand  $\theta_{IV}$  in Definition 1. 95% confidence intervals for  $\hat{\theta}$  across simulations are reported in parenthesis. For each DGP,  $(\alpha^\top, \pi^\top)$  are chosen while maintaining the condition  $\theta = \mathbb{E}\left\{\frac{\partial}{\partial D_i} Y(D_i, X_i, \epsilon_i)\right\} = 1$ . Details of this calculation are at the end of Appendix B.

In Table 2, we observe that for all four pairs of papers, a 95% confidence interval for the estimated treatment effect  $\hat{\theta}$  does not include the ground truth of  $\theta = 1$ , i.e. the average partial treatment effect (APE) in the true structural function  $Y(D_i, X_i, \epsilon_i)$ . For instance, IV estimates tend to be negative when using the models from [Aladangady \(2017\)](#), where the DGP is linear with log-transformed covariates  $X$ . Meanwhile, for the models used by [Helmerts et al. \(2017\)](#) and [Liu et al. \(2023\)](#), where the true structural function follows an exponential and logit design, IV estimates based on instruments that are functions of covariates tend to be close to zero, underestimating the APE. This illustrates how the bias of the IV estimand can be very positive or negative when using product instruments, depending on the underlying DGP.

These findings underscore the sensitivity of these IV estimates  $\hat{\theta}$  to the functional form chosen to model the relationship between covariates  $X$  and the outcome  $Y$  and illustrate their bias under sensible choices for the DGP. Even for reasonable choices of  $Y(D_i, X_i, \epsilon_i)$ , IV estimates of the treatment effect  $\hat{\theta}$  with constructed instruments can differ substantially from the true APE and imply different causal conclusions.

Another natural question is how sensitive causal estimates are to the analyst’s modeling choices when applied to real economic data. To examine this, Table 3 explores how estimates of  $\theta$  vary when fitting different functional forms to data from published empirical studies. Specifically, we consider the subset of

papers published in the *American Economic Review* (AER) since 2014 that employ instrumental variable (IV) strategies relying on product instruments of the form  $\tilde{f}(X) = X_1 \cdot X_2$ .<sup>5</sup> For each of these studies,<sup>6</sup> we re-estimate the parameter  $\tilde{\theta}$  using the linear IV model,

$$Y_i = \tilde{Y}(D_i, X_i, \tilde{\epsilon}_i) = \tilde{\alpha}_0 + \tilde{\theta}D_i + \tilde{\pi}^\top h(X_i) + \tilde{\epsilon}_i, \quad (2)$$

for various nonlinear functions  $h(\cdot)$  and investigate the range of estimates produced.

Table 3: Estimates of  $\tilde{\theta}$  from Equation 2.

$h(X_i) =$	Paper		
	(1)	(2a)	(2b)
$(X_{i,1}, \dots, X_{i,q})^\top$	1.34 (-0.30, 2.99)	5.40 (3.76, 7.05)	-0.72 (-2.37, 0.92)
$(\exp\{X_{i,1} - \bar{X}_1\}, \dots, \exp\{X_{i,q} - \bar{X}_q\})^\top$	1.55 (-0.16, 3.25)	1.58 (-0.03, 3.19)	-0.73 (-2.42, 0.96)
$(\log(1 + X_{i,1}), \dots, \log(1 + X_{i,q}))^\top$	5.54 (-5.12, 16.21)	1.41 (-0.35, 3.18)	-0.60 (-2.24, 1.04)
$\left(\frac{2 \exp\{2X_{i,1}\} - 2 \exp\{-2X_{i,1}\}}{2.5 \exp\{2.5X_{i,1}\} + 2.5 \exp\{-2.5X_{i,1}\}}, \dots\right)^\top$	0.39 (-0.98, 1.76)	0.65 (-11.52, 12.81)	-5.46 (-13.36, 2.43)

Notes: Each row estimates the linear IV estimand  $\tilde{\theta}$  from Equation 2 for the specification listed under  $h(X_i) =$  using 2SLS. We standardized each column by dividing it by the standard deviation of the linear estimate, which corresponds to the IV estimate using the authors' model  $h(X_i) = (X_{i,1}, \dots, X_{i,q})^\top$  for  $q = \dim(X_i)$ . Asymptotic 95% confidence intervals using heteroskedasticity-robust standard errors are in parenthesis.

Within each paper, we find that the estimated causal effect  $\tilde{\theta}$  is sensitive to the choice of functional form for  $h(X_i)$ . In paper (1), for instance, a 95% confidence interval based on the reported linear specification does not include the estimate obtained using a log-transformed specification, where  $h(X_i) = (\log(1 + X_{i,1}), \dots, \log(1 + X_{i,q}))^\top$ . Similarly, in paper (2a), none of the estimates from nonlinear specifications lie within the 95% confidence interval derived from the original linear model.

These results demonstrate that the analyst's estimate of  $\tilde{\theta}$  can vary substantially depending on the functional form used to model the relationship between covariates  $X$  and the outcome  $Y$ . Even among plausible specifications of  $h(X_i)$ , the magnitude and statistical significance of the estimated treatment effect can differ markedly.

Together, this evidence highlights the lack of robustness of IV estimators based on product instruments to nonlinearities in how covariates  $X$  enter the structural function  $Y(D_i, X_i, \eta_i)$ . Substantive conclusions drawn from such models may be sensitive to modeling assumptions.

Overall, our results underscore the importance of functional form choices in IV models where instruments are functions of covariates,  $f(X, Z) = \tilde{f}(X)$ . The selected functional form can significantly influence how the treatment effect is interpreted.

<sup>5</sup>We only use papers for which the data is publicly available.

<sup>6</sup>Paper (1) is Rogall (2021) and paper (2) is Munshi and Rosenzweig (2016).

## 5 Recommendations for practice

In this paper, we have provided theoretical and empirical evidence showing that product instruments, or more generally, instruments that are transformations of covariates, can be very biased. Our framework shows that even when imposing homogeneous linear effects between an endogenous variable and outcome, these IV estimates cannot be interpreted as a causal effect under mild forms of misspecification. Theoretically, we show that there exists a simple DGP with constant linear effects that can lead to unbounded bias of the IV estimand. Empirically, we illustrate the sensitivity of IV estimates that use instruments constructed from nonlinear functions of covariates.

When a researcher does not have access to an excluded, exogenous variable, we recommend caution when discussing the causal interpretability of IV estimators. Alternatively, if the researcher believes their instrumental variable  $f(X, Z) = \tilde{f}(X)$  satisfies the IV assumptions in Section 2, it is important that they strongly defend their choice of model  $\tilde{Y}(\cdot)$ , i.e. the relationship between the outcome  $Y$  and covariates  $X$ , to defend the validity of their estimate.

## References

- Abramitzky, R. (2008). The Limits of Equality: Insights from the Israeli Kibbutz\*. *The Quarterly Journal of Economics*, 123(3):1111–1159.
- Aladangady, A. (2017). Housing Wealth and Consumption: Evidence from Geographically-Linked Microdata. *American Economic Review*, 107(11):3415–3446.
- Andrews, I., Barahona, N., Gentzkow, M., Rambachan, A., and Shapiro, J. M. (2023). Structural Estimation Under Misspecification: Theory and Implications for Practice.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455.
- Bettinger, E. P., Fox, L., Loeb, S., and Taylor, E. S. (2017). Virtual Classrooms: How Online College Courses Affect Student Success. *American Economic Review*, 107(9):2855–2875.
- Bharadwaj, P. (2015). Fertility and rural labor market inefficiencies: Evidence from India. *Journal of Development Economics*, 115:217–232.
- Blandhol, C., Bonney, J., Mogstad, M., and Torgovitsky, A. (2022). When is TSLS Actually LATE?
- Cagé, J., Hervé, N., and Mazoyer, B. (2022). Social Media Influence Mainstream Media: Evidence from Two Billion Tweets.
- Chang, S.-J., Chung, C.-N., and Mahmood, I. P. (2006). When and How Does Business Group Affiliation Promote Firm Innovation? A Tale of Two Emerging Economies. *Organization Science*, 17(5):637–656.
- Clark, J. and Ferrer, A. (2019). The effect of house prices on fertility: evidence from Canada. *Economics*, 13(38):1–33.

- Crawford, R. and Simpson, P. (2020). The impact of house prices on pension saving in early adulthood. Working Paper W20/38, IFS Working Paper.
- Disney, R., Gathergood, J., and Henley, A. (2010). House Price Shocks, Negative Equity, and Household Consumption in the United Kingdom. *Journal of the European Economic Association*, 8(6):1179–1207.
- Gao, W. Y. and Wang, R. (2023). IV Regressions without Exclusion Restrictions.
- Ghose, A., Lee, H. A., Oh, W., and Son, Y. (2024). Leveraging the Digital Tracing Alert in Virus Fight: The Impact of COVID-19 Cell Broadcast on Population Movement. *Information Systems Research*, 35(2):570–589.
- Helmers, C., Patnam, M., and Rau, P. R. (2017). Do board interlocks increase innovation? Evidence from a corporate governance reform in India. *Journal of Banking & Finance*, 80:51–70.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475.
- Liu, H., Liu, L., and Wang, F. (2023). Housing wealth and fertility: evidence from China. *Journal of Population Economics*, 36(1):359–395.
- Munshi, K. and Rosenzweig, M. (2016). Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap. *American Economic Review*, 106(1):46–98.
- Rogall, T. (2021). Mobilizing the Masses for Genocide. *American Economic Review*, 111(1):41–72.
- Shao, L. (2014). Estimating the relationship between calculated financial need and actual aid received using quarter of birth instruments. *Economics of Education Review*, 42:165–174.

# Appendix for “Constructing Instruments as a Function of Included Covariates”

Moses Stewart, *Harvard University*<sup>1</sup>

## A Proofs from main text

*Proof of Lemma 1.*

To prove the forward direction, let  $\text{Var}(f_{\perp}) = \mathbb{E}\{f_{\perp}^2\} = 0$ . This implies that  $f_{\perp} = 0$  almost surely. However, since  $X^{\top}\mathbb{E}\{XX^{\top}\}^{-1}\mathbb{E}\{X\tilde{f}(X)\}$  is a linear projection of  $\tilde{f}(X)$  onto  $X$ , this implies that  $\tilde{f}(X)$  is a linear function of each  $X_j \in X$ .

Similarly to prove the backwards direction, let  $\text{Var}(f_{\perp}) = \mathbb{E}\{f_{\perp}^2\} \neq 0$ . Since  $X^{\top}\mathbb{E}\{XX^{\top}\}^{-1}\mathbb{E}\{X\tilde{f}(X)\}$  is a linear projection of  $\tilde{f}(X)$  onto  $X$ , this implies that  $f_{\perp}$  is not equal to 0 almost surely, so it is not a linear function of each  $X_j \in X$ . The result follows.  $\square$

*Proof of Lemma 2.*

Using the tower property, notice that,

$$\text{Cov}(\tilde{f}(X_i), \tilde{\epsilon}_i) = \mathbb{E}\{\tilde{\epsilon}_i \tilde{f}(X_i)\} - \mathbb{E}\{\tilde{\epsilon}_i\}\mathbb{E}\{\tilde{f}(X_i)\} = \mathbb{E}\{\mathbb{E}\{\tilde{\epsilon}_i | X_i\} \tilde{f}(X_i)\} = 0.$$

$\square$

*Proof of Claim 1.*

We assume that the true DGP matches the analyst’s model, so,

$$Y_i = Y(D_i, X_i, \epsilon_i) = \alpha_i + D_i\theta + \pi^{\top}X_i + \epsilon_i.$$

Then since under Assumption 2, we have that  $\text{Cov}(\epsilon_i, f_{\perp}) = 0$  we can write,

$$\begin{aligned} \theta_{IV} &= \frac{\text{Cov}(Y, f_{\perp})}{\text{Cov}(D, f_{\perp})} = \frac{\text{Cov}(Y(D, X, \epsilon), f_{\perp})}{\text{Cov}(D, f_{\perp})} \\ &= \frac{\text{Cov}(\alpha, f_{\perp})}{\text{Cov}(D, f_{\perp})} + \frac{\text{Cov}(\theta D, f_{\perp})}{\text{Cov}(D, f_{\perp})} + \frac{\text{Cov}(X\pi, f_{\perp})}{\text{Cov}(D, f_{\perp})} + \frac{\text{Cov}(\epsilon, f_{\perp})}{\text{Cov}(D, f_{\perp})} \\ &= \theta \frac{\text{Cov}(D, f_{\perp})}{\text{Cov}(D, f_{\perp})} + \pi \frac{\text{Cov}(X, f_{\perp})}{\text{Cov}(D, f_{\perp})} \\ &= \theta. \end{aligned}$$

The assertion that  $\text{Cov}(X_i, f_{\perp}) = 0$  in the last line follows since  $f_{\perp} = \tilde{f}(X) - X^{\top}\mathbb{E}\{XX^{\top}\}^{-1}\mathbb{E}\{X\tilde{f}(X)\}$  are the residuals from the projection of  $\tilde{f}(X)$  onto  $X$ , and  $\text{Cov}(D, f_{\perp}) \neq 0$  follows from Assumption 3 (relevance).  $\square$

*Proof of Theorem 1.*

To prove Theorem 1, we will construct a continuous function  $h(X_i)$  such that the estimand  $\theta_{IV} = \rho$ . Let

---

<sup>1</sup>E-mail: mosessstewart@g.harvard.edu

$f_{\perp} = \tilde{f}(X) - X\mathbb{E}\{XX^{\top}\}^{-1}\mathbb{E}\{X\tilde{f}(X)\}$  be the residuals from a projection of the instrumental variable  $\tilde{f}(X)$  onto  $X$  and  $g(\rho) = \frac{\text{Cov}(D, f_{\perp})}{\text{Var}(f_{\perp})}\rho$ . Consider the function,

$$Y_i = Y(D_i, X_i, \epsilon_i) = \alpha + \theta D_i + h(X_i) + \epsilon_i$$

for  $h(X_i) = X_i\pi + g(\rho)\tilde{f}(X_i)$ .

Then, we can write the IV estimand as,

$$\begin{aligned} \theta_{IV} &= \frac{\text{Cov}(Y, f_{\perp})}{\text{Cov}(D, f_{\perp})} = \frac{\text{Cov}(\alpha + \theta D_i + h(X_i) + \epsilon_i, f_{\perp})}{\text{Cov}(D, f_{\perp})} \quad (\text{I}) \\ &= \frac{\text{Cov}(\alpha, f_{\perp})}{\text{Cov}(D, f_{\perp})} + \frac{\text{Cov}(\theta D, f_{\perp})}{\text{Cov}(D, f_{\perp})} + \frac{\text{Cov}(X\pi, f_{\perp})}{\text{Cov}(D, f_{\perp})} + \frac{\text{Cov}(g(\rho)\tilde{f}(X_i), f_{\perp})}{\text{Cov}(D, f_{\perp})} + \frac{\text{Cov}(\epsilon, f_{\perp})}{\text{Cov}(D, f_{\perp})} \\ &= \theta \frac{\text{Cov}(D, f_{\perp})}{\text{Cov}(D, f_{\perp})} + \frac{\text{Cov}(X, f_{\perp})}{\text{Cov}(D, f_{\perp})}\pi + g(\rho) \frac{\text{Var}(f_{\perp})}{\text{Cov}(D, f_{\perp})} \quad (\text{II}) \\ &= \theta + g(\rho) \frac{\text{Var}(f_{\perp})}{\text{Cov}(D, f_{\perp})} \quad (\text{III}) \\ &= \rho. \end{aligned}$$

Where (I) follows because we assumed the DGP satisfies  $Y = \alpha + \theta D_i + h(X_i) + \epsilon_i$ .  $\text{Cov}(\epsilon, f_{\perp}) = 0$  in line (II) follows since we impose  $\epsilon \perp X$ .  $\text{Cov}(X, f_{\perp}) = 0$  in line (III) follows since  $f_{\perp} = \tilde{f}(X) - X^{\top}\mathbb{E}\{XX^{\top}\}^{-1}\mathbb{E}\{X\tilde{f}(X)\}$  are the residuals from the projection of  $\tilde{f}(X)$  onto  $X$ , and  $\text{Cov}(D, f_{\perp}) \neq 0$  follows from Assumption 3 (relevance).  $\square$

## B Simulation details

### B.1 Data generating process

We consider four DGPs corresponding to nonlinear models used in Table 2: the **probit** model (Abramitzky (2008)), the **exponential** model (Chang et al. (2006)), the **logit** model (Clark and Ferrer (2019)), and the **log-linear** model (Disney et al. (2010)).

For the first three models, we generate the data following

$$(D_i, X_{i,1}, X_{i,2}, \epsilon_i)^{\top} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma) \quad \text{for } \Sigma = \begin{bmatrix} \sigma_d^2 & \rho & \rho & 0 \\ \rho & \sigma_{x1}^2 & \rho & 0 \\ \rho & \rho & \sigma_{x2}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\epsilon}^2 \end{bmatrix}$$

Using these draws, we define the outcome variable  $Y_i$  as follows:

1. **Probit** (Abramitzky (2008)): Define  $p_i = \Phi(\alpha_1 D_i + X_{i1} + X_{i2})$  and draw  $Y_i \sim \text{Bern}(p_i)$ .
2. **Exponential** (Chang et al. (2006)): Set  $Y_i = \exp(\alpha_1 D_i + X_{i1} + X_{i2}) + \epsilon_i$ .
3. **Logit** (Clark and Ferrer (2019)) Define  $p_i = \frac{\exp\{\alpha_1 D_i + X_{i,1} + X_{i,2}\}}{1 + \exp\{\alpha_1 D_i + X_{i,1} + X_{i,2}\}}$  and draw  $Y_i \sim \text{Bern}(p_i)$ .

For the **log-linear** (Disney et al. (2010)) specification, we simulate a common latent factor  $Z_i \sim \mathcal{N}(0, 1)$  and independent Gamma random variables  $G_{i,1}, G_{i,2} \sim \text{Gamma}(\mu_k, \sigma_k^2 - \rho)$ , where the Gamma distribution is parameterized by its mean and variance. We then generate the variables as,

$$X_{i1} = G_{i1} + \sqrt{\rho}Z_i, \quad X_{i2} = G_{i2} + \sqrt{\rho}Z_i, \quad D_i = \delta_i + \sqrt{\rho}Z_i, \quad \text{with } \delta_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_d^2 - \rho).$$

Finally, for  $\epsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$  we define the outcome as:

$$Y_i = D_i + \pi_1 \log X_{i1} + \pi_2 \log X_{i2} + \epsilon_i$$

## B.2 $\alpha$ -Calibration

Let  $g(\cdot)$  denote the structural function used in the DGP. To standardize the marginal effect of  $D_i$ , we calibrate the parameter  $\alpha_1$  so that the average partial derivative of  $g(\alpha_1 D_i + X_{i1} + X_{i2})$  with respect to  $D_i$  equals one.

Since the sum  $\alpha_1 D_i + X_{i1} + X_{i2}$  is distributed as  $W \sim \mathcal{N}(0, \alpha_1^2(\sigma_d^2 + 2\rho) + \sigma_{x1}^2 + \sigma_{x2}^2)$ , we write,

$$g(\alpha_1 D_i + X_{i1} + X_{i2}) \sim g(W), \quad \text{so} \quad \frac{\partial}{\partial D_i} g(\alpha_1 D_i + X_{i1} + X_{i2}) \sim \alpha_1 g'(W).$$

We therefore choose  $\alpha_1$  to satisfy  $\mathbb{E}\{\alpha_1 g'(W)\} = 1$  by minimizing the objective  $(\alpha_1 - \mathbb{E}\{g'(W)\})^{-1}$ . This calibration ensures that the simulated average partial effect of the treatment variable  $D_i$  is normalized across models.

## C Weak Causality

### C.1 Nonlinear IV and LATE interpretation

In this section, we generalize the theoretical results in Blandhol et al. (2022) to cover continuous random variables  $(Y, D, Z, X)$  and non-linear treatment effects. We then use the results to analyze the local average treatment effects (LATE) interpretation of IV estimators of instruments  $f(X, Z) = \tilde{f}(X)$  that are nonlinear functions of exogenous variables.

*Definition 2* (Weakly causal).  $\beta$  is *weakly causal* if both of the following statements are true,

- a If  $Y(d, x, \epsilon)$  is non-decreasing in  $d$  for all  $d \in \mathcal{D}$  and  $x \in \mathcal{X}$ , then  $\beta \geq 0$
- b If  $Y(d, x, \epsilon)$  is non-increasing in  $d$  for all  $d \in \mathcal{D}$  and  $x \in \mathcal{X}$ , then  $\beta \leq 0$

Definition 2 is a minimal requirement for an estimand  $\beta$  to reflect the causal effect of  $D$  on  $Y$ . It says that if the causal effect of the treatment has the same sign for individual, then the summary estimand  $\beta$  also has that sign. An estimand  $\beta$  which fails to satisfy Definition 2 becomes uninterpretable as a measure of the treatment effect  $D$  on the outcome  $Y$ . Note that this definition of weak causality based on the potential outcomes implies the definition in Blandhol et al. (2022).

**Assumption 4** (Instrument monotonicity). For the potential endogenous variable function  $D_i = D(X_i, Z_i, v_i)$  and instrument  $f(X_i, Z_i)$  assume either,

a  $D(x_i, z_i, v_i)$  weakly increases as  $f(x_i, z_i)$  increases for all  $z \in \mathcal{Z}$  and  $x \in \mathcal{X}$

b  $D(x_i, z_i, v_i)$  weakly decreases as  $f(x_i, z_i)$  increases for all  $z \in \mathcal{Z}$  and  $x \in \mathcal{X}$

Assumption 4 requires that there is a consistent relationship between the endogenous variable  $D$  and the instrument  $f(X, Z)$ . It says that as the observed instrument  $f(X_i, Z_i)$  increases, so does the endogenous variable  $D_i$ . Using assumption 4, we can arrive at proposition 4 of Blandhol et al. (2022) without assuming discrete random variables or constant, linear treatment effects.

**Proposition 1** (Weakly Causal). *Let Assumptions 1, 2, 3, and 4 hold. Then,  $\theta_{IV}$  is weakly causal if and only if  $\mathbb{E}\{f_{\perp} | X\} = 0$ .*

**Corollary 2** (model bias). *For any instrumental variable  $f(X, Z) = \tilde{f}(X)$  that is a function exclusively of included covariates  $X$ ,  $\theta_{IV}$  is not weakly causal.*

Corollary 2 gives an alternate result to Theorem 1 that shows how IV estimands of instrumental variables  $\tilde{f}(X)$  that are only functions of included covariates lack causal interpretability. It relies on the failure of the ‘‘saturated covariates’’ condition ( $\mathbb{E}\{f_{\perp} | X\} = 0$ ) for IV estimands which use an instrument constructed from a nonlinear function of included covariates. Corollary 2 implies that these estimators cannot be interpreted as LATE.

## C.2 Proofs

**Lemma 3** (FKG Inequality). *Let  $(X, \epsilon, \eta)$  be mutually independent random variables, and let  $g(\cdot, \cdot) : \mathbb{R}^2 \mapsto \mathbb{R}$  and  $h(\cdot, \cdot) : \mathbb{R}^2 \mapsto \mathbb{R}$  be non-decreasing functions in their first arguments. Then, it follows that  $\text{Cov}(g(X, \epsilon), h(X, \eta)) \geq 0$ . Similarly, if  $g(\cdot, \cdot)$  is non-decreasing and  $h(\cdot, \cdot)$  is non-increasing, then  $\text{Cov}(g(X, \epsilon), h(X, \eta)) \leq 0$ .*

*Proof of Lemma 3.*

First, we show the result for non-decreasing functions  $g(\cdot)$  and  $h(\cdot)$ .

Define  $X_1, X_2 \stackrel{\text{iid}}{\sim} X$ ,  $\eta_1, \eta_2 \stackrel{\text{iid}}{\sim} \eta$ , and  $\epsilon_1, \epsilon_2 \stackrel{\text{iid}}{\sim} \epsilon$ . Then notice that using the tower property,

$$\begin{aligned} & \mathbb{E}\{(g(X_1, \epsilon_1) - g(X_2, \epsilon_2))(h(X_1, \eta_1) - h(X_2, \eta_2))\} \\ &= \mathbb{E}\{\mathbb{E}\{(g(X_1, \epsilon_1) - g(X_2, \epsilon_2))(h(X_1, \eta_1) - h(X_2, \eta_2)) | X_1 > X_2\}\} \\ &\geq \mathbb{E}\{\mathbb{E}\{(g(X_2, \epsilon_1) - g(X_2, \epsilon_2))(h(X_2, \eta_1) - h(X_2, \eta_2)) | X_1 > X_2\}\} \quad (\text{I}) \\ &= 0. \end{aligned}$$

Where (I) follows from the monotonicity inequality, which says if the random variable  $Y_1 \geq Y_2$  then  $\mathbb{E}\{Y_1\} \geq \mathbb{E}\{Y_2\}$ . Noticing that  $g(\cdot, \cdot)$  and  $h(\cdot, \cdot)$  are non-decreasing in their first argument and letting  $g(X_1, \epsilon_1), h(X_1, \eta_1) = Y_1$  and  $g(X_2, \epsilon_2), h(X_2, \eta_2) = Y_2$  gives the result. Next, we write,

$$\begin{aligned} & \mathbb{E}\{(g(X_1, \epsilon_1) - g(X_2, \epsilon_2))(h(X_1, \eta_1) - h(X_2, \eta_2))\} \\ &= \mathbb{E}\{g(X_1, \epsilon_1)h(X_1, \eta_1)\} - \mathbb{E}\{g(X_1, \epsilon_1)\}\mathbb{E}\{h(X_2, \eta_2)\} - \mathbb{E}\{g(X_2, \epsilon_2)\}\mathbb{E}\{h(X_1, \eta_1)\} + \mathbb{E}\{g(X_2, \epsilon_2)h(X_2, \eta_2)\} \\ &= 2\mathbb{E}\{g(X, \epsilon)h(X, \eta)\} - 2\mathbb{E}\{g(X, \epsilon)\}\mathbb{E}\{h(X, \eta)\} \\ &= 2\text{Cov}(g(X, \epsilon), h(X, \eta)). \end{aligned}$$

So it follows that,

$$\text{Cov}(g(X, \epsilon), h(X, \eta)) = \frac{1}{2} \mathbb{E}\{(g(X_1, \epsilon_1) - g(X_2, \epsilon_2))(h(X_1, \eta_1) - h(X_2, \eta_2))\} \geq 0$$

When  $g(\cdot, \cdot)$  is non-decreasing and  $h(\cdot, \cdot)$  is non-increasing in their first arguments, then it follows that  $-h(\cdot, \cdot)$  is non-decreasing and we can write,

$$\text{Cov}(g(X, \epsilon), h(X, \eta)) = -\text{Cov}(g(X, \epsilon), -h(X, \eta)) \leq 0.$$

□

*Proof of Proposition 1.*

We show that  $\mathbb{E}\{f_{\perp} | X\} = 0$  implies  $\theta_{IV}$  is weakly causal. Assume that  $\mathbb{E}\{f_{\perp} | X\} = 0$ . Using Definition 1 and the law of total covariance, we can write,

$$\begin{aligned} \theta_{IV} &= \frac{\mathbb{E}\{Y f_{\perp}\}}{\mathbb{E}\{D f_{\perp}\}} = \frac{\text{Cov}(Y, f_{\perp})}{\text{Cov}(D, f_{\perp})} \\ &= \frac{\mathbb{E}\{\text{Cov}(Y, f_{\perp} | X)\} + \text{Cov}(\mathbb{E}\{Y | X\}, \mathbb{E}\{f_{\perp} | X\})}{\mathbb{E}\{\text{Cov}(D, f_{\perp} | X)\} + \text{Cov}(\mathbb{E}\{D | X\}, \mathbb{E}\{f_{\perp} | X\})} \\ &= \frac{\mathbb{E}\{\text{Cov}(Y, f(X, Z) | X)\}}{\mathbb{E}\{\text{Cov}(D, f(X, Z) | X)\}}. \end{aligned}$$

Now, under Assumption 1 (excludability), if  $D(x_i, z_i, v_i)$  weakly increases in  $f(x_i, z_i)$ , it follows that  $f(x_i, z_i)$  weakly increases in  $D$ . Therefore, we can express  $f(x_i, z_i)$  as a non-decreasing function in  $D$ , with an additional component that is independent of  $\epsilon$  when conditioned on  $X$  under Assumption 1 (excludability). We can do the same when  $D(x_i, z_i, v_i)$  weakly decreases in  $f(x_i, z_i)$ . Then, consider two cases,

1.  $Y(d, x, \epsilon)$  is non-decreasing in  $d$ . Then, under Assumption 4 (instrument monotonicity), if  $D(x_i, z_i, v_i)$  is weakly-increasing in  $f(x_i, z_i)$ , then Lemma 3 (FKG) implies that both  $\text{Cov}(Y, f(X, Z) | X) \geq 0$  and  $\text{Cov}(D, f(X, Z) | X) \geq 0$ , so  $\theta_{IV} \geq 0$ . If  $D(x_i, z_i, v_i)$  is weakly-decreasing in  $f(x_i, z_i)$ , then both  $\text{Cov}(Y, f(X, Z) | X) \leq 0$  and  $\text{Cov}(D, f(X, Z) | X) \leq 0$ , so  $\theta_{IV} \geq 0$ .
2.  $Y(d, x, \epsilon)$  is non-increasing in  $d$ . Then, under Assumption 4 (instrument monotonicity), if  $D(x_i, z_i, v_i)$  is weakly-increasing in  $f(x_i, z_i)$ , then Lemma 3 (FKG) implies that  $\text{Cov}(Y, f(X, Z) | X) \leq 0$  and  $\text{Cov}(D, f(X, Z) | X) \geq 0$ , so  $\theta_{IV} \leq 0$ . If  $D(x_i, z_i, v_i)$  is weakly-decreasing in  $f(x_i, z_i)$ , then  $\text{Cov}(Y, f(X, Z) | X) \geq 0$  and  $\text{Cov}(D, f(X, Z) | X) \leq 0$ , so  $\theta_{IV} \leq 0$ .

This shows that  $\theta_{IV}$  satisfies Definition 2 (weakly causal).

The proof given in Proposition 4 of Blandhol et al. (2022) shows that if  $\mathbb{E}\{f_{\perp} | X\} \neq 0$ , then weak causality is not satisfied. □

*Proof of Corollary 2.*

Recall from Definition 1,

$$\theta_{IV} = \frac{\mathbb{E}\{Y f_{\perp}\}}{\mathbb{E}\{D f_{\perp}\}} = \frac{\text{Cov}(Y, f_{\perp})}{\text{Cov}(D, f_{\perp})}.$$

Suppose that  $\tilde{f}(X)$  is polynomial in  $X$ , or linear in the basis  $\Phi(X)$ . Lemma 1 then implies that  $\text{Cov}(D, f_{\perp}) = 0$ , so  $\theta_{IV}$  is unidentified.

Therefore suppose that  $\tilde{f}(X)$  is nonlinear in  $X$ . Then,

$$\begin{aligned}\mathbb{E}\{\tilde{f}(X) \mid X = x\} &= \tilde{f}(X) \\ &\neq X\mathbb{E}\{X^{\top}X\}^{-1}\mathbb{E}\{X^{\top}\tilde{f}(X)\}.\end{aligned}$$

So Proposition 1 implies that  $\theta_{IV}$  is not weakly causal. □