Justifications for TSLS and a mostly harmless improvement

Jiafeng Chen[1], Daniel L. Chen[2], and Greg Lewis[3]

Abstract. What assumptions justify two-stage least-squares (TSLS) as an estimator of causal effects when there are covariates? We argue that an natural assumption—in addition to constant and linear effects—is one that we call *partial mean-independence*, and further that other seemingly natural restrictions that rationalize TSLS have various deficiencies. We then analyze this assumption from a semiparametric efficiency perspective and derive efficient and locally efficient estimators. These estimators take the form of plugging in an estimated optimal instrument, which is constrained to be linear in the covariates but may be non-linear in the instrumental variables. Under the TSLS-justifying assumptions that we propose, our approach amounts to allowing a more flexible first stage, thereby strengthening the relevance of instruments and delivering improvements to TSLS that we argue are user-friendly and mostly harmless.

Keywords: Two-stage least-squares, linear instrumental variables, partially linear regression

[1]Department of Economics, Harvard University and Harvard Business School.

[2]Toulouse School of Economics

[3]Amazon

E-mail addresses: jiafeng@stanford.edu; dlchen@nber.org; greglewis.work@gmail.com.

Date: May 10, 2024. This work previously appeared in the Machine Learning and Economic Policy Workshop at NeurIPS 2020 under the title "Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models," which is since heavily revised. The authors thank Isaiah Andrews, Tim Armstrong, Mike Droste, Bryan Graham, Jeff Gortmaker, Sendhil Mullainathan, Ashesh Rambachan, David Ritzwoller, Brad Ross, Jonathan Roth, Suproteem Sarkar, Neil Shephard, Rahul Singh, Jim Stock, Liyang Sun, Jann Spiess, Vasilis Syrgkanis, Chris Walker, Wilbur Townsend, and Elie Tamer for helpful comments.

1. **Introduction**

Two-stage least-squares (TSLS) is often used to estimate some causal effect in empirical work. Over 30% of all NBER working papers and top journal publications considered by Currie, Kleven and Zwiers (2020) include some discussion of instrumental variables, and TSLS is the workhorse estimator for research designs involving instruments. Almost a third of papers using TSLS, surveyed by Blandhol, Bonney, Mogstad and Torgovitsky (2022), explicitly interpret the resulting estimate as one for some treatment effect parameter.

However, there is considerable recent interest in the minimal set of assumptions under which TSLS actually has a causal interpretation. Of course, TSLS is justified under a linear structural model, which places stringent functional-form assumptions on the joint distribution of potential outcomes. The celebrated result by Angrist and Imbens (1995) shows that TSLS is in a sense "model-free": TSLS, even without assumptions on the potential outcomes, estimates a local average treatment effect. However, recent work (Blandhol *et al.*, 2022; Słoczyński, 2022) has pointed out that in empirical implementations that commonly arise—which are often more complex than in Angrist and Imbens (1995), TSLS often does not have reasonable model-free interpretations. In particular, when covariates are included linearly and when the covariates are thought to be necessary for identification, TSLS can estimate weighted averages of treatment effects with negative weights.

Therefore, if one would like to continue to use TSLS or to interpret existing TSLS estimates causally, then one should impose—and assess—functional form restrictions on the potential outcomes. What restrictions *do* justify TSLS with covariates? One could retreat to a textbook-style linear structural model, where the "structural errors" are either assumed to be uncorrelated with or mean-independent of the instruments. However, as we shall see, depending on which error assumption is imposed, such a model either *allows* nonlinear functions of covariates as valid instruments or *disallows* nonlinear functions of instruments as valid instruments. Neither is a reasonable feature.

This paper provides assumptions that justify TSLS with covariates and do not exhibit the above deficiencies of the classical structural equation assumptions. We show that a minimal assumption under these considerations is that the baseline potential outcome is *partially mean-independent* of the instrument given covariates—meaning that a partially linear regression of the potential outcome on nonlinear functions of the instruments and linear functions of the covariates returns a linear function of the covariates.

Having identified such an assumption, we then study the semiparametric properties of this model, applying the general analysis of sequential moment models (Chamberlain, 1992).

This analysis implies efficient and locally efficient[1] estimation procedures which provide efficency improvements over TSLS under the partial mean-independence assumption. Notably, it is possible that the instrument does not linearly predict the endogenous treatment well—leading to a weak instruments problem for TSLS—but nonlinear functions of the instrument do predict the endogenous treatment. Machine learning methods thus hold promise for effectively finding these nonlinear functions and rescuing otherwise weak instruments. Indeed, we show that implementations of these estimators—which may leverage complex machine learning methods—have the usual appealing asymptotic properties under notably weaker assumptions than are typical in semiparametric econometrics. Our Monte Carlo simulations and empirical application confirm that there are some efficiency benefits.

If we view practitioners who use TSLS as implicitly accepting the assumption that we have identified, then such efficient procedures provide a free-lunch—or at least, mostly harmless—improvement for these practitioners. Owing to the overwhelming popularity of TSLS methods in both applied work in causal inference and applied work in structural econometrics, we argue that our analysis is valuable, if at least to clarify the assumptions needed and to suggest "minimally invasive" improvements.

This paper proceeds as follows. Section 2 revisits linear IV specifications and considers several assumptions that relate to TSLS and finds that partial mean-independence is a natural one that satisfies three desiderata. Section 3 then studies efficiency and efficient estimation under the partial mean-independence assumption that we propose. Finally, Section 4 illustrates our theory with a Monte Carlo study and an empirical application.

2. **Linear IV specifications**

Consider some outcome variable Y_i, treatment variable W_i, covariates X_i, and instrument Z_i. We let $\mathbf{D}_i = [1, W_i', X_i']'$ and $\mathbf{Z}_i = [Z_i', X_i']'$ collect the second stage variables and the first stage variables, respectively. To distinguish observed and counterfactual values of Y_i, we let the random variable $Y_i(w)$ denote a potential outcome if individual i is assigned treatment level w, and think of Y_i as the observed outcome $Y_i(W_i)$.[2] Similarly, we let $W_i(z)$ denote a potential treatment level, where $W_i(\cdot)$ describes the *compliance pattern* for unit i.[3] We assume a cross-sectional setting where the structural variables are sampled according to some

[1] An estimator is locally efficient for a given model $\mathcal{P}_0$ at some restriction of the model $\mathcal{P}_1 \subset \mathcal{P}_0$ if it is consistent and asymptotically normal at all $P \in \mathcal{P}_0$ and its asymptotic variance matches the efficiency bound for $\mathcal{P}_0$ at all $P \in \mathcal{P}_1$. See Newey (1990); Graham, de Xavier Pinto and Egel (2012) for related discussions. Note, however, this is distinct from efficiency for $\mathcal{P}_1$, because the efficiency bound for $\mathcal{P}_1$ is in general higher than the efficiency bound for $\mathcal{P}_0$ at a given $P \in \mathcal{P}_1$.

[2] Note that doing so imposes the exclusion restriction that Z has no direct effect on Y.

[3] If Z is binary, then $(W(1), W(0))$ characterizes whether an individual is a complier, always-taker, never-taker, or defier (Angrist and Imbens, 1995).

unknown joint distribution

$$(Y_i(\cdot), W_i(\cdot), X_i, Z_i) \overset{\text{i.i.d.}}{\sim} P^*.$$

We observe $(Y_i, W_i, X_i, Z_i) \overset{\text{i.i.d.}}{\sim} P$ where $W_i = W_i(Z_i)$ and $Y_i = Y_i(W_i)$. Throughout, we assume $(Y_i(w), W_i(z), X_i, Z_i)$ has finite second moments for all w, z.

To estimate causal effects of W on Y, many researchers use the two-stage least-squares (TSLS) specification:[4]

$$Y_i = \alpha + W_i'\beta + X_i'\eta + U_i, \quad \mathbb{E}[U_i \mathbf{Z}_i] = 0. \tag{1}$$

Blandhol *et al.* (2022) construct a sample of journal articles in economics using instrumental variable designs, published in one of the five journals: *American Economic Review*, *Econometrica*, *Quarterly Journal of Economics*, *Journal of Political Economy*, and *Review of Economic Studies*. The vast majority of those (i) estimate (1) with some covariates X_i, (ii) deem having X_i as important for causal identification, and (iii) fail to saturate the covariate term $X_i'\eta$. Blandhol *et al.* (2022) point out that, in such cases, β from (1) typically does not correspond to any reasonable aggregate of conditional local average treatment effects, under the standard nonparametric potential outcomes model and even under additional restrictions.

This paper gives (1)—in settings with (i)–(iii)—some benefit of doubt and asks what structural restrictions rationalize (1) as an estimator of causal or structural parameters. We find that intuitive structural restrictions have certain undesirable properties, and we propose a new restriction that rationalizes TSLS. We then consider efficient estimation of β_0 under this new restriction. This restriction is indeed stronger than those considered by Blandhol *et al.* (2022). However, since TSLS specifications are very popular in practice and often interpreted as estimating causal effects, there is an argument that practitioners are revealed to prefer this assumption and implicitly operate under it.

To have some hope that (1) estimates a treatment effect parameter, it is natural to at least restrict to a constant and linear treatment effects model so that we can define β_0 as the slope of the treatment acting on the potential outcomes. In some cases, the procedures we recommend have LATE interpretations under heterogeneous treatment effects, which we return to in Section 3.3.

Assumption 2.1 (Constant and linear treatment effects)**.** *Fix some baseline level w_0. The mean treatment effect is constant and linear in the treatment level w, regardless of compliance pattern $W_i(\cdot)$ and covariate value X_i:*

$$\mathbb{E}[Y_i(w) - Y_i(w_0) \mid X_i, W_i(\cdot)] = (w - w_0)'\beta_0. \tag{2}$$

[4]Here, we do not think of (1) as a well-specified structural equation. What we mean by (1) is not a substantively restrictive model on P^* or P; rather, it defines an estimand and estimator through the moment condition $\mathbb{E}[(Y_i - \alpha - W_i'\beta - X_i'\eta)\mathbf{Z}_i] = 0$. When $\dim W = \dim Z$, this moment condition is just-identified and a solution always exists subjected to some rank condition. When $\dim Z > \dim W$, TSLS corresponds to some weighted GMM objective.

Moreover, the instrument does not select on heterogeneous treatment effects:

$$\mathbb{E}[Y_i(w) - Y_i(w_0) \mid W_i(\cdot), X_i, Z_i] = (w - w_0)'\beta_0. \tag{3}$$

The restriction (2) is the same as Assumption CLE in Blandhol *et al.* (2022) and imposes that the treatment effects are constant and linear, at least on average, within each covariate-by-compliance-pattern cell. The additional restriction (3) imposes that the instrument similarly does not select on the treatment effect heterogeneity. It is guaranteed by (2) and random assignment ($Z_i \perp\!\!\!\perp (Y_i(\cdot), W_i(\cdot)) \mid X_i$). Here, we impose (3) as a high-level restriction.

Given Assumption 2.1, we consider three reasonable requirements for structural restrictions that rationalize (1):

(a) TSLS (1) should estimate β_0 consistently in settings with features (i)–(iii) above.
(b) Nonlinear functions of X_i should not be valid external instruments.
(c) Nonlinear functions of Z_i can be valid external instruments.

The requirement (a) states that adjusting for X_i linearly in TSLS is sufficient to recover β_0. Imposing (b) is reasonable: Since X_i is typically not saturated, such nonlinear variation is usually available. Practitioners sensibly do not use such variation for identification of (β_0, η_0) and instead opt for an instrument that is exogenous in some sense. Similarly, imposing (c) also conforms with our intuition on sources of causal identification: If researchers believe they have found a good instrument, then they ought to be able to use any transformation of it to identify β_0.

The restriction that satisfies all three requirements is the following. Define the partially linear projection of a random variable Y^* onto $(Z^*; X^*)$ as the minimizer of the squared error prediction loss (assuming a unique minimizer exists) over functions $f_1(Z^*) + \eta' X^*$:

$$\mathbb{PL}[Y^* \mid Z^*; X^*] \equiv \underset{\{f(z,x)=f_1(z)+x'\eta\}}{\arg\min} \mathbb{E}\left[(Y_i^* - f(Z^*; X^*))^2\right].$$

Similarly, define the linear projection $\mathbb{L}[Y^* \mid X^*]$ as the solution to the linear least-squares problem (again, assuming a unique minimizer exists):

$$\mathbb{L}[Y^* \mid X^*] = \underset{\{f(x)=\alpha+x'\eta\}}{\arg\min} \mathbb{E}\left[(Y^* - f(X^*))^2\right]. \tag{4}$$

Definition 2.1. We say that Z^* is *partially mean-independent* of Y^* given X^* if $\mathbb{E}[Y^* \mid Z^*; X^*] = \mathbb{L}[Y^* \mid X^*]$.

We propose the following restriction in terms of partial mean-independence of the instrument Z on the baseline potential outcome $Y(w_0)$ given the covariates X:

Assumption 2.2 (Partial mean independence)**.** *$Y_i(w_0)$ is partially mean independent of Z_i given X_i.*

This assumption means that practitioners choose X_i judiciously so that a partially linear regression of $Y_i(w_0)$ on Z_i and X_i is solely an affine function of X_i, with the nonlinear part of Z_i being a constant function.

This restriction is equivalent to a structural equation model with a partially mean-zero restriction, as the following lemma makes precise.

Lemma 2.2. *Any P^* that satisfies Assumptions 2.1 and 2.2 implies a distribution P for observed variables that satisfies the structural equation*

$$Y_i = \alpha_0 + W_i'\beta_0 + X_i'\eta_0 + U_i \quad \mathbb{PL}[U_i \mid Z_i; X_i] = 0 \tag{5}$$

for some (α_0, η_0). Conversely, any P that satisfies (5) can be rationalized by some P^* that satifies Assumptions 2.1 and 2.2. Moreover, the restriction $\mathbb{PL}[U_i \mid Z_i; X_i] = 0$ is equivalent to the restriction that $\mathbb{E}[U_i X_i] = \mathbb{E}[U_i \mid Z_i] = 0$.

The moment condition (5) immediately shows that Assumption 2.2 satisfies the requirement (a)–(c). For (a), since (5) implies that $\mathbb{E}[U_i \mathbf{Z}_i] = 0$, β_0 from TSLS is consistent subjected to a rank condition (strong first-stage). For (b), since $\mathbb{E}[U_i X_i] = 0$ does not necessarily imply $\mathbb{E}[U_i f(X_i)] = 0$, nonlinear functions of X_i are not valid instruments without further assumption. For (c), since $\mathbb{E}[U_i \mid Z_i] = 0$ implies that $\mathbb{E}[U_i f(Z_i)] = 0$ for any f, nonlinear functions of Z_i do provide identifying variation for β_0.[5]

We can view Assumption 2.2 as a different independence restriction on Z and $Y(w_0)$. Recall that statistical independence restrictions can be recast as prediction problems. A (scalar) random variable Y^* is independent of Z^* conditional on X^* if for all (bounded) functions f, Z^* generates no additional predictive power:[6] $\mathbb{E}[f(Y^*) \mid Z^*, X^*] = \mathbb{E}[f(Y^*) \mid X^*]$. Y^* is mean-independent of Z^* given X^* if the above holds for the identity map $f(t) = t$: $\mathbb{E}[Y^* \mid Z^*, X^*] = \mathbb{E}[Y^* \mid X^*]$. Similarly, Y^* is partially uncorrelated with Z^* given X^* if controlling for X^* in a linear regression removes the coefficient on Z^*: For $\mathbb{L}$ defined as the linear projection operator, $\mathbb{L}[Y^* \mid Z^*, X^*] = \mathbb{L}[Y^* \mid X^*]$. Viewed in this light, Assumption 2.2 is simply replacing these independence restrictions with partial mean-independence.

2.1. **Alternative restrictions and why they do not satisfy (a)–(c).** To further see why Assumption 2.2 is a natural restriction given the requirements (a)–(c), it is helpful to consider several other restrictions that may appear natural.

The simplest restriction such that (1) is consistent for some causal effect is the model where we assume the following outcome model along with Assumption 2.1:

$$\mathbb{E}[Y_i(w_0) \mid X_i, Z_i] = \alpha_0 + \eta_0' X_i \tag{6}$$

[5]To the best of our knowledge, the restriction (5)—and hence Assumption 2.2—is novel, and first proposed by the first draft of this article (https://arxiv.org/abs/2011.06158v1).

[6]To prove this, we can, for instance, let $f(y) = \mathbb{1}(y \leq t)$.

Under random assignment of Z_i, (6) is equivalent to an outcome model where the conditional expectations of the potential outcomes are linear in the covariates (Assumption LIN in Blandhol *et al.* (2022)):

$$\mathbb{E}[Y_i(w_0) \mid X_i] = \alpha_0 + \eta_0' X_i.$$

This restriction is equivalent—in the sense similar to Lemma 2.2—to the familiar linear IV structural equation model, ubiquitous in both classical econometrics and modern structural econometrics[7]

$$Y_i = \alpha_0 + W_i'\beta_0 + X_i'\eta_0 + U_i \quad \mathbb{E}[U_i \mid \mathbf{Z}_i] = 0. \tag{7}$$

As pointed out by, among others, Abadie (2003); Blandhol *et al.* (2022); Angrist and Pischke (2008), this restriction allows for "backdoor identification" through nonlinear functions of X_i, violating requirement (b). Compared to this model, the requirement (5) is a weakening of (7) by removing nonlinear functions of X_i as valid instruments.

The outcome model (6) is too strong to satisfy (b). Weakening the outcome model (6) to allow for nonlinear covariates[8]

$$\mathbb{E}[Y_i(w_0) \mid X_i, Z_i] = f_0(X_i) \tag{8}$$

yields, in structural equation form, *the partially linear IV model* (e.g., (4.5) in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018)):[9]

$$Y_i = W_i'\beta_0 + f_0(X_i) + U_i \quad \mathbb{E}[U_i \mid \mathbf{Z}_i] = 0.$$

Although nonlinear functions of X_i no longer have identifying power for β_0, linear TSLS (1) is in general inconsistent for β_0 in this model; thus, this model would fail the requirement (a).

Further restrictions on the design can restore consistency. If the instrument Z_i is fully randomized (i.e., $Z_i \perp\!\!\!\perp (Y_i(\cdot), W_i(\cdot))$ unconditionally), then TSLS is consistent and can be improved by finding better transformations $g(Z)$ as instruments (Coussens and Spiess, 2021). However, many empirical papers in Blandhol *et al.* (2022)'s sample explicitly or implicitly deem controlling for X_i as important for identification.

To connect our assumption to the partially linear model, we note that under random assignment of Z_i, our Assumption 2.2 is equivalent to an additional restriction on $f_0(\cdot)$:

[7] In Berry, Levinsohn and Pakes (1995), for instance, (7) is present as a model of mean consumer utility Y_i (typically denoted δ), where U_i (typicalled denoted ξ) is a notion of unobserved characteristics of a product (Berry and Haile, 2021).

[8] Under random assignment, (8) does not impose any actual restrictions on the outcomes.

[9] To estimate β_0 in this model, one could work with the double/debiased machine learning moment condition in Chernozhukov *et al.* (2018). One could also let $V_i = f_0(X_i) + U_i$ and note that it is orthogonal to any *recentered* instrument $g(x,z) - \mathbb{E}[g(X,Z) \mid X = x]$ (Borusyak and Hull, 2023):

$$\mathbb{E}[V_i \cdot (g(X_i, Z_i) - \mathbb{E}[g(X_i, Z_i) \mid X_i])] = 0.$$

Proposition 2.3. *Under random assignment* $Z_i \perp\!\!\!\perp Y_i(\cdot) \mid X_i$,[10] *the following are equivalent:*

(1) Assumption 2.2

(2) Z_i *does not predict the nonlinearity of the mean potential outcome given* X_i*: For* $f_0(X_i) \equiv \mathbb{E}[Y_i(w_0) \mid X_i]$,

$$\mathbb{E}[f_0(X_i) - \mathbb{L}[f_0(X_i) \mid X_i] \mid Z_i] = 0.$$

and $\mathbb{PL}[Y_i(w_0) \mid Z_i; X_i] = \mathbb{L}[f_0(X_i) \mid X_i]$.

Proposition 2.3 states that Assumption 2.2 is equivalent to adding the restriction that the nonlinearity of $f_0(X_i)$ is not predicted by Z_i, which is indeed intuitively necessary for (a) to hold.

The specification (8) weakens (6) too much to satisfy (a)–(c). A different way to weaken (6) is by requiring simply that the linear projection of $Y(w_0)$ on Z_i, X_i is $\alpha_0 + \eta_0' X_i$:

$$\mathbb{L}[Y_i(w_0) \mid Z_i, X_i] = \alpha_0 + X_i'\eta_0 \text{ where } \mathbb{L}[Y^* \mid X^*] = \underset{\{f(x)=\alpha+x'\eta\}}{\arg\min} \mathbb{E}\left[(Y^* - f(X^*))^2\right]. \quad (9)$$

This restriction again allows TSLS to estimate β_0.[11] It disallows nonlinear functions of X_i from being used as instruments. However, such a restriction also potentially disallows nonlinear functions of Z_i—which we commonly think of as exogenous—as instruments, thus violating (c). Compared to (4), Assumption 2.2 is exactly equivalent to the version of (9) that holds for all nonlinear functions $g(Z_i)$:

Proposition 2.4. *Assumption 2.2 is equivalent to the following: For any function* $g(\cdot)$,

$$\mathbb{L}[Y_i(w_0) \mid g(Z_i), X_i] = \alpha_0 + X_i'\eta_0.$$

All three alternative models we consider are unsatisfying for TSLS in that they violate one of (a)–(c). In each of the cases, Assumption 2.2 is a natural strengthening or weakening of the model to rescue the failed requirements.

It is a fair critique that Assumption 2.2 can be knife-edge. For instance, exactly what is meant by Assumption 2.2 depends—*necessarily so*—on the vector of X_i that practitioners choose. Then again, we conduct this exercise presuming that TSLS is innocent, and so the fact that Assumption 2.2 is knife-edge reflects difficulties thinking of TSLS as a "model-free" estimator. Nevertheless, we think there is value in studying Assumption 2.2, given that TSLS—and linear IV structural models like (7)—are popular in both causal inference and structural econometrics. Assumption 2.2 clarifies the nature of functional-form assumptions one would have to defend if one uses TSLS, and gives vast amount of existing work interpretation.

[10] The proof only uses mean-independence: $\mathbb{E}[Y_i(\cdot) \mid X_i, Z_i] = \mathbb{E}[Y_i(\cdot) \mid X_i]$.

[11] If Z is binary and randomly assigned, then one way to ensure (9) for all $Y(w_0)$ is if the propensity score is linear: $\mathbb{E}[Z_i \mid X_i] = \mathbb{L}[Z_i \mid X_i]$. This is the "rich covariate" restriction in Blandhol *et al.* (2022).

Accepting Assumption 2.2, our next section on efficient estimation yields free-lunch improvements over TSLS—in the sense that under reasonable settings where TSLS estimates a causal efect, the estimators we propose improve on TSLS.

3. **Efficiency, estimation, and inference**

The rest of the paper is concerned with efficient estimation of β_0 given Assumptions 2.1 and 2.2. We first discuss efficient estimation in population in Section 3.1. Like conventional instrumental variable settings, efficient estimation of β_0 can be viewed as deriving an *optimal instrument*—that is, a choice of $\Upsilon(\mathbf{Z}_i)$ that delivers the smallest possible asymptotic variance—and running TSLS with this optimal instrument. In general, the optimal instrument depends on unknown population quantities that need to be estimated—Section 3.2 then discusses estimating this optimal instrument. Like conventional settings, fully efficient estimation requires weighting by inverse of the conditional heteroskedasticity. However, the form of the optimal instrument under homoskedasticity is considerably simpler and have some additional appealing properties, which we discuss in Section 3.3. Finally, Section 3.4 discusses an implementation of the estimation procedure in Section 3.2.

3.1. **Efficiency.**

Define $\sigma^2(Z_i) = \mathbb{E}[U_i^2 \mid Z_i]$ as the conditional variance of the moment condition given Z_i and

$$\tilde{X}_i = X_i - \frac{\mathbb{E}[X_i U_i^2 \mid Z_i]}{\sigma^2(Z_i)}$$

as the residual of X_i from a weighted projection onto Z_i. Following Chamberlain (1992), the following $\dim(\mathbf{D}_i)$-vector is the optimal instrument for (5):

$$\Upsilon^\star(\mathbf{Z}_i) = \frac{\mathbb{E}[\mathbf{D}_i \mid Z_i]}{\sigma^2(Z_i)} + \mathbb{E}[\mathbf{D}_i \tilde{X}_i'] \mathbb{E}[U_i^2 \tilde{X}_i \tilde{X}_i']^{-1} \tilde{X}_i. \tag{10}$$

Given a sample $(Y_i, X_i, Z_i, W_i)_{i=1}^N$, an (infeasible) efficient estimator of $\theta_0 = (\alpha_0, \beta_0', \eta_0')'$ for (5) is obtained by

$$\hat{\theta}_N^\star = \left(\frac{1}{n} \sum_{i=1}^N \Upsilon^\star(\mathbf{Z}_i) \mathbf{D}_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^N \Upsilon^\star(\mathbf{Z}_i) Y_i. \tag{11}$$

Feasible estimators can be constructed by estimating $\Upsilon^\star(\cdot)$ in a first step. We collect these efficiency results in the following theorem.

Theorem 3.1. *The semiparametric efficiency bound for the* sequential moment restrictions (5) *is*

$$V^\star = \left(\mathbb{E}\left[\frac{1}{\sigma^2(Z_i)} \mathbb{E}[\mathbf{D}_i \mid Z_i] \mathbb{E}[\mathbf{D}_i \mid Z_i]' \right] + \mathbb{E}[\mathbf{D}_i \tilde{X}_i'] \mathbb{E}[U_i \tilde{X}_i \tilde{X}_i']^{-1} \mathbb{E}[\tilde{X}_i \mathbf{D}_i'] \right)^{-1}.$$

Under the assumptions that (i) $\mathrm{Var}(\Upsilon^\star(\mathbf{Z}_i) U_i) < \infty$ *and (ii)* $\mathbb{E}[\Upsilon^\star(\mathbf{Z}_i)\mathbf{D}_i'] < \infty$ *and is full rank,* $\sqrt{n}(\hat{\theta}_N^\star - \theta_0) \xrightarrow{d} \mathcal{N}(0, V^\star)$ *as* $N \to \infty$.

The intuition for the optimal instrument (10) is the following. The two *sequential* moment conditions

$$\mathbb{E}[U_i \mid Z_i] = 0 \quad \mathbb{E}[U_i X_i] = 0$$

are correlated, in the sense that $\operatorname{Cov}(U_i, U_i X_i \mid Z_i) \neq 0$. We note that if we project the second moment onto the first moment, we obtain

$$U_i X_i - \frac{\operatorname{Cov}(U_i, U_i X_i \mid Z_i)}{\operatorname{Var}(U_i \mid Z_i)} U_i X_i = U_i \tilde{X}_i.$$

This transformation decorrelates the moments: $\mathbb{E}[U_i \mid Z_i] = 0$ and $\mathbb{E}[\tilde{X}_i U_i] = 0$ are uncorrelated since $\operatorname{Cov}(U_i, \tilde{X}_i U_i \mid Z_i) = 0$ by construction. The insight of Chamberlain (1992) is that when sequential moment restrictions are uncorrelated in this sense, their information bounds (that is, the inverse of the efficiency bound) can be directly summed. Efficient estimation then amounts to adding up the individual moment conditions, where each individual moment condition is transformed to an unconditional moment condition by plugging in its optimal instrument.

The optimal instrument for $\mathbb{E}[U_i \mid Z_i] = 0$ is $\Upsilon_1^\star(Z_i) = \frac{\mathbb{E}[\mathbf{D}_i \mid Z_i]}{\operatorname{Var}(U_i \mid Z_i)}$ and the optimal instrument for $\mathbb{E}[U_i \tilde{X}_i]$ is $\Upsilon_2^\star = \mathbb{E}[\mathbf{D}_i \tilde{X}_i'] \mathbb{E}[U_i^2 \tilde{X}_i \tilde{X}_i']^{-1}$.[12] Thus the sum of the transformed moments is indeed the moment condition obtained by using $\Upsilon^\star$:

$$\mathbb{E}[\Upsilon_1^\star(Z_i) U_i + \Upsilon_2^\star \tilde{X}_i U_i] = \mathbb{E}[\Upsilon^\star(\mathbf{Z}_i) U_i] = 0.$$

3.2. **Estimation and inference.** We now consider feasible estimators in the form of (11), where we replace the unknown instrument $\Upsilon^\star(\cdot)$ with some feasible estimate $\hat{\Upsilon}_N(\cdot)$. To allow for a wide array of estimators and to avoid empirical process arguments, we consider cross-fitting (Chernozhukov *et al.*, 2018).[13] A cross-fitted estimation procedure splits the sample into K folds, such that, for i in fold k and fixing $\mathbf{Z}_i$, $\hat{\Upsilon}_n(\mathbf{Z}_i)$ depends solely on observations from folds other than fold k. This procedure does introduce additional variability into the final estimates, but this variability can be mitigated by the procedure proposed by Ritzwoller and Romano (2023).

Since our theoretical results generalize immediately to any constant K, we let $K = 2$ and consider a partition $I_1 \cup I_2 = \{1, \ldots, N\}$. For convenience, assume N is even.

[12] Recall that for a conditional moment restriction

$$\mathbb{E}[m(A_i, \vartheta_0) \mid B_i] = 0,$$

Chamberlain (1987) shows that the optimal instrument is

$$-\mathbb{E}\left[\left.\frac{\partial m}{\partial \vartheta}\right|_{\vartheta=\vartheta_0} \mid B_i\right]' \operatorname{Var}(m(A_i, \vartheta_0) \mid B_i)^{-1}.$$

[13] If we restrict estimation to a sufficiently restrictive class of functions (e.g. a class with well-behaved metric entropy), empirical process arguments would allow for similar results without cross-fitting (Vaart and Wellner, 2023).

Theorem 3.2. *For a cross-fitted and partially linear estimate $\hat{\Upsilon}_N(\mathbf{Z}_i)$, assume that*

(1) (Convergence in $L^2(\mathbf{Z})$ to a limiting instrument) $\hat{\Upsilon}_N$ converges to some partially linear $\Upsilon_0(\mathbf{Z}_i) = f_1(Z_i) + X_i'\gamma_0$ in the following out-of-sample sense

$$\lim_{N\to\infty} \mathbb{E}\left[\|\hat{\Upsilon}_N(\mathbf{Z}_i^*) - \Upsilon_0(\mathbf{Z}_i^*)\|^2\right] = \lim_{N\to\infty} \mathbb{E}\left[\|\hat{\Upsilon}_N - \Upsilon_0\|^2_{L^2(\mathbf{Z})}\right] = 0.$$

Here, the expectation integrates over a distribution where $\hat{\Upsilon}_N$ depends on $N/2$ iid copies of (Y_i, X_i, W_i, Z_i) and $\mathbf{Z}_i^*$ is an additional iid copy of $\mathbf{Z}_i$.

(2) (Finite moments) The following moments are finite: $\mathbb{E}[\mathbf{D}_i\mathbf{D}_i'] \leq \infty$, $\mathbb{E}[|U_i|^3\|\Upsilon_0(\mathbf{Z}_i)\|^3] < \infty$, $\mathrm{Var}(U_i \mid \mathbf{Z}_i) \leq M$ uniformly over $\mathbf{Z}_i$.

(3) (Limiting instrument is strong) $\mathbb{E}[\Upsilon_0(\mathbf{Z}_i)\mathbf{D}_i']$ is full rank.

Let

$$\hat{\theta}_N = \left(\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)\mathbf{D}_i'\right)^{-1} \frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)Y_i \tag{12}$$

and let $\hat{\theta}_{0,N} = \left(\frac{1}{N}\sum_i \Upsilon_0(\mathbf{Z}_i)\mathbf{D}_i'\right)^{-1} \frac{1}{N}\sum_i \Upsilon_0(\mathbf{Z}_i)Y_i$.[14] *Then:*

(1) $\hat{\theta}_N$ and $\hat{\theta}_{0,N}$ are asymptotically equivalent

$$\sqrt{N}(\hat{\theta}_N - \hat{\theta}_{0,N}) \xrightarrow{p} 0.$$

and both are asymptotically normal.

(2) The asymptotic variance of $\hat{\theta}_N$ is equal to the efficiency bound in Theorem 3.1 if $\hat{\Upsilon}_N$ is consistent for the optimal instrument: i.e., $\Upsilon_0 = \Upsilon^\star$.

(3) Assuming additionally that $\mathbf{D}_i$ and $\hat{\Upsilon}_N(\mathbf{Z}_i)$ has a uniformly bounded fourth moment conditional on $\hat{\Upsilon}_N(\cdot)$ almost surely, the asymptotic variance is consistently estimated by its empirical counterpart

$$\hat{V}_N = \left(\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)\mathbf{D}_i'\right)^{-1} \left(\frac{1}{N}\sum_i (Y_i - \hat{\theta}_N'\mathbf{D}_i)^2 \hat{\Upsilon}_N(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)'\right) \left(\frac{1}{N}\sum_i \mathbf{D}_i\hat{\Upsilon}_N(\mathbf{Z}_i)'\right)^{-1}.$$

Theorem 3.2 verifies that a plug-in estimator, using an estimated instrument, is asymptotically equivalent to an estimator that uses the limit of the estimated instrument. When that limit equals the optimal instrument, the resulting estimator for θ_0 is also efficient. Theorem 3.2 shows additional robustness in the sense that even if $\hat{\Upsilon}_N$ is inconsistent for $\Upsilon^\star$—but so long as it converges to some strong instrument Υ_0—we still obtain a consistent and asymptotically normal estimator for θ_0.

[14] Since we assume $\mathbb{E}[\Upsilon_0\mathbf{D}_i']$ is full rank, with probability tending to 1 both $\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)\mathbf{D}_i'$ and $\frac{1}{N}\sum_i \hat{\Upsilon}_0(\mathbf{Z}_i)\mathbf{D}_i'$ are full rank. In the event that they are not, we may replace the inverse operation with Moore–Penrose pseudoinverse.

The key assumption for Theorem 3.2 is a mean-square *stability* requirement for $\hat{\Upsilon}_N$.[15] This is weaker than many assumptions in the literature for semiparametric models (Chen, 2007; Chernozhukov *et al.*, 2018), which typically require consistency of at least $o(N^{-1/4})$ in $\|\cdot\|_2$. In contrast, (1) in Theorem 3.2 does not require any rate of convergence. The reason here is that the moment restriction (5) possesses the robustness property that any partially linear function $\Upsilon_0(\cdot)$ are valid instruments, and so bad estimates of $\Upsilon^\star$ do not necessarily cause bias in θ_0. This robustness property is special to *parametric* conditional moment restrictions—the partially linear IV model, for instance, does not possess this robustness and requires $o(n^{-1/4})$-rate conditions (Chernozhukov *et al.*, 2018).[16]

Remark 3.3 ("Mostly harmless" improvement over TSLS)**.** We conclude this section with a discussion that highlights the "mostly harmless" nature of (an improvement of) this estimator. It is sometimes the case that the limiting instrument $\Upsilon_0(\cdot)$ is in fact a poor one. In such cases, using this estimator performs worse than simply using TSLS with Z_i included linearly. We could correct this issue by treating $\tilde{\mathbf{Z}}_i = [1, \hat{\Upsilon}_N(\mathbf{Z}_i)', \mathbf{Z}_i']$ as a vector of (overidentified) instruments, and estimate optimally weighted GMM with respect to the moment condition $\mathbb{E}[\tilde{\mathbf{Z}}_i(Y_i - \theta_0'\mathbf{D}_i)] = 0$.[17] When $\hat{\Upsilon}_N(\mathbf{Z}_i) \approx \Upsilon^\star(\mathbf{Z}_i)$, doing so recovers efficiency, but when $\hat{\Upsilon}_N(\mathbf{Z}_i)$ is poor, doing so at least recovers the optimally weighted GMM estimator using $\mathbf{Z}_i$. Such an approach is asymptotically at least as efficient as the TSLS estimator using $\mathbf{Z}_i$—asymptotically, at least, it is harmless.

Of course, one might worry that adding $\hat{\Upsilon}_N(\cdot)$ might create a finite-sample "many instrument" bias, essentially due to overfitting when using $\tilde{\mathbf{Z}}_i$ to predict $\mathbf{D}_i$,[18] which would make this approach mostly harmful instead. This issue is mitigated—though not eliminated—by sample-splitting and using the predicted instrument as an instrument in a TSLS specification. Under our setup, this approach at most adds dimension $\dim(\mathbf{D}_i) = \dim(\hat{\Upsilon}_N)$ to the vector of instruments. In the locally efficient (homoskedastic) case that we discuss next, this approach only adds $\dim W_i$ to the vector of instruments, where $\dim(W_i)$ is often 1. Thus, applying sample-splitting makes this approach mostly harmless in practice as well. ■

[15] Conditions like (1) are common in machine learning (Bayle, Bayle, Janson and Mackey, 2020; Austern and Syrgkanis, 2021). For instance, assumption A.2. in Lei, G'Sell, Rinaldo, Tibshirani and Wasserman (2018) is a similar restriction (they consider high-probability bounds in $\|\cdot\|_\infty$ as opposed to bounds on the expected value in $\|\cdot\|_2$).

[16] It is also analogous to the fact that optimally weighted parametric GMM simply requires consistent estimates of the optimal weighting matrix, without further rate requirements.

[17] Echoing the projection argument in Hausman (1978), we can view the resulting optimally weighted GMM estimator as the variance-minimizing weighted average of $\hat{\theta}_{0,N}$ in Theorem 3.2 and the optimally weighted GMM estimator for $\mathbb{E}[\mathbf{Z}_i U_i] = 0$, $\hat{\theta}_{\text{OW-GMM}}$. When $\hat{\theta}_{0,N}$ is efficient, the resulting weighted average puts weight solely on $\hat{\theta}_{0,N}$.

[18] For one, estimated standard errors mechanically decrease as more instruments are added to $\tilde{\mathbf{Z}}_i$, but finite-sample performance is typically poorer.

3.3. **Local efficiency.** The expression (10) simplifies when $\mathbb{E}[U_i^2 \mid \mathbf{Z}_i] = \sigma^2(Z_i) = 1$ is a constant. This leads to a locally efficient instrument that has certain additional appealing properties that are worth discussing.

Lemma 3.4. *When $\mathbb{E}[U_i^2 \mid \mathbf{Z}_i] = 1$ is a constant normalized to 1, the optimal instrument for (5) is the partially linear regression of $\mathbf{D}_i$ on Z_i, X_i:*

$$\Upsilon^\star(\mathbf{Z}_i) = \mathbb{PL}[\mathbf{D}_i \mid Z_i; X_i] = [1, \mathbb{PL}[W_i \mid Z_i; X_i]', X_i']'. \tag{13}$$

The estimator that depends on (13) is consistent for (5) and *locally efficient* at (5) when $\mathbb{E}[U_i^2 \mid \mathbf{Z}_i] = \sigma^2(Z_i)$ is constant. Local efficiency means that such an estimator achieves the efficiency bound that solely imposes (5) (see, e.g., Newey, 1990; Graham *et al.*, 2012). However, this efficiency bound does not presume knowledge of $\mathbb{E}[U_i^2 \mid \mathbf{Z}_i] = 1$, and exploiting this restriction would lead to estimators with lower asymptotic variance.

The instrument (13) is easy to implement, as it does not involve estimating and weighting by $\mathrm{Var}(U_i \mid Z_i)$. This leads the instrument to have additional appealing properties. First, since the instrument includes the covariates $[1, X_i']'$, the corresponding TSLS estimator for the subvector β_0 is numerically equivalent to a TSLS estimator of the following (infeasible) specification:

$$\begin{aligned} Y_i &= \alpha_0 + W_i'\beta_0 + X_i'\eta_0 + U_i \\ W_i &= \gamma_0 + \Upsilon(Z_i)'\delta_0 + X_i'\pi_0 + V_i, \end{aligned} \tag{14}$$

where $\Upsilon(Z_i)$ is the nonlinear part of $\mathbb{PL}[W_i \mid Z_i; X_i] = \Upsilon(Z_i) + \kappa_0 X_i$. Feasible implementations would replace $\Upsilon(Z_i)$ with an estimate, possibly coming from a separate sample of the data. This means that, at least if we sample-split, on the held-out set of the data (conditional on the estimate $\hat{\Upsilon}(Z_i)$), we have a conventional, just-identified linear IV problem, for which a set of additional theoretical results (say on weak instruments) apply (Andrews, Stock and Sun, 2019).

Second, the form (14) also clarifies how $\Upsilon^\star(\cdot)$ improves efficiency—it does so by finding a good prediction of W_i. (14) implies that the asymptotic variance of the estimator that uses $\Upsilon(Z_i)$ for β_0 is of the form

$$V_{0,\beta} = \left(\mathbb{E}[\tilde{\Upsilon}(Z_i)\tilde{W}_i']\mathbb{E}[U_i^2\tilde{\Upsilon}(Z_i)\tilde{\Upsilon}(Z_i)']^{-1}\mathbb{E}[\tilde{W}_i\tilde{\Upsilon}(Z_i)']\right)^{-1}$$

where $\tilde{W}_i = W_i - \mathbb{L}[W_i \mid X_i]$ and $\tilde{\Upsilon}(Z_i) = \Upsilon(Z_i) - \mathbb{L}[Z_i \mid X_i]$. If we assume homoskedasticity $\sigma^2(Z_i) = \mathrm{Var}(U_i \mid Z_i) = \sigma^2$ and that the treatment is a scalar $\dim(W) = 1$, then

$$V_{0,\beta} = \sigma^2 \left(\frac{\mathrm{Cov}(\tilde{W}, \tilde{\Upsilon})^2}{\mathrm{Var}(\tilde{W})\,\mathrm{Var}(\tilde{\Upsilon})}\,\mathrm{Var}(\tilde{W})\right)^{-1} = \frac{1}{R^2_{W\sim\Upsilon|X}}\frac{\sigma^2}{\mathrm{Var}(\tilde{W})}, \tag{15}$$

where $R^2_{W\sim\Upsilon|X}$ is the partial R^2 of a regression of W on the instrument Υ, controlling for X. This echoes a familiar intuition about TSLS, where the second-stage regresses on the first-stage predictions of the endogenous variables, and one might expect that—indeed as (15) demonstrates—better predictions deliver better second-stage estimates.[19] TSLS substitutes in a best *linear* prediction. Our assumption (2.2) expands the class of predictions to ones that are *partially linear* in X_i, allowing for nonlinear functions of Z_i and efficient prediction via machine learning methods. The optimal instrument $\Upsilon(Z_i)$ exactly maximizes this partial R^2 since it is the nonlinear part of a partially linear regression of W_i on Z_i, X_i.

Third, since $\Upsilon^\star(\cdot)$ in this case already includes X_i, the mostly harmless approach discussed in Remark 3.3 would simply use $\tilde{\mathbf{Z}}_i = [1, \Upsilon(Z_i)', Z_i', X_i']'$ as the vector of instruments. Under homoskedasticity, it would also use TSLS rather than optimally weighted GMM to estimate θ_0. In this case, the approach in Remark 3.3 simply boils down to estimating a candidate optimal instrument $\Upsilon(Z_i)$ and including it in the TSLS specification. This approach adds at most $\dim(W_i)$ instruments to the specification, and at least in the leading case where W_i is a scalar, it should not contribute too much to many-instrument bias.

Lastly, this additional structure yields some additional nonparametric interpretation of the estimand, formalized in the following proposition.[20] In short, when W_i is binary, even in the absence of Assumptions 2.1 and 2.2, the estimand of (14) has a causal interpretation if (i) Z_i randomly assigned and independent of X_i and (ii) the *strong monotonicity* condition holds: for all values of X_i, $W_i(z)$ has the same ordering over z.

Proposition 3.5. *Suppose W_i is binary and suppose that*

$$W_i(z) = \mathbb{1}(u(z) > V_i(X_i)) \qquad \text{(Strong monotonicity)}$$

for some random $V_i(\cdot)$ with the marginal distribution of $V_i = V_i(X_i)$ normalized to $\text{Unif}[0,1]$. *Suppose $\Upsilon(Z_i)$ is mean zero such that*

$$\mathbb{PL}[W_i \mid Z_i; X_i] = \Upsilon(Z_i) + [1, X_i']\kappa_0.$$

Define the marginal treatment effect at v to be

$$\text{MTE}(v) = \mathbb{E}[Y_i(1) - Y_i(0) \mid V_i = v].$$

[19] We warn that regressing on $\Upsilon(Z_i)$ would *not* deliver consistent estimates of β_0, as confirmed empirically by Lennon, Rubin and Waddell (2022)—this is what Angrist and Pischke (2008) refer to as forbidden regression. Instead, one needs to use $\Upsilon(Z_i)$ as instruments in a just-identified TSLS regression (14).

[20] This proposition is related to Corollary 3.4 in Słoczyński (2022) and section 4 in Heckman and Vytlacil (2005). Relative to Słoczyński (2022), we do not restrict to binary instruments but do restrict to fully randomized instruments. This proposition can be viewed as a corollary of section 4.3 of Heckman and Vytlacil (2005) exploiting the additional observation that $\mathbb{E}[W_i \mid Z_i] = \Upsilon(Z_i) + \mathbb{E}[W_i]$.

Then, suppose Z_i is fully randomly assigned $Z_i \perp\!\!\!\perp (X_i, Y_i(\cdot), W_i(\cdot))$. Then the estimand for β in (14) is a convex *average of marginal treatment effects:*

$$\beta = \int_0^1 \frac{\mathbb{E}[\Upsilon(Z_i)\mathbb{1}(u(Z_i) > v)]}{\mathbb{E}[\Upsilon(Z_i)u(Z_i)]}\mathrm{MTE}(v)\,dv.$$

3.4. **Implementation.** We describe a concrete implementation of $\hat{\Upsilon}_N$. We start from estimating locally efficient instrument (13). Fix a sample splitting fold k and let j index all N_{-k}-observations *not* in fold k. By a transformation of Robinson (1988), note that

$$\mathbb{PL}[W \mid Z; X] = \mathbb{E}[W \mid Z] + \mathbb{L}[(W - \mathbb{E}[W \mid Z]) \mid (X - \mathbb{E}[X \mid Z])].$$

Thus, by the plug-in principle, we may estimate the partially linear regression via the following steps on the data in folds other than k:

(1) Let $\hat{f}_{1N,k}(Z)$ be an estimate of $\mathbb{E}[W \mid Z]$, obtained, for instance, via minimizing squared error with machine learning methods.
(2) Similarly, let $\hat{f}_{2N,k}(Z)$ be an estimate of $\mathbb{E}[X \mid Z]$.
(3) Let $\hat{\beta}_{3N,k}$ be the $\dim(W) \times \dim(X)$ matrix of linear regression coefficients of $W - \hat{f}_{1N,k}(Z)$ on $X - \hat{f}_{2N,k}(Z)$.
(4) Return $\hat{\Upsilon}_{N,k}(Z) = \hat{f}_{1N,k}(Z) - \hat{\beta}_{3N,k}\hat{f}_{2N,k}(Z)$.[21]

For observations i in fold k, we can then define

$$\hat{\Upsilon}_N(\mathbf{Z}_i) = [1, \hat{\Upsilon}'_{N,k}(Z_i), X_i']'.$$

and plug into (12) to obtain an estimate $\hat{\theta}_N$.

The fully optimal instrument (10) requires estimating more nuisance parameters, but it is conceptually similar via the plug-in principle. On folds other than k:

(1) Let $\hat{f}_{N,k}(Z) = [1, \hat{f}_{1N,k}(Z)', \hat{f}_{2N,k}(Z)']'$ be an estimate of $\mathbb{E}[\mathbf{D}_j \mid Z_j] = [1, \mathbb{E}[W_j \mid Z_j]', \mathbb{E}[X_j \mid Z_j]']$.
(2) Obtain a first step consistent estimate $\tilde{\theta}_N$ of θ_0, via, say, TSLS.
(3) Let $\hat{U}_j = Y_j - \mathbf{D}_j'\tilde{\theta}_N$.
(4) Obtain an estimate $\hat{\sigma}^2_{kN}(Z)$ of $\mathbb{E}[\hat{U}_j^2 \mid Z_j]$ and $\hat{f}_{3N,k}(Z)$ of $\mathbb{E}[\hat{U}_j^2 X_j \mid Z_j]$.
(5) Let $\check{X}_i = X_j - \hat{f}_{3N,k}(Z_j)/\hat{\sigma}^2_{kN}(Z_j)$ be an analogue of $\tilde{X}_j$ and compute

$$\hat{\beta}_{4N,k} = \left(\frac{1}{N_{-k}}\sum_j \mathbf{D}_j\check{X}_j'\right)\left(\frac{1}{N_{-k}}\sum_j \hat{U}_j^2\check{X}_j\check{X}_j'\right)^{-1}.$$

(6) Return $\hat{\Upsilon}_{N,k}(\mathbf{Z}) = \frac{\hat{f}_{N,k}(Z)}{\hat{\sigma}^2_{N,k}(Z)} + \hat{\beta}_{4N,k}(X - \hat{f}_{3N,k}(Z)/\hat{\sigma}^2_{kN}(Z))$.

For observations i in fold k, we can then define

$$\hat{\Upsilon}_N(\mathbf{Z}_i) = \hat{\Upsilon}_{N,k}(\mathbf{Z}_i).$$

[21] We are omitting a $\hat{\beta}'X$ term since X already enters the vector of instruments.

4. **Empirical applications**

4.1. **Monte Carlo.** To illustrate our theoretical results on efficiency, we consider a Monte Carlo setting where the benefits of machine learning methods are salient. We consider the following data-generating process, where $\mathbb{E}[W \mid Z]$ is a complicated function of Z, so that machine learning methods are likely more effective than traditional methods in learning this function:

$$\begin{aligned}
&\epsilon_1 \sim \mathcal{N}(0,1) \quad \epsilon_2 \sim \mathcal{N}(0,1) \quad U_1 = \Phi\left(\frac{\epsilon_1+\epsilon_2}{\sqrt{2}}\right) \sim \text{Unif}[0,1] \\
&Z = \mathcal{N}(0, I_3) \quad \overline{Z} = \frac{1}{\sqrt{3}}\sum_j Z_j \sim \mathcal{N}(0,1) \\
&X = m(\overline{Z}) + s(\overline{Z})\epsilon_1 \text{ where } m(t) = \frac{1}{1+e^{-t}} - 0.5 \text{ and } s(t) = \sqrt{1-m(t)^2} \\
&W = \mathbb{1}(U_1 < \pi(Z)) \quad \pi(z_1,z_2,z_3) = \frac{1}{1+e^{-3q(z_1,z_2)}}\sin^2(2z_3) \\
&q(z_1,z_2) = \begin{cases} (z_1^2+z_2^2) & z_1z_2 > 0, z_1^2+z_2^2 < 1 \\ -(z_1^2+z_2^2) & z_1z_2 < 0, z_1^2+z_2^2 < 1 \\ 0.1 & z_1^2+z_2^2 > 1. \end{cases} \\
&Y = W + 0.5(X^2 - X) + (0.5\overline{Z})\epsilon_2. \qquad (16)
\end{aligned}$$

Since X is symmetric about zero, $\mathbb{L}[X^2 \mid X] = \mathbb{E}[X^2]$. Note too that $\mathbb{E}[X^2 \mid Z] = \mathbb{E}[X^2] = \mathbb{L}[X^2 \mid X]$ is constant. Thus, this data-generating process satisfies Assumption 2.2.

The performances of various methods are recorded in Table 1. The first two rows of Table 1 illustrates that the optimal instrument (10) does deliver more efficient estimates than a locally efficient instrument, as the true data-generating process is heteroskedastic. The next two rows illustrate the performance of feasible methods. Somewhat disappointingly, it appears that both feasible methods fail to fully accurately estimate the optimal instruments. As a result, they deliver estimates that are less precise. Despite this, they still deliver efficiency improvements over TSLS and have undistorted inferences, which demonstrates the robustness in Theorem 3.2. Finally, due to the complex first-stage relationship in this case, traditional TSLS methods deliver estimates that are orders of magnitudes noisier than these feasible machine learning methods.

4.2. **Empirical applications.** Our empirical application is to Dustmann, Fasani and Speciale (2017). They study the effect of legal immigration status (W_i) on immigrants' consumption behavior (Y_i, measured in log expenditure) in Italy. One of their empirical strategies instrument for legal status with rainfall shocks in the immigrants' home country (Z_i) at the time of migration. If rainfall affects an immigrant's home-country income, it plausibly

TABLE 1. Relative performance of various methods in Monte Carlo design

	Relative MSE (MSE / Oracle MSE)	Coverage
Oracle	1.00	
Oracle locally efficient	1.10	
Estimated efficient	1.54	0.94
Estimated locally efficient	1.56	0.94
TSLS	64.32	0.96
TSLS with polynomials	55.98	0.96

Notes. We estimate several methods on the Monte Carlo design (16) and average over 1000 draws of the data-generating process, with $N = 10,000$. The first column shows the relative MSE of estimating the coefficient on W_i (which is equal to 1 under (16)). Relative MSE is the ratio of MSE against the MSE of the oracle efficient instrument (which is 0.00136). The second coverage shows the empirical coverage of Wald confidence intervals using estimated standard errors. Here, Oracle is the method that uses $\Upsilon^\star$ (10) without having to estimate it (we use $0.75 + 0.25\overline{Z}^2$ as $\mathbb{E}[U^2 \mid Z]$, which is very close to the true function); Oracle locally efficient is the method that uses $\mathbb{PL}[W \mid Z; X]$ without having to estimate it. Estimated efficient and Estimated locally efficient are feasible counterparts to Oracle and Oracle locally efficient, respectively, where we use a 3-fold sample split procedure. Nonparametric estimation of $\mathbb{E}[W \mid Z], \mathbb{E}[X \mid Z]$ is through `lightGBM`, and nonparametric estimation of $\mathbb{E}[U^2 \mid Z]$ is through 50-nearest neighbors. Finally, TSLS is two-stage least-squares with Z, and TSLS with polynomials transforms Z into the basis functions $Z, Z_1^2, Z_2^2, Z_3^2, Z_1Z_2, Z_2Z_3, Z_1Z_3, Z_1Z_2Z_3$. □

affects their migration decisions. Table 5, column 3 in their paper implements one such IV specification, which includes some baseline covariates X_i. This specification finds that illegal status reduces consumption by 0.58 log points (SE 0.2). This specification is replicated in the last row of Table 2. As an illustration, we consider whether specifications nonlinear in Z_i have appreciable effect on the precision of the estimates.

We implement our procedure described in Section 3.3. We choose a simple estimator for the partially linear regression of W_i on Z_i, X_i. Namely, we discretize Z_i into k equipercentile bins, and view the ensuing binscatter as an estimator for partially linear regression (Cattaneo, Crump, Farrell and Feng, 2019). We estimate TSLS with the estimated instrument $\hat{\Upsilon}$ (as in Section 3.3) and the original instrument Z_i. We implement a three-fold cross-fitting procedure. However, it turns out that if we split on clusters (the migrants' origin country in this case), the partially linear regressions generalize poorly onto the hold-out set. As an illustration here, we will ignore the fact that the instrument is assigned clusterwise and treat this setting as a cross-sectional setting, and implement the sample-splits without taking into account clusters.[22]

[22]We still compute clustered SE along with heteroskedasticity-robust SE in the end. We conjecture that, given the binned regression estimator is quite simple here, our results should extend to this case—at least for

TABLE 2. Empirical exercise with Dustmann *et al.* (2017)

Number of bins	Clustered SE (ratio)	EHW SE (ratio)	Estimate	Partial R-sq
5	0.992	0.991	-0.571	0.013
11	0.738	0.832	-0.486	0.018
15	0.882	0.891	-0.557	0.017
20	0.763	0.860	-0.603	0.018
50	0.497	0.583	-0.283	0.042
80	0.468	0.547	-0.328	0.049
Original	0.202	0.129	-0.578	0.012

Notes. For each k, we implement a 3-fold cross-fitting procedure, where the sample is split across units (and not clusters), and output the median across 500 draws of the sample-splits for each k. The estimated IMSE optimal number of bins through Cattaneo *et al.* (2019) is 11. The first two columns show the ratio between the estimated standard errors and their counterparts in the original specification (Table 5, column 3 in Dustmann *et al.* (2017)). The first column is the cluster-robust SEs, and the second column is the heteroskedasticity-robust SEs. The third column shows the point estimate of TSLS. The last column shows the partial cross-validated R^2, computed by taking the average of the out-of-sample partial $R^2_{W\sim\hat{\Upsilon}|X}$ across the three folds. As a note on the variability of these estimates across *sample splits* on the same data: The standard deviation of the estimates across sample splits is about 0.03, which is small compared to the standard errors (on the order of 0.1) and the standard deviation of the standard errors across splits is about 0.01. □

Table 2 shows the median results across 500 sample splits of this exercise. We normalize the standard errors of our procedure by the estimated standard errors in the orginal design. Across different values of k, all values of this ratio are less than 1, meaning that the standard errors of the proposed procedure are smaller than the standard errors in the original design. For the bin size chosen by the procedure of Cattaneo *et al.* (2019) ($k = 11$), our standard errors are only 70%-80% of the original standard errors. Since one needs $(1/0.75)^2 = 1.8$ times larger sample sizes to achieve a 25% reduction in standard errors, this is potentially substantively significant.

Confirming our intuition in Section 3.3, this gain is coming from a higher partial R^2 of predicting W_i from the instruments Z_i, given the covariates X_i, shown in the last column of Table 2. Interestingly, the point estimates remain stable for small values of k, but becomes closer to zero for large values of k. This potentially suggests treatment effect heterogeneity and the violation of Assumption 2.1 in this empirical setting.

small k—without needing to split across clusters; after all, a procedure that just runs TSLS with additional k indicator variables for the quantile bins of Z_i would be valid so long as k is small.

5. Conclusion

Two-stage least-squares with linear covariates is a workhorse specification in empirical economics. This paper considers several restrictions on the potential outcomes model that purport to justify TSLS, and finds that partial mean-independence is a restriction that satisfies three natural requirements. This is a positive result that complements Blandhol *et al.* (2022)'s negative results. Taken as given that practitioners implicitly impose partial mean-independence when they estimate TSLS specifications, we consider efficient estimation of the target slope parameter via computing optimal instruments in the sense of Chamberlain (1987, 1992). We find that estimators that plug in estimated optimal instruments are consistent and asymptotically normal under stability conditions that are notably weaker than typical rate conditions. These estimators, since they operate under the same set of assumptions that justify TSLS, provide a mostly harmless free lunch to practitioners.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, **113** (2), 231–263. 7

Andrews, I., Stock, J. H. and Sun, L. (2019). Weak instruments in iv regression: Theory and practice. In *Annual Review of Economics*. 13

Angrist, J. and Imbens, G. (1995). Identification and estimation of local average treatment effects. 2, 3

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press. 7, 14

Austern, M. and Syrgkanis, V. (2021). Asymptotics of the bootstrap via stability with applications to inference with model selection. *Advances in Neural Information Processing Systems*, **34**, 10705–10717. 12

Bayle, P., Bayle, A., Janson, L. and Mackey, L. (2020). Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, **33**, 16339–16350. 12

Berry, S., Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, **63** (4), 841–890. 7

Berry, S. T. and Haile, P. A. (2021). Foundations of demand estimation. In *Handbook of industrial organization*, vol. 4, Elsevier, pp. 1–62. 7

Blandhol, C., Bonney, J., Mogstad, M. and Torgovitsky, A. (2022). *When is TSLS actually late?* Tech. rep., National Bureau of Economic Research. 2, 4, 5, 7, 8, 19

Borusyak, K. and Hull, P. (2023). Nonrandom exposure to exogenous shocks. *Econometrica*, **91** (6), 2155–2185. 7

Cattaneo, M. D., Crump, R. K., Farrell, M. H. and Feng, Y. (2019). On binscatter. *arXiv preprint arXiv:1902.09608*. 17, 18

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, **34** (3), 305–334. 10, 19

— (1992). Comment: Sequential moment restrictions in panel data. *Journal of Business & Economic Statistics*, **10** (1), 20–26. 2, 9, 10, 19, 24

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, **6**, 5549–5632. 12

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. 7, 10, 12

Coussens, S. and Spiess, J. (2021). Improving inference from simple instruments through compliance estimation. *arXiv preprint arXiv:2108.03726*. 7

Currie, J., Kleven, H. and Zwiers, E. (2020). Technology and big data are changing economics: mining text to track methods. In *AEA Papers and Proceedings*, vol. 110, pp. 42–48. 2

Dustmann, C., Fasani, F. and Speciale, B. (2017). Illegal migration and consumption behavior of immigrant households. *Journal of the European Economic Association*, **15** (3), 654–691. 16, 18

Graham, B. S., de Xavier Pinto, C. C. and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, **79** (3), 1053–1079. 3, 13

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, pp. 1251–1271. 12

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, **73** (3), 669–738. 14, 30

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, **113** (523), 1094–1111. 12

Lennon, C., Rubin, E. and Waddell, G. R. (2022). Machine learning (too much) in 2sls: Insights from a bias decomposition. 14

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, **5** (2), 99–135. 3, 13

Ritzwoller, D. M. and Romano, J. P. (2023). Reproducible aggregation of sample-split statistics. 10

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pp. 931–954. 15, 29

Słoczyński, T. (2022). When should we (not) interpret linear iv estimands as late? 2, 14

Vaart, A. v. d. and Wellner, J. A. (2023). Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, pp. 127–384. 10

Appendix A. Proofs

Lemma 2.2. *Any P^* that satisfies Assumptions 2.1 and 2.2 implies a distribution P for observed variables that satisfies the structural equation*

$$Y_i = \alpha_0 + W_i'\beta_0 + X_i'\eta_0 + U_i \quad \mathbb{PL}[U_i \mid Z_i; X_i] = 0 \tag{5}$$

for some (α_0, η_0). Conversely, any P that satisfies (5) can be rationalized by some P^* that satifies Assumptions 2.1 and 2.2. Moreover, the restriction $\mathbb{PL}[U_i \mid Z_i; X_i] = 0$ is equivalent to the restriction that $\mathbb{E}[U_i X_i] = \mathbb{E}[U_i \mid Z_i] = 0$.

Proof. We first show that $\mathbb{PL}[U \mid Z; X] = 0$ is equivalent to $\mathbb{E}[U \mid Z] = 0$ and $\mathbb{E}[XU] = 0$. For the $\implies$ direction, note that

$$\mathbb{E}[(U - \mathbb{E}[U \mid Z])^2] \leq \mathbb{E}[U^2]$$

with equality only if $\mathbb{E}[U \mid Z] = 0$. The fact that $\mathbb{PL}[U \mid Z; X] = 0$ means that equality is achieved. Similarly,

$$\mathbb{E}[(U - \mathbb{L}[U \mid X])^2] \leq \mathbb{E}[U^2].$$

with equality only if $\mathbb{L}[U \mid X] = 0$. Thus $\mathbb{L}[U \mid X] = 0$ and thus $\mathbb{E}[UX] = 0$. For the reverse direction, note that $\mathbb{E}[U \mid Z] = 0$ and $\mathbb{E}[XU] = 0$ satisfies the first-order condition of the optimization problem in defining $\mathbb{PL}$. Indeed, for the problem

$$\min_{f,\eta} \mathbb{E}(Y - f(Z_i) - \eta' X_i)^2,$$

at $\eta = 0$, the optimal choice is $f(Z_i) = 0$. At $f(Z_i) = 0$, the optimal choice is $\eta = 0$. Since this problem is convex in (f, η), we conclude that this certifies that $f(\cdot) = \eta = 0$ is the solution.

Take some P^* that satisfies Assumptions 2.1 and 2.2. Note that the observed outcome can then be written as

$$\begin{aligned}\mathbb{E}[Y_i \mid W_i(\cdot), Z_i = z, X_i] &= \mathbb{E}[Y_i(W_i(z)) - Y_i(w_0) \mid W_i, Z_i = z, X_i] + \mathbb{E}[Y_i(w_0) \mid W_i(\cdot), X_i, Z_i = z] \\ &= (W_i - w_0)'\beta_0 + \mathbb{E}[Y_i(w_0) \mid W_i(\cdot), X_i, Z_i = z]\end{aligned}$$

and hence

$$Y_i = (W_i - w_0)'\beta_0 + \mathbb{E}[Y_i(w_0) \mid W_i(\cdot), X_i, Z_i = z] + U_{i0} \quad \mathbb{E}[U_{i0} \mid W_i(\cdot), X_i, Z_i] = 0.$$

Observe that

$$\mathbb{E}[Y_i(w_0) \mid W_i(\cdot), X_i, Z_i] = U_{i1} + \mathbb{E}[Y_i(w_0) \mid X_i, Z_i] \quad \mathbb{E}[U_{i1} \mid X_i, Z_i] = 0$$

By Assumption 2.2, note that

$$\mathbb{PL}[\mathbb{E}[Y_i(w_0) \mid X_i, Z_i] \mid Z_i; X_i] = \mathbb{PL}[Y_i(w_0) \mid Z_i; X_i] = \mathbb{L}[Y_i(w_0) \mid X_i]$$

and we can write

$$\mathbb{E}[Y_i(w_0) \mid X_i, Z_i] = \mathbb{L}[Y_i(w_0) \mid X_i] + U_{i2} \quad \mathbb{PL}[U_{i2} \mid Z_i; X_i] = 0.$$

Thus, note that for some α_0, η_0, we can write

$$Y_i = W_i'\beta_0 + \alpha_0 + X_i'\eta_0 + U_i$$

for $U_i = U_{i0} + U_{i1} + U_{i2}$ with $\mathbb{PL}[U_i \mid Z_i; X_i] = 0$.

Conversely, let P satisfy (5). Consider a potential outcomes model in which $Y_i(w) = w'\beta_0 + \alpha_0 + X_i'\eta_0 + U_i$ where the distribution of $W_i(\cdot)$ is chosen such that

$$(W_i(Z_i) \mid Z_i, U_i) \sim (W_i \mid Z_i, U_i).$$

Then this potential outcomes model is consistent with P, in the sense that it generates the same joint distribution $(Y_i, X_i, W_i, Z_i) \sim P$. It is also immediate to check that this potential outcomes model satisfies Assumptions 2.1 and 2.2. □

Proposition 2.3. *Under random assignment $Z_i \perp\!\!\!\perp Y_i(\cdot) \mid X_i$,[23] the following are equivalent:*

(1) Assumption 2.2

(2) Z_i does not predict the nonlinearity of the mean potential outcome given X_i: For $f_0(X_i) \equiv \mathbb{E}[Y_i(w_0) \mid X_i]$,

$$\mathbb{E}[f_0(X_i) - \mathbb{L}[f_0(X_i) \mid X_i] \mid Z_i] = 0.$$

and $\mathbb{PL}[Y_i(w_0) \mid Z_i; X_i] = \mathbb{L}[f_0(X_i) \mid X_i]$.

Proof. ($\implies$) Note that

$$f_0(X_i) = \mathbb{E}[Y_i(w_0) \mid X_i, Z_i]$$

by random assignment. First, observe that

$$\mathbb{PL}[Y_i(w_0) \mid Z_i; X_i] = \mathbb{PL}[\mathbb{E}[Y_i(w_i) \mid Z_i, X_i] \mid Z_i; X_i] = \mathbb{PL}[f_0(X_i) \mid Z_i, X_i] = \alpha_0 + X_i'\eta_0$$

for some α_0, η_0 by Assumption 2.2. Since $\mathbb{PL}$ minimizes squared error and happens to be an affine function solely of X_i, we know that $\mathbb{L}[f_0(X_i) \mid X_i] = \alpha_0 + X_i'\eta_0$ and

$$\mathbb{E}[f_0(X_i) - \alpha_0 - X_i'\eta_0 \mid Z_i] = 0.$$

($\impliedby$) We can write

$$Y_i(w_0) = f_0(X_i) + U_{i1} = (f_0(X_i) - \mathbb{L}[f_0 \mid X_i]) + \mathbb{L}[f_0 \mid X_i] + U_{i1} \quad \mathbb{E}[U_{i1} \mid X_i] = \mathbb{E}[U_{i1} \mid Z_i, X_i] = 0.$$

Let $U_{i2} = f_0(X_i) - \mathbb{L}[f_0 \mid X_i]$, where by assumption $\mathbb{E}[U_{i2} \mid Z_i] = 0$ and $\mathbb{E}[UX_i] = 0$. By the second part of Lemma 2.2, we know that $\mathbb{PL}[U_{i2} \mid Z_i; X_i] = 0$. Thus, Assumption 2.2 is

[23] The proof only uses mean-independence: $\mathbb{E}[Y_i(\cdot) \mid X_i, Z_i] = \mathbb{E}[Y_i(\cdot) \mid X_i]$.

satisfied:

$$\mathbb{PL}[Y_i(w_0) \mid Z_i; X_i] = \mathbb{L}[f_0 \mid X_i] = \mathbb{L}[Y_i(w_0) \mid X_i].$$

□

Proposition 2.4. *Assumption 2.2 is equivalent to the following: For any function $g(\cdot)$,*

$$\mathbb{L}[Y_i(w_0) \mid g(Z_i), X_i] = \alpha_0 + X_i'\eta_0.$$

Proof. ($\Longleftarrow$) Suppose $\mathbb{PL}[Y_i(w_0) \mid Z_i; X_i] = g_0(Z_i) + X_i'\eta_0$. Then

$$\mathbb{L}[Y_i(w_0) \mid g_0(Z_i), X_i] = g_0(Z_i) + X_i'\eta_0.$$

By assumption, then $g_0(Z_i)$ must be zero.

($\Longrightarrow$) Note that given any g, the best partially linear function $\mathbb{PL}[Y_i(w_0) \mid Z_i; X_i] = \alpha_0 + \eta_0' X_i$ is an affine function of $g(Z_i), X_i$. Hence it is automatically equal to $\mathbb{L}[Y_i(w_0) \mid Z_i, X_i]$. □

Theorem 3.1. *The semiparametric efficiency bound for the* sequential moment restrictions (5) *is*

$$V^\star = \left(\mathbb{E}\left[\frac{1}{\sigma^2(Z_i)} \mathbb{E}[\mathbf{D}_i \mid Z_i] \mathbb{E}[\mathbf{D}_i \mid Z_i]' \right] + \mathbb{E}[\mathbf{D}_i \tilde{X}_i'] \mathbb{E}[U_i \tilde{X}_i \tilde{X}_i']^{-1} \mathbb{E}[\tilde{X}_i \mathbf{D}_i'] \right)^{-1}.$$

Under the assumptions that (i) $\mathrm{Var}(\Upsilon^\star(\mathbf{Z}_i)U_i) < \infty$ *and (ii)* $\mathbb{E}[\Upsilon^\star(\mathbf{Z}_i)\mathbf{D}_i'] < \infty$ *and is full rank,* $\sqrt{n}(\hat{\theta}_N^\star - \theta_0) \xrightarrow{d} \mathcal{N}(0, V^\star)$ *as* $N \to \infty$.

Proof. By Theorem 1 in Chamberlain (1992), the information bound for the sequential moment restriction (5) is

$$(V^\star)^{-1} = J_1 + J_2$$

where J_1, J_2 are the information bounds for $\mathbb{E}[\tilde{X}_i U_i] = 0$ and $\mathbb{E}[U_i \mid Z_i] = 0$, respectively.

We recall that for a conditional moment restriction $\mathbb{E}[m(\theta_0, A_i) \mid B_i] = 0$, the information bound is of the form

$$J = \mathbb{E}\left[\mathbb{E}[dm/d\theta \mid B_i]' \,\mathrm{Var}(m(\theta_0, A_i) \mid B_i)^{-1} \mathbb{E}[dm/d\theta \mid B_i] \right].$$

The information bound J_1 is then

$$J_1 = \mathbb{E}[\mathbf{D}_i \tilde{X}_i'] \mathbb{E}[U_i^2 \tilde{X}_i \tilde{X}_i']^{-1} \mathbb{E}[\tilde{X}_i \mathbf{D}_i']$$

and J_2 is

$$J_2 = \mathbb{E}\left[\mathbb{E}[\mathbf{D}_i \mid Z_i] \mathbb{E}[U_i^2 \mid Z_i]^{-1} \mathbb{E}[\mathbf{D}_i' \mid Z_i] \right] = \mathbb{E}\left[\frac{1}{\sigma^2(Z_i)} \mathbb{E}[\mathbf{D}_i \mid Z_i] \mathbb{E}[\mathbf{D}_i' \mid Z_i] \right].$$

This finishes the proof of the first part.

For the second part, it is easy to see that the IV estimator is asymptotically normal and has asymptotic variance equal to

$$\mathbb{E}[\Upsilon^\star(\mathbf{Z}_i)\mathbf{D}_i']^{-1}\mathbb{E}[U_i^2\Upsilon^\star(\mathbf{Z}_i)(\Upsilon^\star(\mathbf{Z}_i))']\mathbb{E}[\Upsilon^\star(\mathbf{Z}_i)\mathbf{D}_i']^{-1}.$$

Now,

$$\mathbb{E}[\Upsilon^\star(\mathbf{Z}_i)\mathbf{D}_i] = \mathbb{E}\left[\frac{1}{\sigma^2(Z_i)}\mathbb{E}[\mathbf{D}_i \mid Z_i]\mathbb{E}[\mathbf{D}_i \mid Z_i]'\right] + \mathbb{E}[\mathbf{D}_i\tilde{X}_i]\mathbb{E}[U_i^2\tilde{X}_i\tilde{X}_i']\mathbb{E}[\tilde{X}_i\mathbf{D}_i'] = (V^\star)^{-1}.$$

and

$$\begin{aligned}
\mathbb{E}[U_i^2\Upsilon^\star(\mathbf{Z}_i)(\Upsilon^\star(\mathbf{Z}_i))'] &= \mathrm{Var}(U_i\Upsilon^\star(\mathbf{Z}_i)) \\
&= \mathrm{Var}\left(U_i\frac{\mathbb{E}[\mathbf{D}_i \mid Z_i]}{\sigma^2(Z_i)}\right) + \mathrm{Var}\left(\mathbb{E}[\mathbf{D}_i\tilde{X}_i]\mathbb{E}[U_i^2\tilde{X}_i\tilde{X}_i']^{-1}U_i\tilde{X}_i\right) \\
&\qquad\qquad (\mathrm{Cov}(U_i\tilde{X}_i, U_i \mid Z_i) = 0) \\
&= (V^\star)^{-1}.
\end{aligned}$$

Hence the asymptotic variance is equal to $V^\star$. □

Theorem 3.2. *For a cross-fitted and partially linear estimate $\hat{\Upsilon}_N(\mathbf{Z}_i)$, assume that*

(1) (Convergence in $L^2(\mathbf{Z})$ to a limiting instrument) $\hat{\Upsilon}_N$ converges to some partially linear $\Upsilon_0(\mathbf{Z}_i) = f_1(Z_i) + X_i'\gamma_0$ in the following out-of-sample sense

$$\lim_{N\to\infty}\mathbb{E}\left[\|\hat{\Upsilon}_N(\mathbf{Z}_i^*) - \Upsilon_0(\mathbf{Z}_i^*)\|^2\right] = \lim_{N\to\infty}\mathbb{E}\left[\|\hat{\Upsilon}_N - \Upsilon_0\|_{L^2(\mathbf{Z})}^2\right] = 0.$$

Here, the expectation integrates over a distribution where $\hat{\Upsilon}_N$ depends on $N/2$ iid copies of (Y_i, X_i, W_i, Z_i) and $\mathbf{Z}_i^*$ is an additional iid copy of $\mathbf{Z}_i$.

(2) (Finite moments) The following moments are finite: $\mathbb{E}[\mathbf{D}_i\mathbf{D}_i'] \le \infty$, $\mathbb{E}[|U_i|^3\|\Upsilon_0(\mathbf{Z}_i)\|^3] < \infty$, $\mathrm{Var}(U_i \mid \mathbf{Z}_i) \le M$ uniformly over $\mathbf{Z}_i$.

(3) (Limiting instrument is strong) $\mathbb{E}[\Upsilon_0(\mathbf{Z}_i)\mathbf{D}_i']$ is full rank.

Let

$$\hat{\theta}_N = \left(\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)\mathbf{D}_i'\right)^{-1}\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)Y_i \tag{12}$$

and let $\hat{\theta}_{0,N} = \left(\frac{1}{N}\sum_i \Upsilon_0(\mathbf{Z}_i)\mathbf{D}_i'\right)^{-1}\frac{1}{N}\sum_i \Upsilon_0(\mathbf{Z}_i)Y_i$.[24] Then:

(1) $\hat{\theta}_N$ and $\hat{\theta}_{0,N}$ are asymptotically equivalent

$$\sqrt{N}(\hat{\theta}_N - \hat{\theta}_{0,N}) \xrightarrow{p} 0.$$

and both are asymptotically normal.

[24] Since we assume $\mathbb{E}[\Upsilon_0\mathbf{D}_i']$ is full rank, with probability tending to 1 both $\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)\mathbf{D}_i'$ and $\frac{1}{N}\sum_i \hat{\Upsilon}_0(\mathbf{Z}_i)\mathbf{D}_i'$ are full rank. In the event that they are not, we may replace the inverse operation with Moore–Penrose pseudoinverse.

(2) The asymptotic variance of $\hat{\theta}_N$ is equal to the efficiency bound in Theorem 3.1 if $\hat{\Upsilon}_N$ is consistent for the optimal instrument: i.e., $\Upsilon_0 = \Upsilon^\star$.

(3) Assuming additionally that $\mathbf{D}_i$ and $\hat{\Upsilon}_N(\mathbf{Z}_i)$ has a uniformly bounded fourth moment conditional on $\hat{\Upsilon}_N(\cdot)$ almost surely, the asymptotic variance is consistently estimated by its empirical counterpart

$$\hat{V}_N = \left(\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)\mathbf{D}_i'\right)^{-1}\left(\frac{1}{N}\sum_i (Y_i - \hat{\theta}_N'\mathbf{D}_i)^2\hat{\Upsilon}_N(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)'\right)\left(\frac{1}{N}\sum_i \mathbf{D}_i\hat{\Upsilon}_N(\mathbf{Z}_i)'\right)^{-1}.$$

Proof. (1) We first show that

$$\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)\mathbf{D}_i' = \frac{1}{N}\sum_i \Upsilon_0(\mathbf{Z}_i)\mathbf{D}_i' + o_p(1)$$

and

$$\frac{1}{\sqrt{N}}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)U_i = \frac{1}{\sqrt{N}}\sum_i \Upsilon_0(\mathbf{Z}_i)U_i + o_p(1).$$

For the first claim, note that for the Frobenius norm $\|\cdot\|_F$

$$\begin{aligned}
\left\|\frac{1}{N}\sum_i (\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))\mathbf{D}_i'\right\|_F &\le \frac{1}{N}\sum_i \|(\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))\mathbf{D}_i'\|_F \\
&\le \frac{1}{N}\sum_i \|\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i)\|\|\mathbf{D}_i\| \\
&\le \left(\frac{1}{N}\sum_i \|\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i)\|^2\right)^{1/2}\left(\frac{1}{N}\sum_i \|\mathbf{D}_i\|^2\right)^{1/2} \\
&\qquad \text{(Cauchy–Schwarz)}
\end{aligned}$$

Note that since

$$\mathbb{E}\left[\|\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i)\|^2\right] \to 0,$$

Markov's inequality implies that

$$\frac{1}{N}\sum_i \|\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i)\|^2 = o_p(1).$$

By the finiteness of $\mathbb{E}[\mathbf{D}_i\mathbf{D}_i']$, $\frac{1}{N}\sum_i \|\mathbf{D}_i\|^2 = O_p(1)$. Hence,

$$\left\|\frac{1}{N}\sum_i (\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))\mathbf{D}_i'\right\|_F = o_p(1).$$

For the second claim, note that by Chebyshev's inequality, for one fold,

$$\frac{1}{\sqrt{N}}\sum_{i\in\mathcal{I}_1}\left((\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))U_i\right) = O_p\left(\sqrt{\mathrm{Var}\left[(\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))U_i \mid \hat{\Upsilon}_N\right]}\right)$$

$$= O_p\left(\sqrt{\mathbb{E}[\mathbb{E}[U_i^2 \mid \mathbf{Z}_i](\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))^2 \mid \hat{\Upsilon}_N]}\right)$$
$$= O_p\left(\sqrt{\mathbb{E}\left[(\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))^2 \mid \hat{\Upsilon}_N\right]}\right)$$
$$= o_p(1).$$

This implies that

$$\frac{1}{\sqrt{N}}\sum_i \left((\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))U_i\right) = o_p(1).$$

Since $\mathbb{E}[\Upsilon_0 \mathbf{D}_i']$ is full rank, this implies that

$$\sqrt{N}\left(\hat{\theta}_n - \hat{\theta}_{0,N}\right) = o_p(1).$$

It suffices to then show that

$$\sqrt{N}\left(\hat{\theta}_{0,N} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, V)$$

This is a simple application of the central limit theorem for

$$\frac{1}{\sqrt{N}}\sum_i \Upsilon_0(\mathbf{Z}_i)U_i,$$

which is satisfied say under $\mathbb{E}[|U_i|^3\|\Upsilon_0(\mathbf{Z}_i)\|^3] < \infty$. By standard results on TSLS, the variance matrix is

$$V = \mathbb{E}[\Upsilon_0(\mathbf{Z}_i)\mathbf{D}_i']^{-1}\mathbb{E}[U_i^2\Upsilon_0(\mathbf{Z}_i)\Upsilon_0(\mathbf{Z}_i)']\mathbb{E}[\Upsilon_0(\mathbf{Z}_i)\mathbf{D}_i']^{-1}.$$

(2) We showed in the proof to Theorem 3.1 that the efficiency bound is equal to

$$\mathbb{E}[\Upsilon^\star(\mathbf{Z}_i)\mathbf{D}_i']^{-1}\mathbb{E}[U_i^2\Upsilon^\star(\mathbf{Z}_i)\Upsilon^\star(\mathbf{Z}_i)']\mathbb{E}[\Upsilon^\star(\mathbf{Z}_i)\mathbf{D}_i']^{-1},$$

which is equal to V if $\Upsilon_0 = \Upsilon^\star$.

(3) We have already shown that

$$\frac{1}{N}\sum_i \hat{\Upsilon}_N(\mathbf{Z}_i)\mathbf{D}_i' = \frac{1}{N}\sum_i \hat{\Upsilon}_0(\mathbf{Z}_i)\mathbf{D}_i' + o_p(1) \xrightarrow{p} \mathbb{E}[\Upsilon_0\mathbf{D}_i'].$$

It suffices to show that

$$\frac{1}{N}\sum_i (Y_i - \hat{\theta}_N'\mathbf{D}_i)^2\hat{\Upsilon}_N(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)' \xrightarrow{p} \mathbb{E}[U_i^2\Upsilon_0(\mathbf{Z}_i)\Upsilon_0(\mathbf{Z}_i)'].$$

Now, observe that

$$(Y_i - \hat{\theta}_N'\mathbf{D}_i)^2 - U_i^2 = (\theta_0 - \hat{\theta}_N)'\mathbf{D}_i(2U_i + (\theta_0 - \hat{\theta}_N)'\mathbf{D}_i)$$

Thus

$$\begin{aligned}
&\frac{1}{N}\sum_i (Y_i - \hat{\theta}_N'\mathbf{D}_i)^2 \hat{\Upsilon}_N(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)' \\
&= \frac{1}{N}\sum_i U_i^2 \hat{\Upsilon}_N(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)' + (\theta_0 - \hat{\theta}_N)' \underbrace{\frac{1}{N}\sum_i (2U_i + (\theta_0 - \hat{\theta}_N)'\mathbf{D}_i)\mathbf{D}_i \hat{\Upsilon}_N(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)'}_{=O_p(1), \text{ by the existence of the fourth moment.}} \\
&= \frac{1}{N}\sum_i U_i^2 \hat{\Upsilon}_N(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)' + o_p(1).
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{1}{N}\sum_i U_i^2 \hat{\Upsilon}_N(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)' &= \frac{1}{N}\sum_i U_i^2 (\hat{\Upsilon}_N(\mathbf{Z}_i) - \Upsilon_0(\mathbf{Z}_i))\hat{\Upsilon}_N(\mathbf{Z}_i)' + \frac{1}{N}\sum_i U_i^2 \Upsilon_0(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)' \\
&= o_p(1) + \frac{1}{N}\sum_i U_i^2 \Upsilon_0(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)'
\end{aligned}$$

by Cauchy–Schwarz and the mean-square convergence condition.

Similarly,

$$\frac{1}{N}\sum_i U_i^2 \Upsilon_0(\mathbf{Z}_i)\hat{\Upsilon}_N(\mathbf{Z}_i)' = \frac{1}{N}\sum_i U_i^2 \Upsilon_0(\mathbf{Z}_i)\Upsilon_0(\mathbf{Z}_i)' + o_p(1).$$

This completes the proof.

□

Lemma 3.4. *When $\mathbb{E}[U_i^2 \mid \mathbf{Z}_i] = 1$ is a constant normalized to 1, the optimal instrument for (5) is the partially linear regression of $\mathbf{D}_i$ on Z_i, X_i:*

$$\Upsilon^\star(\mathbf{Z}_i) = \mathbb{PL}[\mathbf{D}_i \mid Z_i; X_i] = [1, \mathbb{PL}[W_i \mid Z_i; X_i]', X_i']'. \tag{13}$$

Proof. Recall that

$$\Upsilon^\star(\mathbf{Z}_i) = \frac{\mathbb{E}[\mathbf{D}_i \mid Z_i]}{\sigma^2(Z_i)} + \mathbb{E}[\mathbf{D}_i \tilde{X}_i']\mathbb{E}[U_i^2 \tilde{X}_i \tilde{X}_i']^{-1}\tilde{X}_i.$$

Under these assumptions,

$$\begin{aligned}
\tilde{X}_i &= X_i - \mathbb{E}[X_i \mid Z_i] \\
\sigma^2(Z_i) &= 1 \\
\mathbb{E}[U_i^2 \tilde{X}_i \tilde{X}_i'] &= \mathbb{E}[\tilde{X}_i \tilde{X}_i'] \\
\mathbb{E}[\mathbf{D}_i \tilde{X}_i'] = \mathbb{E}[(\mathbf{D}_i - \mathbb{E}[\mathbf{D}_i \mid Z_i])\tilde{X}_i'] &= \begin{bmatrix} 0 \\ \mathbb{E}[(W_i - \mathbb{E}[W_i \mid Z_i])(X_i - \mathbb{E}[X_i \mid Z_i])'] \\ \mathbb{E}[\tilde{X}_i \tilde{X}_i'] \end{bmatrix}
\end{aligned}$$

$$\mathbb{E}[\mathbf{D}_i \mid Z_i] = \begin{bmatrix} 1 \\ \mathbb{E}[W_i \mid Z_i] \\ \mathbb{E}[X_i \mid Z_i] \end{bmatrix}$$

Thus

$$\mathbb{E}[\mathbf{D}_i \tilde{X}_i'] \mathbb{E}[U_i^2 \tilde{X}_i \tilde{X}_i']^{-1} \tilde{X}_i = \begin{bmatrix} 0 \\ \beta_{\tilde{W} \sim \tilde{X}} \tilde{X}_i \\ \tilde{X}_i \end{bmatrix}$$

where $\beta_{\tilde{W} \sim \tilde{X}}$ is the $\dim W \times \dim X$ matrix of regression coefficients of $W - \mathbb{E}[W \mid Z]$ on $X - \mathbb{E}[X \mid Z]$. Thus

$$\Upsilon^\star(\mathbf{Z}) = \begin{bmatrix} 1 \\ \mathbb{E}[W_i \mid Z_i] + \beta_{\tilde{W} \sim \tilde{X}} \tilde{X}_i \\ X_i \end{bmatrix}$$

Finally,

$$\mathbb{E}[W_i \mid Z_i] + \beta_{\tilde{W} \sim \tilde{X}} = \mathbb{PL}[W_i \mid Z_i; X_i]$$

by Robinson (1988). □

Proposition 3.5. *Suppose W_i is binary and suppose that*

$$W_i(z) = \mathbb{1}(u(z) > V_i(X_i)) \qquad \text{(Strong monotonicity)}$$

for some random $V_i(\cdot)$ with the marginal distribution of $V_i = V_i(X_i)$ normalized to Unif$[0, 1]$. *Suppose $\Upsilon(Z_i)$ is mean zero such that*

$$\mathbb{PL}[W_i \mid Z_i; X_i] = \Upsilon(Z_i) + [1, X_i'] \kappa_0.$$

Define the marginal treatment effect at v to be

$$\text{MTE}(v) = \mathbb{E}[Y_i(1) - Y_i(0) \mid V_i = v].$$

Then, suppose Z_i is fully randomly assigned $Z_i \perp\!\!\!\perp (X_i, Y_i(\cdot), W_i(\cdot))$. Then the estimand for β in (14) is a convex *average of marginal treatment effects:*

$$\beta = \int_0^1 \frac{\mathbb{E}[\Upsilon(Z_i) \mathbb{1}(u(Z_i) > v)]}{\mathbb{E}[\Upsilon(Z_i) u(Z_i)]} \text{MTE}(v) \, dv.$$

Proof. Since Z_i is randomly assigned, note that $\tilde{\Upsilon}(Z_i) = \Upsilon(Z_i) - \mathbb{L}[\Upsilon \mid X_i] = \Upsilon(Z_i)$. Thus the estimand for β in (14) is

$$\beta = \frac{\mathbb{E}[\Upsilon(Z_i) Y_i]}{\mathbb{E}[\Upsilon(Z_i) W_i]} = \frac{\mathbb{E}[\Upsilon(Z_i) Y(0)] + \mathbb{E}[\Upsilon(Z_i) W_i (Y(1) - Y(0))]}{\mathbb{E}[\Upsilon(Z_i) W_i]} = \frac{\mathbb{E}[\Upsilon(Z_i) W_i (Y(1) - Y(0))]}{\mathbb{E}[\Upsilon(Z_i) W_i]}$$

This is the estimand of a TSLS regression of Y_i on $1, W_i$, instrumenting with $1, \Upsilon(Z_i)$. Note that

$$\mathbb{E}[W_i \mid Z_i] = u(Z_i) = \Upsilon(Z_i) + \mathbb{E}[W_i]$$

The reason is that the condition

$$\mathbb{PL}[W_i \mid Z_i; X_i] = \Upsilon(Z_i) + [1, X_i']\eta_0$$

implies that $\mathbb{E}[W_i - [1, X_i']\eta_0 \mid Z_i] = \Upsilon(Z_i)$. Since $\mathbb{E}[[1, X_i']\eta_0 \mid Z_i]$ is a constant, we conclude that it must equal to $\mathbb{E}[W_i]$. The random assignment of Z_i means that $(W_i, V_i, Y_i(1), Y_i(0), \Upsilon(Z_i))$ satisfies the assumptions of the MTE model in Heckman and Vytlacil (2005). By section 4.3 of Heckman and Vytlacil (2005),

$$\beta = \int_0^1 \omega(v)\mathbb{E}[Y_i(1) - Y_i(0) \mid V_i = v]\, dv$$

where

$$\omega(v) = \frac{\mathbb{E}[\Upsilon(Z_i)\mathbb{1}(u(Z_i) > v)]}{\mathbb{E}[\Upsilon(Z_i)u(Z_i)]} \geq 0$$

since

$$\mathbb{E}[\Upsilon(Z_i)\mathbb{1}(u(Z_i) > v)] = \mathbb{E}[u(Z_i)\mathbb{1}(u(Z_i > v))] - \mathbb{E}[u(Z_i)]\,\mathrm{P}(u(Z_i) > v) \geq 0.$$

□