Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models

Jiafeng Chen
Harvard Business School
Boston, MA
jchen@hbs.edu

Daniel L. Chen
Toulouse School of Economics
Toulouse, France
dlchen@nber.org

Greg Lewis
Microsoft Research
Cambridge, MA
glewis@microsoft.com

Abstract

We provide some simple theoretical results that justify incorporating machine learning in a standard linear instrumental variable setting, prevalent in empirical research in economics. Machine learning techniques, combined with sample-splitting, extract nonlinear variation in the instrument that may dramatically improve estimation precision and robustness by boosting instrument strength. The analysis is straightforward in the absence of covariates. The presence of linearly included exogenous covariates complicates identification, as the researcher would like to prevent nonlinearities in the covariates from providing the identifying variation. Our procedure can be effectively adapted to account for this complication, based on an argument by Chamberlain (1992). Our method preserves standard intuitions and interpretations of linear instrumental variable methods and provides a simple, user-friendly upgrade to the applied economics toolbox. We illustrate our method with an example in law and criminal justice, examining the causal effect of appellate court reversals on district court sentencing decisions.

1 Introduction

Instrumental variable (IV) designs are a popular method in empirical economics. Over 30% of all NBER working papers and top journal publications considered by Currie *et al.* (2020) include some discussion of instrumental variables. The vast majority of IV designs used in practice are linear IV estimated via two-stage least squares (TSLS), a familiar technique in standard graduate introductions to econometrics and causal inference (e.g. Angrist and Pischke, 2008). Standard TSLS, however, leaves on the table some variation provided by the instruments that may improve precision of estimates, as TSLS only exploits variation that is linearly related to the endogenous regressors. In the event that the instrument has a low linear correlation with the endogenous variable, but nevertheless predicts the endogenous variable well through a nonlinear transformation, we should expect TSLS to perform poorly in terms of estimation precision and inference robustness. In particular, in some cases, TSLS would provide spuriously precise but biased estimates (due to weak instruments, see Andrews *et al.*, 2019). Such nonlinear settings become increasingly plausible when exogenous variation includes high dimensional data or alternative data, such as text, images, or other complex attributes. We show that off-the-shelf machine learning techniques provide a general-purpose toolbox for leveraging such complex variation, improving instrument strength and estimate quality.

Of course, replacing the first stage linear regression with more flexible specifications does not come without cost in terms of identifying assumptions. The validity of TSLS hinges only upon the exclusion restriction that the instrument is linearly uncorrelated with unobserved disturbances in the response variable. Relaxing the linearity requires that endogenous residuals are mean zero conditional on the exogenous instruments, which is stronger. However, it is rare that a researcher has a compelling reason to believe the weaker non-correlation assumption, but rejects the slightly stronger mean-independence

Preprint. Under review.

assumption. In fact, by not exploiting the nonlinearities, TSLS may accidentally make a strong instrument weak, and deliver spuriously precise inference: Dieterle and Snell (2016) and references therein find that several applied microeconomics papers have conclusions that are sensitive to the specification (linear vs. quadratic) of the first-stage.

A more serious identification concern to leveraging machine learning in the first-stage comes from the parametric functional form in the second stage. When there are exogenous covariates that are included in the parametric structural specification, nonlinear transformations of these covariates could in principle be valid instruments, and provide variation that pins down the parameter of interest. For example, in the standard IV setup of $Y = D\tau + X\beta + U$ where X is an exogenous covariate, imposing $\mathbb{E}[U|X] = 0$ would formally result in X^2, X^3, etc. being valid "excluded" instruments. However, given that the researcher's stated source of identification comes from excluded instruments, such "identifying variation" provided by covariates is more of an artifact of parametric specification than any serious information from the data that relates to the researcher's scientific inquiry.

One principled response to the above concern is to make the second stage structural specification likewise nonparametric or semiparametric, thereby including an infinite dimensional parameter to estimate, making the empirical design a *nonparametric instrumental variable* (NPIV) design. Significant theoretical and computational progress have been made in this regard (Newey and Powell, 2003; Ai and Chen, 2003, 2007; Horowitz and Lee, 2007; Severini and Tripathi, 2012; Ai and Chen, 2012; Hartford *et al.*, 2017; Dikkala *et al.*, 2020; Chen *et al.*, 2020a,b, and references therein). However, regrettably, NPIV has received relatively little attention in applied work in economics, potentially due to theoretical complications regarding identification,[1] difficulty in interpretation and troubleshooting, and computational scalability. Moreover, in some cases parametric restrictions on structural functions come from theoretical considerations or techniques like log-linearization, where estimated parameters have intuitive theoretical interpretation and policy relevance; in these cases the researcher may have compelling reasons to stick with parametric specifications.

In the spirit of being user-friendly to practitioners, this paper considers estimation and inference in an instrumental variable model where the second stage structural relationship is linear, while allowing for as much nonlinearity in the instrumental variable as possible, without creating unintended and spurious identifying variation from included covariates. We document some simple results via elementary techniques, which provide intuition and justification for using machine learning methods in instrumental variable designs. We show that with sample-splitting, under weak consistency conditions, a simple estimator that uses the predicted values of endogenous and included regressors as technical instruments is consistent, asymptotically normal, and semiparametrically efficient; the constructed instrumental variable also readily provides weak instrument diagnostics and robust procedures, such as the Anderson–Rubin procedure. Moreover, standard diagnostics like out-of-sample prediction quality are directly related to quality of estimates. In the presence of exogenous covariates that are parametrically included in the second-stage structural function, adapting machine learning techniques requires caution to avoid spurious identification from functional forms of the included covariates. To that end, we formulate and analyze the problem as a sequential moment restriction, and develop estimators that utilize machine learning for extracting nonlinear variation from and only from instruments.

We conclude the introduction by relating our work to the technical and applied literatures. The core techniques that allow for the construction of our estimators follow from Chamberlain (1987, 1992). The ideas in our proofs are also familiar in the double machine learning (Chernozhukov *et al.*, 2018; Belloni *et al.*, 2012) and semiparametrics literatures (e.g. Liu *et al.*, 2020); our arguments, however, follow from elementary techniques that are familiar to graduate students and are self-contained. Our proposed estimator is, of course, similar to the split-sample IV or jackknife IV estimators in Angrist *et al.* (1999), but we do not restrict ourselves to linear settings or linear smoothers. Using machine learning in instrumental variable settings is considered by Hansen and Kozbur (2014) (for ridge), Belloni *et al.* (2012) (for lasso), and Bai and Ng (2010) (for boosting), among others; and our work can be viewed as providing a simple, unified analysis for practitioners, much in the spirit of Chernozhukov *et al.* (2018). To the best of our knowledge, we are the first to formally explore practical complications of making the first stage nonlinear in a context with exogenous covariates. Finally, we view our work as counterpoint to the recent work by Angrist and Frandsen (2019), which

[1]The NPIV estimation problem is an instance of an *ill-posed inverse problem*, meaning that very different parameter values could lead to similar values in the optimization objective.

is more pessimistic about combining machine learning with instrumental variables—a point we explore in detail in Section 2.3.

2 Main theoretical results

We consider the standard cross-sectional setup where $(Y_i, D_i, X_i, W_i)_{i=1}^N \overset{\text{i.i.d.}}{\sim} P$ are sampled from some infinite population. Y_i is some outcome variable, D_i is a set of endogenous treatment variables, X_i is a set of exogenous controls (which includes a constant), and W_i is a set of instrumental variables. The researcher is willing to argue that W_i is exogenously or quasi-experimentally assigned. Moreover, the researcher believes that W_i provides a source of variation that "identifies" an effect of D_i on Y_i. We denote the endogenous variables and covariates as $T_i = [D_i, X_i]$ and the excluded instrument and covariates as the *technical instruments* $Z_i = [W_i, X_i]$.

A typical specification in empirical economics is the linear IV setup:

$$\begin{aligned} Y_i &= D_i^\intercal \tau + X_i^\intercal \beta + U_i \equiv T_i^\intercal \theta + U_i \\ D_i &= W_i \Pi + X_i \Gamma + V_i \equiv Z_i \Lambda + V_i \qquad \iff \qquad \mathbb{E}[Z_i(Y_i - T_i^\intercal \theta)] = 0. \\ (U_i, V_i) &\perp (X_i, Z_i), \quad \mathbb{E}[U_i] = \mathbb{E}[V_i] = 0 \end{aligned} \tag{Linear IV}$$

where $A \perp B$ means $\text{Cov}(A, B) = 0$.

We argue that, in many settings, the researcher is willing to assume more than that U, V are uncorrelated from X, Z. Common introductions of instrumental variables (Angrist and Pischke, 2008; Angrist and Krueger, 2001) stress that instruments induce variation in D and is otherwise unrelated to U, and that a common source of instruments is natural experiments. We argue that these narratives imply a stronger form of exogeneity than TSLS requires—the researcher is perhaps willing to assume mean independence $\mathbb{E}[U_i|W_i] = 0$ beyond $\text{Cov}(U_i, W_i) = 0$. After all, a symmetric mean-zero random variable S is uncorrelated with S^2, but one would hardly be comfortable justifying S^2 as unrelated to S. We demonstrate that strengthening the exogeneity assumption to mean-independence allows the researcher to extract more identifying variation from instruments, but doing so requires more flexible machinery for dealing with the first stage.

2.1 Simple case: homoskedasticity and no covariates

Let us first consider the case in which X_i contains only a constant, and we have homoskedastic errors, i.e. $\mathbb{E}[U_i^2|Z_i] = \sigma^2$. Our mean-independence restrictions give rise to a conditional moment restriction, $\mathbb{E}[Y_i - T_i^\intercal \theta \mid W_i] = 0$. By definition of conditional expectations, the conditional moment restriction encodes an infinite set of unconditional moment restrictions in the form of orthogonalities from the prediction errors:

$$\text{For all square integrable } \upsilon : \mathbb{E}[\upsilon(W_i)(Y_i - T_i^\intercal \theta)] = 0.$$

Chamberlain (1987) finds that all relevant statistical information in a conditional moment restriction is in fact contained in a single unconditional moment restriction indexed by an *optimal instrument* Υ, and the unconditional moment restriction with the optimal instrument delivers semiparametrically efficient estimation and inference.[2] In our case, $\Upsilon(W_i) = \mathbb{E}[T_i \mid W_i] = [\mathbb{E}[D_i|W_i], 1]^\intercal$. It is then natural to estimate Υ with $\hat{\Upsilon}$ and form a plug-in estimator for θ:

$$\hat{\theta} = \left(\frac{1}{N} \sum_{i=1}^N \hat{\Upsilon}(W_i) T_i^\intercal \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N \hat{\Upsilon}(W_i) Y_i \right).$$

Note that estimating Υ amounts to predicting D_i with W_i, which is well-suited to machine learning techniques. One might worry that the preliminary estimation of $\hat{\Upsilon}$ complicates asymptotic analysis of $\hat{\theta}$. Under a simple sampling-splitting scheme, however, we state a high-level condition for

[2] That is, for a conditional moment restriction of the form $\mathbb{E}[m(Y, T, \theta) \mid W] = 0$, it is sufficient to consider the unconditional moment restriction $\mathbb{E}[\Upsilon(W) m(Y, T, \theta)] = 0$, where the optimal instrument takes the form $\Upsilon(W) \propto \mathbb{E}[m^2(Y, T, \theta) \mid W]^{-1} \mathbb{E}\left[\partial m / \partial \theta \mid W\right]$. Intuitively, the optimal instrument takes a signal-to-noise ratio form: Larger values of $\mathbb{E}\left[\partial m(Y, T, \theta)/\partial \theta \mid W\right]$ indicates that the moment condition is sensitive to θ at W and hence represents a notion of signal, and $\mathbb{E}[m^2|W]$ represents some notion of noise.

consistency, normality, and efficiency of $\hat{\theta}$. Though it simplifies the proof and weakens regularity conditions, sample-splitting does reduce the size of data used to estimate the optimal instrument Υ, but such problems can be effectively mitigated by k-fold sample-splitting: 20-fold sample-splitting, for instance, limits the loss of data to 5% at the cost of 20 computations that can be effectively parallelized. Such concerns notwithstanding, we focus our exposition to two-fold sample-splitting.

Specifically, assume $N = 2n$ for simplicity and let $S_1, S_2 \subset [N]$ be the two subsamples with size n. For $j \in \{1, 2\}$, let $\hat{\Upsilon}_j$ be an estimated prediction function predicting D_i with W_i, estimated with data from the other sample, S_{-j}. $\hat{\Upsilon}$ may be a neural network or a random forest trained via empirical risk minimization, or a penalized linear regression such as elastic net.[3] The estimated instrument $\hat{\Upsilon}(W_i)$ is then formed by evaluating $\hat{\Upsilon}_j$ at $W_i, i \in S_j$. Crucially, the sampling-splitting allows that, conditional on S_{-j}, $\hat{\Upsilon}_j$ may be viewed as nonrandom, and $(Y_i, D_i, \hat{\Upsilon}(W_i))$ may be viewed as independently distributed over i. We term the resulting estimator the machine learning split-sample estimator (MLSS) estimator.

Theorem 2.1 shows that the MLSS estimator is consistent, asymptotically normal, and semiparametrically efficient when the first stage estimation of $\mathbb{E}[D|W]$ is consistent in L^2 norm; the MLSS estimator remains consistent and asymptotically normal when the L^2 consistency condition fails, but a set of weaker conditions hold, which govern the limiting behavior of sample means involving $\hat{\Upsilon}(W_i)$. The L^2 consistency condition is not strong—in particular, it is weaker than the L^2 consistency at $o(N^{-1/4})$-rate condition commonly used in the double machine learning literature (Chernozhukov *et al.*, 2018), where such conditions are considered mild.[4]

Theorem 2.1. Let $\hat{\theta}_{\text{MLSS}}$ be the machine learning split-sample estimator described above. Under Condition 1 in the appendix, which governs the concentration of various terms, $\hat{\theta}_{\text{MLSS}}$ is consistent and asymptotically normal for some G, Ω defined by Condition 1:

$$(\sigma^2 G \Omega G^\intercal)^{-1/2} \sqrt{N} (\hat{\theta}_{\text{MLSS}} - \theta) \rightsquigarrow \mathcal{N}(0, 1).$$

Moreover, if $\mathbb{E}[D_p|W] \neq 0$ for all entries p and the machine learning estimator $\hat{\Upsilon}$ is consistent in L^2, i.e. $\mathbb{E}[\|\hat{\Upsilon}_j(W_i) - \Upsilon(W_i)]^2 \mid S_{-j}\| \xrightarrow{p} 0$, then parts 1 and 3 in Condition 1 automatically hold with $G^{-1} = \mathbb{E}[\Upsilon(W_i) T_i^\intercal]$, $\Omega = \mathbb{E}[\Upsilon(W_i)\Upsilon^\intercal(W_i)]$, and the asymptotic variance $\sigma^2 G \Omega G^\intercal$ achieves the semiparametric efficiency bound.

We provide some intuition on the proof for Theorem 2.1, where we assume $\dim D = 1$ and $X = \emptyset$ for simplicity, so that Υ outputs a scalar. The estimation error is of the form $\hat{G} N^{-1/2} \sum_{i=1}^N \hat{\Upsilon}(W_i) U_i$. Within each sample, conditioned on the other sample, the terms $\hat{\Upsilon}(W_i) U_i$ are i.i.d., and we should expect a central limit theorem to hold. Furthermore, suppose the first-stage estimation is L^2 consistent, i.e. $\mathbb{E}[[\hat{\Upsilon}(W_i) - \Upsilon(W_i)]^2] = o(1)$. Note that, under the heuristic implications of the central limit theorem, the distance from the feasible estimator and the oracle estimator with known $\Upsilon(W_i)$ is

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N [\hat{\Upsilon}(W_i) - \Upsilon(W_i)] U_i \overset{d}{\approx} \mathcal{N}\left(0, \sigma^2 \frac{1}{N} \sum_{i=1}^N [\hat{\Upsilon}(W_i) - \Upsilon(W_i)]^2\right).$$

The variance term is approximately the L^2 distance between $\hat{\Upsilon}$ and Υ, and thus vanishes by assumption. Thus the asymptotic distribution of the MLSS estimator is the same as the oracle IV estimator when we know the efficient instrument $\Upsilon(W_i)$, and MLSS is semiparametrically efficient.

2.2 Covariates and heteroskedasticity

The presence of covariates X_i complicates the analysis considerably. Under the researcher's model, both W_i and X_i are considered exogenous, and thus we may assume $\mathbb{E}[U_i \mid Z_i] = 0$ and use it as a conditional moment restriction, under which the efficient instrument is $\text{Var}(U_i|Z_i)^{-1}\mathbb{E}[T_i|Z_i]$ and our analysis from the previous section continues to apply *mutatis mutandis*. However, if the researcher maintains a linear specification $Y_i = T_i^\intercal \theta + U_i$, estimating θ based on the conditional

[3] With k-fold sample-splitting, S_{-j} is the union of all samples other than the j-th one.

[4] The nuisance parameter $\mathbb{E}[D|W]$ in this setting enjoys higher-order orthogonality property described in Mackey *et al.* (2018). In particular, it is infinite-order orthogonal, thereby requiring no rate condition to work. Intuitively, estimation error in $\Upsilon(\cdot)$ has no effect on the moment condition $\mathbb{E}[\Upsilon(Y - T^\intercal \theta)] = 0$ holding, and this feature of the problem makes the estimation robust to estimation of Υ.

moment restriction $\mathbb{E}[U_i \mid Z_i] = 0$ may inadvertently "identify" θ through nonlinear behavior in X_i rather than the variation in W_i. Such a specification may allow the researcher to precisely estimate θ even when the instrument W_i is completely irrelevant, when, say, higher-order polynomial terms in the scalar X_i, X_i^2, X_i^3, are strongly correlated with D_i, perhaps due to misspecification of the linear moment condition. There may well be compelling reasons why these nonlinear terms in X_i allow for identification of τ under an economic or causal model; however, they are likely not the researcher's stated source of identification, and allowing their influence to leak into the estimation procedure undermines credibility of the statistical exercise.

One idea to resolve such a conundrum is to make the structural function nonparametric as well, and convert the model to a nonparametric instrumental variable regression (Newey and Powell, 2003; Ai and Chen, 2003) (See Appendix B for discussion). Another idea, which we undertake in this paper, is to weaken the moment condition and rule nonlinearities in X_i as inadmissible for inference.

The moment condition $\mathbb{E}[U_i \mid Z_i] = 0$ is equivalent to

$$\text{For all (square integrable) } \upsilon,\ \mathbb{E}[\upsilon(W_i, X_i)(Y - T_i^\intercal \theta)] = 0,$$

which is too strong, since it allows nonlinear transforms of X to be valid instruments. A natural idea is to restrict the class of allowable instruments $\upsilon(W_i, X_i)$ to those that are partially linear in X_i, $\upsilon(W_i, X_i) = h(W_i) + X_i^\intercal \ell$, thereby deliberately discarding information from nonlinear transformations of X. Doing so yields the following

$$\text{For all (square integrable) } \upsilon,\ \mathbb{E}[\upsilon(W_i)(Y - T_i^\intercal \theta)] = \mathbb{E}[X_i(Y_i - T_i^\intercal \theta)] = 0,$$

which is equivalent to the *sequential moment restriction*

$$\mathbb{E}[X_i(Y_i - T_i^\intercal \theta)] = \mathbb{E}[Y_i - T_i^\intercal \theta \mid W_i] = 0. \tag{1}$$

We see that (1) is a natural interpolation between the usual unconditional moment condition, $\mathbb{E}[(X_i, W_i)^\intercal \cdot U_i] = 0$, and the conditional moment restriction that may be spurious $\mathbb{E}[U_i \mid X_i, W_i] = 0$, by only allowing nonlinear information in W_i to be used for estimation and inference.

Due to conditioning over different random variables, efficient estimation in sequential moment restrictions is not straightforward. Sequential moment restrictions are analyzed in Chamberlain (1992), and we now review his argument, which motivates the construction of our estimator. Chamberlain (1992)'s characterization of efficient estimation proceeds from the conditional moment restriction with the finest conditioning, and sequentially orthogonalizes coarser moment restrictions against the finer ones, so as to extract the additional, independent information that each moment restriction provides. For each orthogonalized moment condition, the efficient instrument is readily available, and the moment conditions are converted to unconditional ones.

In our setting, let $\Sigma(W_i) = \operatorname{Var}(U_i \mid W_i)$ and let $\Gamma(W) = \mathbb{E}[X_i U_i^2 | W]\Sigma(W)^{-1}$ be the projection of $X_i U_i$ on U_i—note that $\Gamma(W)$ takes a familiar $\operatorname{Cov}(\cdot)\operatorname{Var}(\cdot)^{-1}$ form. We obtain the orthogonalized moment conditions by subtracting the projection $\Gamma(W_i)U_i$ from the moment condition $X_i U_i$: $\mathbb{E}[U_i | W_i] = \mathbb{E}[X_i U_i - \Gamma(W_i) U_i] = 0$. By construction, the two moments are orthogonal when conditioned on W_i:

$$\mathbb{E}[U_i \cdot (X_i U_i - \Gamma(W_i) U_i) \mid W_i] = \mathbb{E}[X_i U_i^2 | W_i] - \Gamma(W_i)\mathbb{E}[U_i^2 | W_i] = 0.$$

The orthogonality means that the moment conditions now provide independent information for θ; therefore, we may simply consider optimal instruments for each moment condition and form a combined moment restriction. The optimal instrument for the conditional moment restriction $\mathbb{E}[U_i \mid W_i] = 0$ is $\frac{\mathbb{E}[T_i | W_i]}{\Sigma(W_i)}$ and that for the unconditional moment restriction is the constant scaling matrix $C := \mathbb{E}[T_i(X_i - \Gamma(W_i))^\intercal]\mathbb{E}[U_i^2(X_i - \Gamma(W_i))(X_i - \Gamma(W_i))^\intercal]^{-1}$.

The combined moment condition is then

$$\mathbb{E}[\Upsilon_i(Y_i - T_i^\intercal \theta)] = 0 \qquad \Upsilon_i := \Sigma(W_i)^{-1}\mathbb{E}[T_i \mid W_i] + C(X_i - \Gamma(W_i))$$

whose solution takes the familiar form $\theta = \mathbb{E}\left[\Upsilon_i T_i^\intercal\right]^{-1} \mathbb{E}\left[\Upsilon_i Y_i\right]$. This characterization then motivates a split-sample plug-in procedure as before, though the nuisance parameters are more complex—

as $C, \Sigma, \Gamma, \mathbb{E}[T_i|W_i]$ all require estimation.[5] We discuss the details in Appendix A. Analogously, we may state high-level conditions for consistency and normality of $\hat{\theta}_{\text{MLSS}}$ in Theorem 2.2. Additionally, Lemma D.1 presents some sufficient conditions for Theorem 2.2 as well.

Theorem 2.2. Let $G = \mathbb{E}[\Upsilon_i T_i^\intercal]^{-1}, \Omega = \mathbb{E}[U_i^2 \Upsilon_i \Upsilon_i^\intercal]$. Let $\hat{\theta}_{\text{MLSS}}$ be defined as in (2) in Appendix A. Under Condition 2, which governs the L^2-consistency of $\hat{\Upsilon}$ and concentration of various terms, the estimator is consistent and asymptotically normal $\sqrt{N}(\hat{\theta}_{\text{MLSS}} - \theta) \rightsquigarrow \mathcal{N}(0, G\Omega G^\intercal)$.

2.3 Discussion

We now provide some discussion in light of Theorems 2.1 and 2.2.

Estimation of standard errors. The variance-covariance matrix of $\hat{\theta}_{\text{MLSS}}$ can be readily estimated by plugging in $\hat{\Upsilon}$ and $\hat{\theta}_{\text{MLSS}}$. In principle, under strong identification, inference may be conducted via bootstrap, but machine learning methods tend to be somewhat computationally expensive and bootstrap may not be an attractive option for large scale problems.

"Forbidden regression." Nonlinearities in the first stage are often discouraged due to a "forbidden regression," where the researcher regresses Y on $\hat{D}$ estimated via nonlinear methods, motivated by a heuristic explanation for TSLS. As Angrist and Krueger (2001) point out, this regression is incorrect, and correct inference follows from using $\hat{D}$ as an intrument for D, as we do, rather than replacing D with $\hat{D}$—in the TSLS setting, the two estimates are numerically equivalent, but not in general.

Connection between first-stage fitting and estimate quality. Additionally, we connect the quality of the first-stage fitting to the quality of the final estimation in a heuristic argument here. Consider the homoskedastic, no-covariate case where D is a scalar and, in a slight abuse of notation, let $\Upsilon = \mathbb{E}[D|W]$ be the optimal instrument that excludes the constant. IV estimators of τ, in this case, broadly take the form of $\text{Cov}_n(\boldsymbol{Q}, \boldsymbol{Y}) / \text{Cov}_n(\boldsymbol{Q}, \boldsymbol{D})$ for some constructed instrument $\boldsymbol{Q} = [Q_1, \ldots, Q_N]^\intercal \in \mathbb{R}^N$, where Cov_n are sample covariance: Just-identified linear IV estimator corresponds to $\boldsymbol{Q} = \boldsymbol{W}$ and the (infeasible) efficient estimator corresponds to $\boldsymbol{Q} = \boldsymbol{\Upsilon}$. The estimation error then takes the form of $\text{Cov}_n(\boldsymbol{Q}, \boldsymbol{U}) / \text{Cov}_n(\boldsymbol{Q}, \boldsymbol{D})$. The central limit theorem implies that $\text{Cov}_n(\boldsymbol{Q}, \boldsymbol{U}) = O_p\left(N^{-1/2} \text{Var}_n(\boldsymbol{Q})\sigma^2\right)$. Thus the estimation error is of order

$$\text{Cov}_n(\boldsymbol{Q}, \boldsymbol{U}) = O_p\left(\frac{1}{\sqrt{N}} \frac{\sqrt{\text{Var}_n(\boldsymbol{Q})}\sigma}{\text{Cov}_n(\boldsymbol{Q}, \boldsymbol{D})}\right) = O_p\left(\frac{\sigma}{\sqrt{N}\sqrt{\text{Var}_n(\boldsymbol{D})}} \frac{1}{R}\right) \text{ where } R = \text{Corr}_n(\boldsymbol{Q}, \boldsymbol{D}).$$

The estimation error is thus decreasing in the quality of prediction in the first stage, as measured by R^2, which reflects the intuition that a first-stage prediction with better quality should deliver IV estimates that are more precise.[6] The out-of-sample R^2, which can be readily computed from a split-sample procedure, therefore offers a useful indicator for quality of estimation. In particular, if one is comfortable with the strengthened identification assumptions, there is little reason not to use the model that achieves the best out-of-sample prediction performance on the split-sample. In many settings, this best-performing model would be linear regression, but in many settings it may not be, and using more complex tools may deliver benefits.

Moreover, much of the discussion on using machine learning for instrumental variables analysis has been focused on *selecting* instruments (Belloni *et al.*, 2012; Angrist and Frandsen, 2019) assuming some level of sparsity, motivated by statistical difficulties encountered when the number of instruments is high. In light of the heuristic above, a more precise framing is perhaps *combining* instruments to form a better prediction of the endogenous regressor, as noted by Hansen and Kozbur (2014).

Weak IV robust inference. A major practical motivation for our work, following Bai and Ng (2010), is to use machine learning to rescue otherwise weak instruments due to a lack of linear correlation; nonetheless, the instrument may be irredeemably weak, and providing weak-instrument robust inference is important in practice. Identification-robust inference in this nonlinear framework is formally considered in Antoine and Lavergne (2019)—we provide a simple construction here which is valid but not necessarily optimal. The independence induced by sample-splitting readily

[5]In the case where we assume homoskedasticity or choose to forgo efficiency, we only need to estimate $\mathbb{E}[T_i|W_i]$.

[6]This heuristic, of course, should not be taken literally, since the heuristic assumes that $\mathbb{E}[\text{Cov}(\boldsymbol{Q}, \boldsymbol{U})] = 0$, which can fail if Q estimated from the data and suffers from overfitting.

allows for weak-instrument robust inference, a point that Staiger and Stock (1994) briefly touch upon and is used in recent works such as Mikusheva and Sun (2020). On each subsample S_j, conditional on subsamples S_{-j}, the triplet $(\hat{\Upsilon}_i, T_i, Y_i)$ are independently distributed. Thus, conditioning on samples S_{-j}, weak-instrument robust procedures, such as the Anderson–Rubin test (Anderson *et al.*, 1949) and the Kleibergen (2002) Lagrange multiplier test, continue to have correct coverage. The tests over each subsample S_j can be combined via a simple Bonferroni correction.[7] Moreover, since the weak IV robust procedures take instruments, covariates, endogenous variables, and response variables as input, robust inference does not require possibly expensive re-fitting of the machine learning procedures. Similarly, weak identification robust inference applies out-of-the-box in the more complex setting with covariates below as well.

(When) is machine learning useful? We conclude this section by discussing our work relative to Angrist and Frandsen (2019), who note that using lasso and random forest methods in the first stage does not seem to provide large performance benefits in practice, on a simulation design based on the data of Angrist and Krueger (1991). We note that, per our discussion above in the connection between first-stage fitting and estimate quality, a good heuristic summary for the estimation precision is the R^2 between the fitted instrument and the true optimal instrument—$\mathbb{E}[D|W]$ in the homoskedastic case. It is quite possible that in some settings, the conditional expectation $\mathbb{E}[D|W]$ is estimated well with linear regression, and lasso or random forest do not provide large benefits in terms of out-of-sample prediction quality. Since Angrist and Krueger (1991)'s instruments are quarter-of-birth interactions and are hence binary, it is in fact likely that predicting D with linear regression performs well relative to nonlinear or complex methods in the setting. Whether or not machine learning methods work well relative to linear methods is something that the researcher may verify in practice, via evaluating performance on a hold-out set, which is standard machine learning practice but is not yet widely adopted in empirical economics. Indeed, we observe that in both real (Section 3) and Monte Carlo (Appendix C) settings where the out-of-sample prediction quality of more complex machine learning methods out-perform linear regression, MLSS estimators perform better than TSLS.

3 Empirical Application

We consider an empirical application in the criminal justice setting of Ash *et al.* (2019), where we consider the causal effect of appellate court decisions at the U.S. circuit court level on lengths of criminal sentences at the U.S. district courts under jurisdiction of the circuit court. Ash *et al.* (2019) exploit the fact that appellate judges are randomly assigned, and use the characteristics of appellate judges—including age, party affiliation, education, and career backgrounds—as instrumental variables. In criminal justice cases, plaintiffs rarely appeal, as it would involve trying the defendant twice for the same offense—generally not permitted in the United States; therefore, an appellate court reversal would typically be in favor of defendants, and we may posit a causal channel in which such reversals affect sentencing; for instance, the district court may be more lenient as a result of a reversal, as would be naturally predicted if the reversal sets up a precedent in favor of the defendant.

To connect the set up with our notation, the outcome variable Y is the change in sentencing length before and after an appellate decision, measured in months, where larger values of Y indicates that sentences become longer after the appellate court decision. The endogenous treatment variable D is whether or not an appellate decision reverses a district court ruling. The instrument W is the characteristics of the randomly assigned circuit judge presiding over the appellate case in question, and covariates X contain textual features from the circuit case, represented by Doc2Vec embeddings (Le and Mikolov, 2014).[8]

We present our results in Table 1, comparing an MLSS estimator to a linear IV estimator, with or without sample-splitting. The first three columns of Table 1 display the point estimate, standard errors, and Wald (1.96-s.e.) confidence intervals. The point estimate of MLSS estimators fall in the range -1.7 to -0.6, suggesting a small effect of appellate court reversals—typically in favor of an appealing

[7] More precisely, for each $\tilde{\theta}$ value on a grid, we test $H_0 : \theta = \tilde{\theta}$ by Bonferroni-correcting the Anderson–Rubin or Kleibergen LM tests over the subsamples S_j. A confidence interval is formed by collected $\tilde{\theta}$'s that are not rejected.

[8] Empirically, the instrument does not seem to predict the covariates very well, and so we make the convenient assumption that $\mathbb{E}[X|Z]$ is constant in implementing the estimator with covariates, and the MLSS estimator with covariates under identity weighting is simply the TSLS estimator with the MLSS-estimated instrument $\hat{\Upsilon}$ and covariates.

defendant—on district court leniency. The Wald confidence intervals are of the range -3 to 0.8, which fail to reject the null hypothesis of zero effect. These results do not vary wildly across sample splits and are robust to inclusion of covariates. In the fourth column, we display our preferred inference method, the identification-robust Anderson–Rubin confidence interval. For MLSS estimators, the AR interval mostly agrees with the Wald interval, suggesting that the estimated instrument $\hat{\Upsilon}$ is in fact strong, despite that judge characteristics are a relatively weak predictor for appellate court decisions.

Linear IV-based methods, however, suffer considerably from weak identification. Standard errors are significantly larger for linear IV estimates, and the Wald confidence intervals are dramatically different from the Anderson–Rubin confidence intervals. Indeed, the Wald intervals are often significantly shorter than AR intervals, which can sometimes even be $(-\infty, \infty)$, suggesting that the instruments are uninformative of the treatment. In this case, Wald confidence intervals for linear IV estimators lead to overall imprecise inference and are sometimes quite misleading—Wald intervals for the non-split-sample linear IV estimator are quite spuriously narrow, masking identification issues.

Table 1: Estimates and confidence intervals of the treatment effect of appeals decision reversal on criminal sentence length (months)

Split-sample ID	Estimate	SE	Wald	Anderson-Rubin	R^2	Estimator
0	-1.71	0.70	[-3.08, -0.35]	[-3.04, -0.23]	0.0199	MLSS (Random Forest)
0	5.35	5.17	[-4.78, 15.47]	[$-\infty, \infty$]	-0.0137	Split-sample linear IV
1	-0.65	0.61	[-1.85, 0.55]	[-2.34, 0.80]	0.0261	MLSS (Random Forest)
1	1.14	2.74	[-4.24, 6.52]	[-11.04, 11.58]	-0.0078	Split-sample linear IV
2	-0.91	0.67	[-2.22, 0.40]	[-3.06, 0.98]	0.0250	MLSS (Random Forest)
2	2.31	4.10	[-5.73, 10.35]	[$-\infty, \infty$]	-0.0114	Split-sample linear IV
3	-1.26	0.65	[-2.54, 0.02]	[-2.64, -0.09]	0.0250	MLSS (Random Forest)
3	5.74	5.17	[-4.39, 15.87]	(-∞, -33.49] $\cup$ [-8.54 ,∞)	-0.0132	Split-sample linear IV
4	-1.06	0.65	[-2.34, 0.22]	[-2.96, 0.76]	0.0263	MLSS (Random Forest)
4	-3.24	5.51	[-14.05, 7.56]	[-10.88, 79.99]	-0.1174	Split-sample linear IV
0	-1.76	0.82	[-3.37, -0.16]	[-3.04, -0.33]		MLSS with covariates
1	-0.68	0.75	[-2.14, 0.78]	[-2.48, 0.89]		MLSS with covariates
2	-0.93	0.77	[-2.44, 0.59]	[-2.95, 1.05]		MLSS with covariates
3	-1.31	0.77	[-2.82, 0.21]	[-2.71, 0.01]		MLSS with covariates
4	-1.09	0.76	[-2.57, 0.39]	[-3.12, 0.99]		MLSS with covariates
Full-sample	0.8300	1.0600	[-1.25, 2.91]	[-1.00, 14.06]	—	Linear IV

Notes: Parameter estimates and confidence intervals over five random sample-splits shown. For each split-sample, we further split S_1 to form a validation sample 10% the size for tuning hyperparameters. Hyperparameters of the random forest are chosen to minimize validation error on this subsample of S_1. All Anderson–Rubin intervals for split-sample designs are applied to the estimated instrument, and thus are just-identified.

4 Conclusion

In this paper, we provide a simple and user-friendly analysis of incorporating machine learning techniques into instrumental variable analysis in a manner that is familiar to applied researchers. In particular, we document via elementary techniques that a split-sample IV estimator with machine learning methods as the first stage inherits classical asymptotic and optimality properties of usual instrumental regression, requiring only weak conditions governing the consistency of the first stage prediction. In the presence of covariates, we also formalize moment conditions for instrumental regression that continues to leverage nonlinearities in the excluded instrument without creating spurious identification from the nonlinearities in the included covariates. Leveraging such nonlinearity in the first stage allows the user to extract more identifying variation from the instrumental variables and can have the potential of rescuing seemingly weak instruments into strong ones, as we demonstrate with simulated data and real data from a criminal justice context. Conventional accoutrements to an instrumental variable analysis, such as identification-robust confidence sets, extend seamlessly in the presence of machine learning first stage. We believe that machine learning in IV settings is a mostly harmless addition to the empiricist's toolbox.

References

AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71** (6), 1795–1843.

— and — (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, **141** (1), 5–43.

— and — (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, **170** (2), 442–457.

ANDERSON, T. W., RUBIN, H. *et al.* (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, **20** (1), 46–63.

ANDREWS, I., STOCK, J. H. and SUN, L. (2019). Weak instruments in iv regression: Theory and practice. In *Annual Review of Economics*.

ANGRIST, J. and FRANDSEN, B. (2019). *Machine labor*. Tech. rep., National Bureau of Economic Research.

ANGRIST, J. D., IMBENS, G. W. and KRUEGER, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, **14** (1), 57–67.

— and KRUEGER, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106** (4), 979–1014.

— and — (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, **15** (4), 69–85.

— and PISCHKE, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

ANTOINE, B. and LAVERGNE, P. (2019). Identification-robust nonparametric inference in a linear iv model.

ASH, E., CHEN, D., ZHANG, X., HUANG, Z. and WANG, R. (2019). Deep iv in law: Analysis of appellate impacts on sentencing using high-dimensional instrumental variables. In *Advances in Neural Information Processing Systems (Causal ML Workshop)*.

BAI, J. and NG, S. (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory*, pp. 1577–1606.

BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80** (6), 2369–2429.

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

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

CHEN, J., CHEN, X., LIAO, Y. and TAMER, E. (2020a). *Deep Learning Inference on Semi-Parametric Models with Weakly Dependent Data*. Tech. rep.

—, — and TAMER, E. (2020b). *Efficient Estimation of Expectation Functionals of Nonparametric Conditional Moments via Neural Networks*. Tech. rep.

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.

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.

DIETERLE, S. G. and SNELL, A. (2016). A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument. *Labour Economics*, **42**, 76–86.

DIKKALA, N., LEWIS, G., MACKEY, L. and SYRGKANIS, V. (2020). Minimax estimation of conditional moment models. *arXiv preprint arXiv:2006.07201*.

ESCANCIANO, J. C. and LI, W. (2020). Optimal linear instrumental variables approximations. *Journal of Econometrics*.

HANSEN, C. and KOZBUR, D. (2014). Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, **182** (2), 290–308.

HARTFORD, J., LEWIS, G., LEYTON-BROWN, K. and TADDY, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pp. 1414–1423.

HOROWITZ, J. L. and LEE, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, **75** (4), 1191–1208.

KLEIBERGEN, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, **70** (5), 1781–1803.

LE, Q. and MIKOLOV, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196.

LIU, R., SHANG, Z. and CHENG, G. (2020). On deep instrumental variables estimate.

MACKEY, L., SYRGKANIS, V. and ZADIK, I. (2018). Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, PMLR, pp. 3375–3383.

MIKUSHEVA, A. and SUN, L. (2020). Inference with many weak instruments. *arXiv preprint arXiv:2004.12445*.

NEWEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, **71** (5), 1565–1578.

SEVERINI, T. A. and TRIPATHI, G. (2012). Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, **170** (2), 491–498.

STAIGER, D. and STOCK, J. H. (1994). *Instrumental variables regression with weak instruments*. Tech. rep., National Bureau of Economic Research.

A Definition of the MLSS estimator in the presence of covariates

As before, we split the sample randomly in half and let S_1, S_2 denote the two halves. For efficiency purposes, we would need to estimate the heteroskedasticity $\Sigma(U_i)$, which is unnecessary if we are not concerned with efficiency or we are willing to assume homoskedasticity. We refer to estimating $\Sigma(U_i)$ as *efficiency weighting*, and we refer to simply plugging in $\Sigma(U_i) = 1$ as *identity weighting*. Since the heteroskedasticity $\Sigma(U_i)$ depends on θ, we need preliminary estimates $\hat{\theta}$, which we may obtain via TSLS or identity weighting. Let $\hat{\theta}_j$ be a preliminary, consistent estimate for θ using data only from S_{-j}. Let $\hat{f}_j$ be an estimate for $\mathbb{E}[D_i|W_i]$ trained using data on S_{-j} and, similarly, let $\hat{g}_j$ be an estimate for $\mathbb{E}[R_i|W_i]$, trained on $i \in S_{-j}$, where

$$R_i = \begin{cases} X_i & \text{under identity weighting} \\ X_i(Y_i - T_i^\intercal \hat{\theta}_j)^2 & \text{under efficient weighting.} \end{cases}$$

Under efficiency weighting, let $\hat{\Sigma}_j(W_i)$ be an estimate of the heteroskedasticity, approximated by $\mathbb{E}[(Y_i - T_i^\intercal \hat{\theta}_j)^2 \mid W_i]$ where $i \in S_{-j}$, and let $\hat{\Sigma}_j(W_i) = 1$ under identity weighting. Lastly, let $\hat{U}_i = Y_i - T_i^\intercal \hat{\theta}_j$ and let

$$\hat{C}_j = \left[\frac{1}{n}\sum_{i\in S_j} T_i \left(X_i - \frac{\hat{g}_j(W_i)}{\hat{\Sigma}_j(W_i)}\right)^\intercal\right]\left[\frac{1}{n}\sum_{i\in S_j}\left(X_i - \frac{\hat{g}_j(W_i)}{\hat{\Sigma}_j(W_i)}\right)\left(X_i - \frac{\hat{g}_j(W_i)}{\hat{\Sigma}_j(W_i)}\right)^\intercal \hat{U}_i^2\right]^{-1}$$

be a plug-in estimate of C. Note that under homoskedasticity $\mathbb{E}[U^2|X,W] = \sigma^2$, we have a much simpler form

$$\hat{C}_j = \left[\frac{1}{n}\sum_{i\in S_j} T_i \left(X_i - \hat{g}_j(W_i)\right)^\intercal\right]\left[\frac{1}{n}\sum_{i\in S_j}\left(X_i - g_j(W_i)\right)\left(X_i - \hat{g}_j(W_i)\right)^\intercal\right]^{-1} \quad \hat{g} = \hat{\mathbb{E}}[X_i|W_i].$$

Finally, for $i \in S_j$, we may form an estimate of the optimal instrument and the MLSS estimator

$$\hat{\Upsilon}_i := \frac{\hat{f}_j(W_i)}{\hat{\Sigma}_j(W_i)} + \hat{C}_j \cdot \left(X_i - \frac{\hat{g}_j(W_i)}{\hat{\Sigma}_j(W_i)}\right) \quad \hat{\theta}_{\text{MLSS}} := \left(\frac{1}{N}\sum_{i=1}^N \hat{\Upsilon}_i T_i\right)^{-1} \frac{1}{N}\sum_{i=1}^N \hat{\Upsilon}_i Y_i \tag{2}$$

B Discussion related to NPIV

A principled modeling approach is the NPIV model, which treats the unknown structural function g as an infinite dimensional parameter and considers the model

$$\mathbb{E}[Y - g(T)|Z] = 0. \tag{NPIV}$$

The researcher may be interested in g itself, or some functionals of g, such as the average derivative $\theta = \mathbb{E}\left[\frac{\partial g}{\partial D}(T) \mid Z\right]$ or the best linear approximation $\beta = \mathbb{E}[TT^\intercal]^{-1}\mathbb{E}[Tg(T)]$. One might wonder whether choosing a parametric functional form in place of $g(T)$ is without loss of generality. Linear regression of Y on T, for instance, yields the best linear approximation to the structural function $\mathbb{E}[Y|T]$, and thus has an attractive nonparametric interpretation; it may be tempting to ask whether an analogous property holds for IV regression. If an analogous property does hold, we may have license in being more blasé about linearity in the second stage.

Unfortunately, modeling g as linear does not produce the best linear approximation, at least not with respect to the L^2-norm. Escanciano and Li (2020) show that the best linear approximation can be written as a particular IV regression estimand

$$\beta = \mathbb{E}[h(Z)T^\intercal]^{-1}\mathbb{E}[h(Z)Y]$$

where h has the property that $\mathbb{E}[h(Z)|T] = T$. Note that with efficient instrument in a homoskedastic, no-covariate linear IV context as we consider in Section 2.1, the optimal instrument is $\hat{D}(W) = \mathbb{E}[D|W]$. A sufficient condition, under which the IV estimand based on the optimal instrument is equal to the best linear approximation β, is the somewhat strange condition that the projection onto D of *predicted* D is linear in D itself: For some invertible A, $\mathbb{E}[\hat{D}(W)|D] = AD$. The condition would hold, for instance, in a setting where D, W are jointly Gaussian and all conditional expectations are linear, but it is difficult to think it holds in general. As such, linear IV would not recover the best linear approximation to the nonlinear structural function in general.

A simple calculation can nevertheless characterize the bias of the linear approach if we take the estimand to be the best linear approximation to the structural function. Suppose we form an instrumental variable estimator that converges to an estimand of the form

$$\gamma = \mathbb{E}[f(Z)X^\intercal]^{-1}\mathbb{E}[f(Z)Y].$$

It is easy to see that

$$\gamma - \beta = \langle g - \mathbb{E}^*[g|T], \mu - \mathbb{E}^*[\mu|T] \rangle,$$

where $\langle A, B \rangle = \mathbb{E}[AB]$, $\mu(T) = \mathbb{E}[f(Z)|T]$, and $\mathbb{E}^*[A|B]$ is the best linear projection of A onto B. This means that the two estimands are identical if and only if at least one of $\mu(\cdot)$ or $g(\cdot)$ are linear, and all else equal the bias is smaller if μ or g is more linear. Importantly, $\mu - \mathbb{E}^*[\mu|T]$ are objects that we could empirically estimate since they are conditional means, and in practice the researcher may estimate $\mu - \mathbb{E}^*[\mu|T]$, which delivers bounds on $\gamma - \beta$ through assumptions on linearity of g.

C Monte Carlo example

We report in Table 2 the results of a Monte Carlo experiment corresponding to the simple case without covariates. We let $N = 500$, $W_i \sim \mathcal{N}(0, I_p)$ for $p = 5$, and let $D = f(W) + V$ where $f(\cdot)$ is linear in W and element-wise squares and cubes of W. We let $Y = D\tau + U$ for $U = \rho V + \sqrt{1-\rho^2}S$, where the endogeneity is controlled by $\rho = 0.8$ and $S, V \sim \mathcal{N}(0, 1)$ independently.

We consider the performance of the traditional TSLS estimator, an MLSS estimator (where nuisance estimation is performed via two-layer ReLU feedforward network), and an oracle estimator (the efficient instrument $\Upsilon(W) = \mathbb{E}[D|W]$ is known) in two settings. In one setting ("quadratic only"), f is symmetric about zero (only containing quadratic terms in W), and thus W is a weak instrument from the linear IV perspective. In the other setting, W is a strong instrument for linear IV as well.

Unsurprisingly, MLSS performs significantly better in terms of estimation quality compared to linear TSLS when the instrument is linearly irrelevant. In the other setting, when the prediction performance of MLSS and TSLS are similar, we see that the performance in terms of coverage and RMSE is similar as well. However, this only indicates the performance of a specific machine learning estimator with minimal tuning; the first stage R^2 of the oracle estimator indicates that neither the linear regression nor the prediction function chosen by the MLSS estimator is close to the true conditional expectation function, and we should expect better machine learning methods, in terms of out-of-sample prediction quality, being able to do better in terms of RMSE of the structural parameter.

We see that Wald intervals for all there estimators cover at approximately the nominal level.[9] The oracle estimator seems to confirm that we should expect mild coverage distortions in finite samples, relative to the 95% nominal level, as the oracle estimator also undercover by 2–3%.

Table 2: Monte Carlo experiment results

		RMSE	Coverage	R^2
Quadratic only				
True	MLSS	0.0192	0.9150	0.7555
	TSLS	0.1422	0.9700	-0.1625
	Oracle	0.0168	0.9250	0.8897
False	MLSS	0.0041	0.9400	0.4777
	TSLS	0.0038	0.9400	0.4732
	Oracle	0.0030	0.9200	0.9959

Notes: Summary statistics across 200 experiments shown. *RMSE* refers to the root mean-square error of the parameter of interest. *Coverage* refers to the coverage of a nominal 95% Wald interval. R^2 refers to the statistic $R^2 \equiv 1 - \frac{\text{Mean Squared Prediction Error}}{\text{Var } Y}$ calculated out-of-sample, which can be negative. *Quadratic only* refers to whether $\mathbb{E}[D|W]$ has only quadratic terms in W—if so TSLS would be weak since the best linear approximation to $\mathbb{E}[D|W]$ when W is symmetric about zero is a constant function.

D Omitted Proofs and Regularity Conditions

Condition 1. 1. (Strong identification) For $j = \{1, 2\}$,

$$\frac{1}{N} \sum_i \hat{\Upsilon}(W_i) T_i^\intercal \xrightarrow{p} G^{-1}$$

for some nonsingular G^{-1}.

[9]This is a fortunate fact for the TSLS estimator in the quadratic only scenario, where the coverage happens to be above nominal despite the nonidentification.

2. (Lyapunov condition) $\mathbb{E}|U_i|^{2+\epsilon} < \infty$ for some $\epsilon > 0$. Furthermore, none of the weights $\hat{\Upsilon}_j$ dominates the others

$$\frac{\max_{i \in S_j} \|\hat{\Upsilon}_j(W_i)\|^2}{\sum_{i \in S_j} \|\hat{\Upsilon}_j(W_i)\|^2} \xrightarrow{p} 0.$$

3. $\frac{1}{n}\sum_{i \in S_j} \hat{\Upsilon}_j(W_i)\hat{\Upsilon}_j^{\intercal}(W_i) \xrightarrow{p} \Omega$ for some Ω, and there exists Z_1, Z_2 such that $Z_j = \frac{1}{\sqrt{n}}\sum_{i \in S_j} \hat{\Upsilon}_j(W_i)U_i + o_p(1)$ and $\text{Cov}(Z_1, Z_2) \to 0$.

Proof of Theorem 2.1. Observe that, since $Y_i = T_i^{\intercal}\theta + U_i$,

$$\sqrt{N}\,(\hat{\theta}_{\text{MLSS}} - \theta) = \left(\frac{1}{N}\sum_{i=1}^{N} \hat{\Upsilon}_i T_i^{\intercal}\right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \hat{\Upsilon}_i U_i.$$

By conditions 2 and 3, $\tilde{Z}_j := \frac{1}{\sqrt{n}}\sum_{i \in S_j} \hat{\Upsilon}_j(W_i)U_i \rightsquigarrow \mathcal{N}(0, \sigma^2\Omega)$. jointly for $j = 1, 2$. The asymptotic covariance is zero by condition 3. Thus $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \hat{\Upsilon}(W_i)U_i = \frac{\tilde{Z}_1 + \tilde{Z}_2}{\sqrt{2}} \rightsquigarrow \mathcal{N}(0, \sigma^2\Omega)$. The first part of the claim then follows from condition 1.

The second part of the claim follows if we show

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\Upsilon}_i T_i^{\intercal} = \frac{1}{N}\sum_{i=1}^{N} \Upsilon_i T_i^{\intercal} + o_p(1) \qquad \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \hat{\Upsilon}_i U_i = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \Upsilon_i U_i + o_p(1)$$

This argument is carried out in the proof to Theorem 2.2, where the condition needed is exactly the L^2 consistency condition. □

Condition 2. 1. $\hat{C}_j \xrightarrow{p} C$ and certain empirical moments are bounded:

$$\frac{1}{n}\sum_{i \in S_j} \left\| X_i - \frac{\hat{g}_j(W_i)}{\hat{\Sigma}_j(W_i)} \right\|^2 = O_p(1) \quad \mathbb{E}U_i^2 < \infty \quad \mathbb{E}\|T_i\|^2 < \infty$$

2. Let $\tilde{\Upsilon}_i$ be $\hat{\Upsilon}_i$ but with C replacing $\hat{C}_j$. The constructed instrument $\tilde{\Upsilon}_i$ is consistent in L^2 to its population counterpart: $\forall j \in \{1, 2\} : \mathbb{E}_{i \in S_j}\left[(\tilde{\Upsilon}_i - \Upsilon_i)^2 \mid S_{-j}\right] \xrightarrow{p} 0$,
3. (Strong identification) $\mathbb{E}[\Upsilon_i T_i] < \infty$ and $\mathbb{E}[\Upsilon_i T_i]$ is not singular, and
4. (Lyapunov CLT) For some $\epsilon > 0$, $\mathbb{E}[|\Upsilon_i U_i|^{2+\epsilon}] < \infty$.

Proof of Theorem 2.2. Observe that, since

$$Y_i = T_i^{\intercal}\theta + U_i, \quad \sqrt{N}\,(\hat{\theta}_{\text{MLSS}} - \theta) = \left(\frac{1}{N}\sum_{i=1}^{N} \hat{\Upsilon}_i T_i^{\intercal}\right)^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \hat{\Upsilon}_i U_i.$$

We proceed to analyze the right-hand side with elementary techniques. Note that if we show

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\Upsilon}_i T_i^{\intercal} = \frac{1}{N}\sum_{i=1}^{N} \Upsilon_i T_i^{\intercal} + o_p(1) \tag{3}$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \hat{\Upsilon}_i U_i = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \Upsilon_i U_i + o_p(1), \tag{4}$$

then, $\sqrt{N}(\hat{\theta}_{\text{MLSS}} - \theta) = \left(\frac{1}{N}\sum_{i=1}^{N} \Upsilon_i T_i^{\intercal}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \Upsilon_i U_i + o_p(1)$ under strong identification (Condition 3). By a law of large numbers (Condition 3), $\left(\frac{1}{N}\sum_{i=1}^{N} \Upsilon_i T_i^{\intercal}\right)^{-1} \xrightarrow{p} G$. By a central limit theorem (Condition 4) $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \Upsilon_i U_i \rightsquigarrow \mathcal{N}(0, \Omega)$, and thus we have the desired asymptotic normality $\sqrt{N}(\hat{\theta}_{\text{MLSS}} - \theta) \rightsquigarrow \mathcal{N}(0, G\Omega G^{\intercal})$.

Note that it suffices to check (3) and (4) for each subsample, and by symmetry it suffices to check them for S_1:

$$\frac{1}{n}\sum_{i \in S_1} \hat{\Upsilon}_i T_i^{\intercal} = \frac{1}{n}\sum_{i \in S_1} \Upsilon_i T_i^{\intercal} + o_p(1) \qquad \frac{1}{\sqrt{n}}\sum_{i \in S_1} \hat{\Upsilon}_i U_i = \frac{1}{\sqrt{n}}\sum_{i \in S_1} \Upsilon_i U_i + o_p(1).$$

We first check that we may replace the $\hat{\Upsilon}_i$ in the left-hand side with $\tilde{\Upsilon}_i$ (defined in Condition 2) by checking that

$$\left\|\frac{1}{n}(\hat{C}_j - C)\sum_{i \in S_j}\left(X_i - \frac{\hat{g}_j(W_i)}{\hat{\Sigma}_j(W_i)}\right)T_i^\intercal\right\|_F = o_p(1) \quad \left\|\frac{1}{n}(\hat{C}_j - C)\sum_{i \in S_j}\left(X_i - \frac{\hat{g}_j(W_i)}{\hat{\Sigma}_j(W_i)}\right)U_i\right\| = o_p(1),$$

which immediately follows from Condition 1.

Next, we first relate both quantities to $\mathbb{E}_2\left\|\Upsilon_i - \tilde{\Upsilon}_i\right\|^2$, where the expectation operator $\mathbb{E}_2$ is taken conditional on S_2, and so $\hat{f}, \hat{g}$ may be viewed as fixed: [10]

$$\begin{aligned}
\mathbb{E}_2\left\|\frac{1}{n}\sum_{i \in S_1}(\tilde{\Upsilon}_i - \Upsilon_i)T_i^\intercal\right\|_F &\le \mathbb{E}_2\left[\frac{1}{n}\sum_{i \in S_1}\left\|\tilde{\Upsilon}_i - \Upsilon_i\right\|\|T_i^\intercal\|\right] && (\|AB\|_F \le \|A\|_F\|B\|_F) \\
&\le \sqrt{\mathbb{E}_2\left[\left\|\Upsilon_i - \tilde{\Upsilon}_i\right\|^2\right]\mathbb{E}_2\|T_i\|^2} && \text{(Cauchy–Schwarz)} \\
&= O\left(\sqrt{\mathbb{E}_2\left[\left\|\Upsilon_i - \tilde{\Upsilon}_i\right\|^2\right]}\right) && (\mathbb{E}\|T_i\|^2 < \infty)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_2\left\|\frac{1}{\sqrt{n}}\sum_{i \in S_1}(\tilde{\Upsilon}_i - \Upsilon_i)U_i\right\|^2 &= \operatorname{tr}\left\{\operatorname{Var}_2\left(\frac{1}{\sqrt{n}}\sum_{i \in S_1}(\tilde{\Upsilon}_i - \Upsilon_i)U_i\right)\right\} \\
&= \operatorname{tr}\left\{\operatorname{Var}_2\left((\tilde{\Upsilon}_i - \Upsilon_i)U_i\right)\right\} \\
&= \operatorname{tr}\left\{\mathbb{E}_2\left[(\tilde{\Upsilon}_i - \Upsilon_i)(\tilde{\Upsilon}_i - \Upsilon_i)^\intercal\right]\mathbb{E}_2[U_i^2]\right\} && \text{(Independence)} \\
&= O\left(\mathbb{E}_2\left[\left\|\tilde{\Upsilon}_i - \Upsilon_i\right\|^2\right]\right),
\end{aligned}$$

where $\|A\|_F$ denotes A's Frobenius norm, and $\|v\|$ is the ℓ_2-norm of v. L_2 consistency (Condition 2) then guarantees that both quantities are $o_p(1)$. □

Lemma D.1. The following is a sufficient condition for L^2 consistency of $\tilde{\Upsilon}_i$ and $\hat{C}_j \xrightarrow{p} C$ conditions in Condition 2:

1. $C < \infty$
2. There exists $M, \eta > 0$ such that f, g is bounded above, Σ is bounded below, and $\hat{\Sigma}$ is bounded below almost surely: $\|f\|_\infty \le M$, $\|g\|_\infty \le M$, $\inf_w \mathbb{E}[U^2 | W = w] > \eta$, and $\Pr(\hat{\Sigma}(W) \ge \eta) = 1$.
3. L^2-consistency of the individual components: $\mathbb{E}_{-j}\left\|\hat{f}_j - f\right\|^2 = o_p(1)$, $\mathbb{E}_{-j}\left\|\hat{g}_j - g\right\|^2 = o_p(1)$, and $\mathbb{E}_{-j}\left|\hat{\Sigma}_j - \Sigma\right|^2 = o_p(1)$.
4. $\mathbb{E}\|T_i\|^3 < \infty, \mathbb{E}U_i^4 < \infty, \mathbb{E}\left[U_i^4\|T_i\|^2\right] < \infty$.

Proof. **Checking 2 in Condition 2.** Note that

$$\mathbb{E}_2\left[\|\Upsilon_i - \tilde{\Upsilon}_i\|^2\right] = \mathbb{E}_2\left[\|P + Q\|^2\right] \le 2\mathbb{E}_2\left[\|P\|^2 + \|Q\|^2\right]$$

where

$$P = \frac{\hat{f}_1(W_i)}{\hat{\Sigma}_1(W_i)} - \frac{f(W_i)}{\Sigma(W_i)} = \frac{\hat{f}_1}{\hat{\Sigma}_1\Sigma}\left(\Sigma - \hat{\Sigma}_1\right) + \frac{1}{\Sigma}(\hat{f}_1 - f)$$

$$Q = C\left[\frac{\hat{g}_1(W_i)}{\hat{\Sigma}_1(W_i)} - \frac{g(W_i)}{\Sigma(W_i)}\right]$$

[10]Precisely speaking, $\mathbb{E}_2[X_i] = \mathbb{E}[X_i \mid S_2]$ where $i \in S_1$, and similarly for Var_2

We may control $\mathbb{E}\|P\|^2$:

$$\begin{aligned}\frac{1}{2}\mathbb{E}_2[\|P\|^2] &\le \mathbb{E}_2\left[\frac{\|\hat{f}_1\|}{\hat{\Sigma}\Sigma}\left(\Sigma-\hat{\Sigma}\right)\right]^2 + \mathbb{E}_2\left[\frac{1}{\Sigma}\left\|\hat{f}_1 - f\right\|\right]^2 \\ &\le M\eta^{-2}\mathbb{E}_2[(\Sigma-\hat{\Sigma})^2] + \eta^{-2}\mathbb{E}_2\|\hat{f}_1 - f\|^2 \\ &= o_p(1)\end{aligned}$$

Similarly, $\mathbb{E}[\|Q\|^2] = o_p(1)$.

Checking $\hat{C}_j \xrightarrow{p} C$. By continuous mapping theorem, it suffices to check that both multiplicands in $\hat{C}_j$ converge in probability to the analogous multiplicand in C. Note that

$$\mathbb{E}_2\left\|\frac{1}{n}\sum_{i\in S_1}\left(\frac{\hat{g}_j}{\hat{\Sigma}_j} - \frac{g_j}{\Sigma_j}\right)T_i^\intercal\right\|_F \le \left(\mathbb{E}_2\|T_i\|^2 \mathbb{E}_2 \underbrace{\left\|\frac{\hat{g}_j}{\hat{\Sigma}_j} - \frac{g_j}{\Sigma_j}\right\|^2}_{Q}\right)^{1/2} = o_p(1)$$

and thus

$$\frac{1}{n}\sum_{i\in S_1} T_i\left(\frac{\hat{g}_j}{\hat{\Sigma}_j} - \frac{g_j}{\Sigma_j}\right) = o_p(1). \qquad \text{(Cauchy–Schwarz)}$$

The second term is analogously bounded

$$\left\|\frac{1}{n}\sum_{i\in S_1}\hat{U}_i^2\hat{A}_i\hat{A}_i^\intercal - U_i^2 A_i A_i^\intercal\right\|_F \le \frac{1}{n}\sum_{i\in S_1}\left|\hat{U}_i^2 - U_i^2\right| \|\hat{A}_i\hat{A}_i^\intercal\|_F + \frac{1}{n}\sum_{i\in S_1}U_i^2(\|\hat{A}_i\| + \|A_i\|)\|A_i - \hat{A}_i\|$$

where $A_i = X_i - \frac{g(W_i)}{\Sigma(W_i)}$. Via Cauchy–Schwarz and U_i having four moments, taking $\mathbb{E}_2$ of the second term shows that the second term is

$$O_p(\mathbb{E}_2\|\hat{A}_i - A_i\|) = O_p\left(\mathbb{E}\left[\frac{\hat{g}_1}{\hat{\Sigma}_1} - \frac{g_1}{\Sigma_1}\right]^2\right) = o_p(1).$$

The first term can be written as

$$\begin{aligned}\frac{1}{n}\sum_{i\in S_1}\left|\hat{U}_i^2 - U_i^2\right| \|\hat{A}_i\hat{A}_i^\intercal\|_F &\le \|\hat{\theta}_j - \theta\| \cdot \frac{1}{n}\sum_{i\in S_1}\|T_i\||2Y_i - T_i^\intercal(\theta + \hat{\theta}_j)|\|\hat{A}_i\hat{A}_i^\intercal\|_F \\ &= o_p(1)\left(\frac{1}{n}\sum_{i\in S_1}\|T_i\|\left(2|U_i| + \|T_i\|o_p(1)\right)\left(\|A_iA_i^\intercal\|_F + \|A_i\|o_p(1) + o_p(1)\right)\right) \\ &= o_p(1)\end{aligned}$$

by the last condition. □