Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

March 5, 2025

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

DML for IV and Proxy Controls Models and Robust DML Inference under Weak Identification

13

"Better LATE than nothing."

– Guido Imbens [1].

Here, we specialize DML methods to partially linear models with instruments, arising either through endogeneity of the policy variable or through the use of proxy controls as outlined in Chapter 12. We also present DML methods for LATE parameters in the fully nonlinear model with a binary endogenous treatment and binary instrument. We further examine how DML inference method can be modified to cope with weak instruments and weak identification in generic moment problems through the use of Neyman orthogonal scores and Neyman's $C(\alpha)$ statistic.

13.1 DML Inference in Partially Linear IV Models

Here we consider estimation of parameters that obey the following instrumental variable exclusion restriction:

$$\mathrm{E}[\epsilon \tilde{Z}] = 0,$$

where

$$\epsilon := \tilde{Y} - \theta_0' \tilde{D},$$

and

$$\begin{aligned} \tilde{Y} &= Y - \ell_0(X), \quad \ell_0(X) = \mathrm{E}[Y \mid X], \\ \tilde{D} &= D - r_0(X), \quad r_0(X) = \mathrm{E}[D \mid X], \\ \tilde{Z} &= Z - m_0(X), \quad m_0(X) = \mathrm{E}[Z \mid X]. \end{aligned}$$

Here we take the dimension of $\tilde{Z}$ to be the same as that of $\tilde{D}$ for simplicity.

Two key examples leading to this statistical structure are

- the partially linear instrumental variable model and
- the partially linear model with proxy controls.

We discussed these examples and showed they fit into this structure in Chapter 12.

To estimate and perform inference on θ_0, we can apply the general DML algorithm with the score

$$\psi(W; \theta, \eta) := (Y - \ell(X) - \theta'(D - r(X)))(Z - m(X)), \quad (13.1.1)$$

where $W = (Y, D, X, Z)$ and $\eta = (\ell, m, r)$ with ℓ, m, and r being P-square-integrable functions mapping the support of X to $\mathbb{R}$. Under the exclusion restriction and using the definition of the nuisance functions, we have that

$$\mathrm{E}[\psi(W; \theta_0, \eta_0)] = 0.$$

It is not difficult to check that the Neyman orthogonality condition,

$$\partial_\eta \mathrm{E}[\psi(W; \theta_0, \eta_0)] = 0,$$

holds at the true value $\eta_0 = (\ell_0, m_0, r_0)$ of the nuisance parameters.

Verify Neyman-orthogonality for yourself as an exercise.

We now explicitly restate the DML algorithm specialized to this case of partially linear IV models.

DML for Partially Linear IV and Proxy Models

1. Partition data indices into k folds of approximately equal size: $\{1, ..., n\} = \cup_{k=1}^{K} I_k$. For each fold $k = 1, ..., K$, compute ML estimators $\hat{\ell}_{[k]}(X)$, $\hat{m}_{[k]}(X)$, $\hat{r}_{[k]}(X)$ of the best predictors $\ell_0(X)$, $m_0(X)$, $r_0(X)$, leaving out the k-th block of data, and obtain the cross-fitted residuals for each $i \in I_k$:

$$\check{Y}_i = Y_i - \hat{\ell}_{[k]}(X_i),$$
$$\check{D}_i = D_i - \hat{r}_{[k]}(X_i),$$
$$\check{Z}_i = Z_i - \hat{m}_{[k]}(X_i).$$

2. Compute the standard IV regression of $\check{Y}_i$ on $\check{D}_i$ using $\check{Z}_i$ as the instrument. That is, obtain $\hat{\theta}$ as the root in θ of the following equation:

$$\mathbb{E}_n[(\check{Y} - \theta'\check{D})\check{Z}] = 0.$$

3. Construct standard errors and confidence intervals as in the standard linear instrumental variables regression theory.

In what follows it will be convenient to use the following notation

$$\|h\|_{L^2} := \sqrt{\mathrm{E}_X[h^2(X)]},$$

where, as before, E_X computes the expectation over values of X.

Theorem 13.1.1 (Adaptive Inference in the Partially Linear IV Model) *Impose technical regularity conditions as in [2] which include the following key conditions: (1) the instruments are strong – namely, the singular values of* $\mathrm{E}[\tilde{D}\tilde{Z}]$ *are well-separated from zero – and (2) the estimators* $\hat{\ell}_{[k]}(X)$, $\hat{m}_{[k]}(X)$, *and* $\hat{r}_{[k]}(X)$ *provide high-quality approximations to the best predictors* $\ell_0(X)$, $m_0(X)$, *and* $r_0(X)$ *– namely,*

$$n^{1/4}\|\hat{\ell}_{[k]} - \ell_0\|_{L^2} \approx 0, \quad n^{1/4}\|\hat{m}_{[k]} - m_0\|_{L^2} \approx 0,$$

and

$$n^{1/4}\|\hat{r}_{[k]} - r_0\|_{L^2} \approx 0.$$

Then the estimation error in $\check{D}_i$ *and* $\check{Y}_i$ *has no first order effect on*

the behavior of $\hat{\theta}$:

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx (\mathbb{E}_n[\tilde{D}\tilde{Z}])^{-1}\sqrt{n}\mathbb{E}_n[\tilde{Z}\epsilon],$$

As a result, $\hat{\theta}$ concentrates in a $1/\sqrt{n}$ neighborhood of θ with deviations approximated by the Gaussian law:

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{a}{\sim} N(0, \mathsf{V}),$$

where

$$\mathsf{V} = (\mathrm{E}[\tilde{D}\tilde{Z}'])^{-1}\mathrm{E}[\tilde{Z}\tilde{Z}'\epsilon^2](\mathrm{E}[\tilde{Z}\tilde{D}])^{-1}.$$

The standard error of $\hat{\theta}$ is estimated as $\sqrt{\hat{\mathsf{V}}/n}$, where $\hat{\mathsf{V}}$ is an estimator of V based on the plug-in principle. The result implies that the confidence interval

$$[\hat{\theta} - 2\sqrt{\hat{\mathsf{V}}/n}, \hat{\theta} + 2\sqrt{\hat{\mathsf{V}}/n}]$$

covers θ for approximately 95% of the realizations of the sample. In other words, if our sample is not atypical, the interval covers the truth.

The Effect of Institutions on Economic Growth

The Notebooks 13.5.2 implement the AJR example.

To demonstrate DML estimation of partially linear structural equation models with instrumental variables, we consider estimating the effect of institutions on aggregate output following the work of Acemoglu, Johnson, and Robinson (2001) [3] (AJR).

We use the same set of 64 country-level observations as AJR. The outcome variable, Y, is the logarithm of GDP per capita and the endogenous explanatory variable, D, is an index which measures protection against expropriation risk that is used as a proxy for the strength of institutions. To deal with endogeneity, we use an instrumental variable Z, which is mortality rates for early European settlers. Our raw set of control variables, X, include distance from the equator and dummy variables for Africa, Asia, North America, and South America.

Estimating the effect of institutions on output is complicated by the clear potential for simultaneity between institutions and output: Better institutions may generate higher incomes, but higher incomes may also lead to the development of better institutions. To help overcome this simultaneity, AJR use mortality rates for early European settlers as an instrument for institution quality. The validity of this instrument hinges on the argument that settlers set up better institutions in places

where they were more likely to establish long-term settlements, that where they were likely to settle for the long term is related to settler mortality at the time of initial colonization, and that institutions are highly persistent. The exclusion restriction for the instrumental variable is then motivated by the argument that GDP, while persistent, is unlikely to be strongly influenced by mortality in the previous century, or earlier, except through institutions.

In their paper, AJR note that their instrumental variable strategy will be invalidated if other factors are also highly persistent and related to the development of institutions within a country and to the country's GDP. A leading candidate for such a factor, as they discuss, is geography. AJR address this by assuming that the confounding effect of geography is adequately captured by a linear term in distance from the equator and a set of continent dummy variables. Using DML allows us to relax this assumption and replace it by a weaker assumption that geography can be sufficiently controlled by an unknown function of distance from the equator and continent dummies which can be learned by ML methods.

We present the verbal identification argument above in the form of a DAG in Figure 13.1. In the DAG, Y is wealth, O the quality of early institutions, D the quality of modern institutions, X observed measures of geography, Z early settler mortality, A the present day latent factors jointly determining modern institutions and wealth, and L early latent factors affecting early settler mortality. Applying the IV method here requires the identification of the causal effect of $Z \to D$ and $Z \to Y$. From the DAG, we see that X blocks the backdoor paths from Y to Z and from $D \to Z$. This means that the instrument satisfies the required exogeneity condition conditional on X.

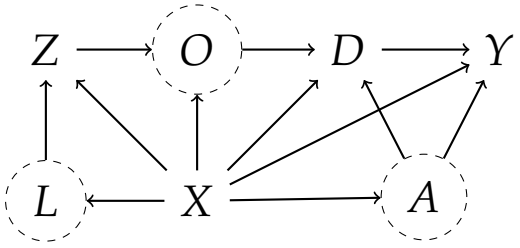

Figure 13.1: DAG for the Effect of Quality of Institutions on Wealth.

We think the story sounds plausible, but it is always important to consider threats to identification. The direct threat to identification would be if L directly affected Z and either O, D, or Y, or, in words, if early latent factors directly affected early settler mortality and either present-day quality of institutions or present day wealth. In such cases we would need to include

Lasso	Forest	Best
0.73	0.86	0.86
(0.19)	(0.33)	(0.33)

Note: Estimated coefficient from DML based estimation of a the partially linear instrumental variables model in the AJR example. Column labels denote the method used to estimate nuisance functions. The random forest produces lower cross-fit RMSEs for predicting each of Y, D, and Z, so "Best" and "Random Forest" are identical.

Table 13.1: DML Estimates of the Effect of Institutions on Output

L as additional controls. L could represent many different latent factors. For example, one might conjecture that the religion of early European settlers (e.g., Catholic vs Protestant) is related to the type of institutions they would establish and to their mortality rates upon colonization. In their original study, AJR did examine this threat by checking robustness of their result with respect to the inclusion of religion variables. They also examined the use of other additional controls to assess robustness to other potential sources of confounding.[1]

1: It is good to revisit their analysis using ML tools. See their Data archive to get started.

We report results from applying DML following the procedure outlined in Section 9.4 in Table 13.1. For cross-fitting, we use 5 folds. Here we just tried two methods, Lasso with plug-in tuning and Random Forests with package defaults, for learning the nuisance functions η. As predictors in the Lasso estimates, we used a dictionary formed by taking latitude and latitude^2 interacted with continent dummies as technical controls. For the Random Forest estimates, we simply include latitude and continent dummies as raw controls. The Random Forest predicts outcomes Y, D, and Z better than Lasso. The resulting best DML estimate is therefore based on DML with Random Forest used in all ML steps.

In this example, we see uniformly large and positive point estimates across all procedures considered, and estimated effects are statistically significant at the 5% level in all cases. We note the estimates are somewhat smaller than the baseline estimates reported in AJR – an estimated coefficient of 1.10 with estimated standard error of 0.46 ([3], Table 4, Panel A, column 7) – but are qualitatively similar, indicating a strong and positive effect of institutions on output.

13.2 DML Inference in the Interactive IV Regression Model (IRM)

DML Inference on LATE

In this section, we consider estimation of local average treatment effects (LATE) with a binary treatment variable, $D \in \{0, 1\}$, and a binary instrument, $Z \in \{0, 1\}$. As before, Y denotes the outcome variable, and X is the vector of covariates. Consider the following statistical parameter:

$$\theta_0 = \frac{\mathrm{E}[\mathrm{E}[Y \mid Z = 1, X] - \mathrm{E}[Y \mid Z = 1, X]]}{\mathrm{E}[\mathrm{E}[D \mid Z = 1, X] - \mathrm{E}[D \mid Z = 0, X]]}.$$

This parameter is the ratio of the average predictive effects of Z on Y and of D on Y. Under the assumptions laid out in Chapter 12, this statistical parameter is a causal parameter – the average treatment effect for compliers (LATE).

To set up estimation, define the regression functions:

$$\begin{aligned}
\mu_0(Z, X) &= \mathrm{E}[Y \mid Z, X] \\
m_0(Z, X) &= \mathrm{E}[D \mid Z, X] \\
p_0(X) &= \mathrm{E}[Z \mid X].
\end{aligned}$$

Define the nuisance parameter $\eta = (\mu, m, p)$ to denote square-integrable functions μ, m, and p, with μ mapping the support of (Z, X) to $\mathbb{R}$ and m and p respectively mapping the support of (Z, X) and X to $(\varepsilon, 1 - \varepsilon)$ for some $\varepsilon \in (0, 1/2)$. The true value of the nuisance parameter is $\eta_0 = (\mu_0, m_0, p_0)$, the regression functions defined above.

The DML estimator of θ_0 employs the orthogonal score

$$\begin{aligned}
\psi(W; \theta, \eta) := \; & \mu(1, X) - \mu(0, X) + H(p)(Y - \mu(Z, X)) \\
& - \Big(m(1, X) - m(0, X) + H(p)(D - m(Z, X))\Big)\theta,
\end{aligned}$$

for $W = (Y, D, X, Z)$ and

$$H(p) := \frac{Z}{p(X)} - \frac{(1 - Z)}{1 - p(X)}.$$

It is easy to verify that this score satisfies the moment condition

$$\mathrm{E}[\psi(W; \theta_0, \eta_0)] = 0$$

and also the Neyman orthogonality condition

$$\partial_\eta \mathrm{E}[\psi(W; \theta_0, \eta_0)] = 0$$

at the true value $\eta_0 = (\mu_0, m_0, p_0)$ of the nuisance parameter.

Verifying both claims is a good exercise.

Therefore we can apply the generic ML algorithm to this problem, including the selection of the best ML methods to estimate the nuisance parameters.

Theorem 13.2.1 (DML for LATE) *Suppose conditions specified in [2] hold. In particular, suppose that the overlap condition holds; namely, for some $\epsilon > 0$ with probability 1,*

$$\epsilon < p_0(X) < 1 - \epsilon.$$

Further, suppose $\epsilon < \hat{p}_{[k]}(X) < 1 - \epsilon$ and that estimators $\hat{p}_{[k]}$, $\hat{m}_{[k]}$, $\hat{\mu}_{[k]}$ provide high-quality approximation to p_0, m_0, and μ_0 in the sense that

$$n^{1/2} \|\hat{p}_0 - p_0\|_{L^2} \times \left(\|\hat{\mu}_0 - \mu_0\|_{L^2} + \|\hat{m}_0 - m_0\|_{L^2} \right) \approx 0.$$

Then estimation of the nuisance parameters does not affect the behavior of the estimator to the first order; namely,

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \sqrt{n} \mathbb{E}_n[\varphi_0(W)],$$

where

$$\varphi_0(W) = -J_0^{-1} \psi(W; \theta_0, \eta_0), \quad J_0 := \mathrm{E}\Big[m_0(1, X) - m_0(0, X) \Big].$$

Consequently, $\hat{\theta}$ concentrates in a $1/\sqrt{n}$-neighborhood of θ_0 and the sampling error $\sqrt{n}(\hat{\theta} - \theta_0)$ is approximately normal

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{a}{\sim} N(0, \mathsf{V}), \quad \mathsf{V} := \mathrm{E}[\varphi_0(W)\varphi_0(W)'].$$

Variance estimation and confidence intervals are constructed as in the generic DML algorithm.

The Effect of 401(k) Participation on Net Financial Assets

The Notebooks 13.5.3 implement DML estimation of the LATE of 401(k) participation.

Here we continue to re-analyze the effects of 401(k)'s on household financial assets, picking up from Section 9.3. In this section, we report the LATE where we take the endogenous treatment variable to be *participating* in a 401(k) plan using 401(k) *eligibility*

	Lasso	Tree	Forest	Boost	Best	Ensemble
Estimate	-5151	11320	11921	11153	11575	11471
Std. Error	(19243)	(1795)	(2023)	(1652)	(1625)	(1623)
RMSE D	0.275	0.285	0.282	0.274	0.274	0.274
RMSE Z	0.448	0.457	0.459	0.443	0.443	0.443
RMSE Y	60980	57293	54855	55133	54855	54178

Note: Estimated LATE and standard errors from the fully interactive IV model. Column labels denote the method used to estimate nuisance functions. For Lasso, we report results based on using ℓ_2 penalized logistic regression to estimate E[$D|X$] and E[$Z|X$]. The first row provides the point estimate of the LATE, and the second row provides the standard error. Rows RMSE D, RMSE Z, and RMSE Y respectively report the cross-fitted RMSE for predicting D, Z, and Y.

Table 13.2: Estimated Effect of 401(k) Participation on Net Financial Assets

as instrument. Even after controlling for features related to job choice, it seems likely that the actual choice of whether to participate in an offered plan would be endogenous. Of course, we can use eligibility for a 401(k) plan as an instrument for participation in a 401(k) plan under the conditions that were used to justify the exogeneity of eligibility for a 401(k) plan outlined in Section 9.3.

We report DML results of estimating the LATE of 401(k) participation using 401(k) eligibility as an instrument in Table 13.2. We employ the procedure outlined in Section 13.2 using the same ML estimators to estimate the quantities used to form the orthogonal estimating equation as we employed to estimate the ATE of 401(k) eligibility in Section 9.3. Looking at the results, we see that the Lasso does a very poor job predicting the outcome relative to the other considered learners and returns a negative and very imprecise coefficient estimate. The remaining learners all have similar predictive performance for each of Y, D, and Z and return similar estimates and standard errors. It is reassuring that the results obtained from the different flexible methods with similar predictive performance are broadly consistent with each other.

It is also interesting that the results based on flexible ML methods are broadly consistent with, though somewhat attenuated relative to, those obtained by applying the same specification for controls as used in [4] and [5] and using a linear IV model which returns an estimated effect of participation of $13,102 with estimated standard error of (1922). The attenuation may suggest that the simple intuitive control specification used in the original baseline specification is not sufficiently flexible.

13.3 DML Inference with Weak Instruments

Motivation

As a simple motivating example, consider an instrumental variables model with a one-dimensional endogenous variable D when there are either no controls or we are able to partial them out perfectly. In this case, the IV estimator takes the form

$$\hat{\theta} = \mathbb{E}_n[\tilde{Z}\tilde{Y}]/\mathbb{E}_n[\tilde{Z}\tilde{D}],$$

and we have that

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\mathbb{E}_n[\tilde{Z}\epsilon]/\mathbb{E}_n[\tilde{Z}\tilde{D}].$$

When $\mathbb{E}_n[\tilde{Z}\tilde{D}]$ is well-separated away from zero, we invoke the approximation

$$\sqrt{n}\mathbb{E}_n[\tilde{Z}\epsilon]/\mathbb{E}_n[\tilde{Z}\tilde{D}] \overset{a}{\sim} N(0, \mathrm{E}[\tilde{Z}^2\epsilon^2])/\mathrm{E}[\tilde{Z}\tilde{D}]. \tag{13.3.1}$$

However, this approximation is not reliable when instruments are "weak" – when $\mathbb{E}_n[\tilde{Z}\tilde{D}]$ appears close to zero. Intuitively, we may worry that small changes in a sample that result in relatively small changes in $\mathbb{E}_n[\tilde{Z}\tilde{D}]$ may still have large impacts on the estimator $\hat{\theta}$ when $\mathbb{E}_n[\tilde{Z}\tilde{D}]$ is near zero because $\mathbb{E}_n[\tilde{Z}\tilde{D}]$ shows up in the denominator. That is, (13.3.1), which essentially ignores sampling variation in $\mathbb{E}_n[\tilde{Z}\tilde{D}]$, may provide a very poor approximation to the actual finite sample sampling behavior of the IV estimator in this case.

"Weak identification" (or "weak instruments" in IV models) refers to settings in which we cannot confidently conclude a testable identifying assumption holds in our data. In our simple IV model, the parameter θ is not identified when $\mathrm{E}[\tilde{Z}\tilde{D}] = 0$ as solving the population moment condition requires solving $\mathrm{E}[\tilde{Z}\tilde{D}]\theta = \mathrm{E}[\tilde{Z}\tilde{Y}]$.

We illustrate the potential poor performance of the usual asymptotic approximation (13.3.1) in Figure 13.2 which reports results from a simulation experiment in which $\mathrm{E}[\tilde{Z}\tilde{D}]$ is close to zero. Here we see the sampling distribution (given by the blue curve) of the IV estimator deviates strongly from the normal approximation (given by the red curve). Note that by varying how close $\mathrm{E}[\tilde{Z}\tilde{D}]$ is to zero, one can make the differences more or less pronounced.

In principle, we can detect the weak instrument problem by testing whether $\beta = 0$ in the projection equation

$$\tilde{D} = \beta\tilde{Z} + U, \quad \mathrm{E}[\tilde{Z}\tilde{D}].$$

The usual asymptotic approximation relies on $\mathrm{E}[\tilde{Z}\tilde{D}]$ being so far from zero that we can ignore finite sample variability in the

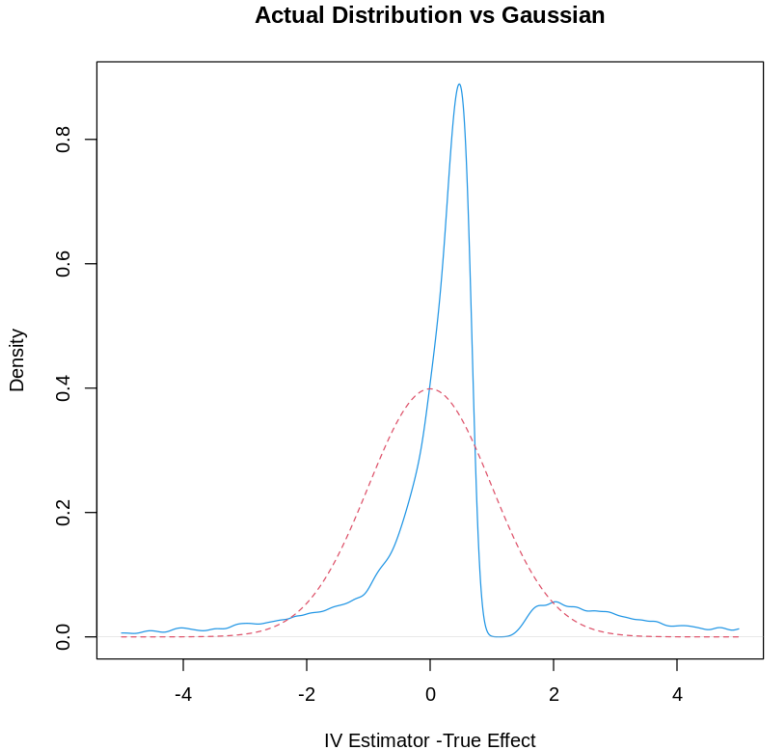

Figure 13.2: Actual sampling distribution of the IV estimator (blue) vs the normal approximation of the IV Estimator (red) in a simulation experiment using a weak instrument.

empirical analog of this expectation. Note that this statement essentially says that we need to be sure that $\beta \neq 0$ before we use the usual asymptotic approximation.

These are rules of thumb as they are based on simulations rather than formal justification.

There are a variety of "rules of thumb" for when to conclude instruments are strong in the literature. Staiger and Stock (1997) [6] suggested the most common rule of thumb used in practice. In the one endogenous variable, one instrument case, this rule of thumb corresponds to using the usual asymptotic approximation when the first stage t-statistic for testing the null hypothesis that $\beta = 0$, $|\hat{\beta} - \beta|/\text{se}(\hat{\beta})$, is bigger than $\sqrt{10} \approx 3.16$. This rule of thumb can unfortunately be very optimistic in that confidence intervals based on the usual asymptotic approximation may have coverage far from the stated coverage level – e.g. a 95% confidence interval may cover the true parameter value in far fewer than 95% of repeated samples – even when this condition is satisfied. Hansen, Hausman, and Newey [7] suggest a different rule of thumb which reduces to using the usual asymptotic approximation when the first stage t-statistic for testing the null hypothesis that $\beta = 0$ is greater than 5.6 in the one endogenous variable one instrument case. More recent work, e.g. Andrews (2018) [8], suggests that one should be cautious in applying any such rule of thumb.

All of these results suggest that the usual asymptotic approxi-

A related caution is that basing inference on the usual asymptotic approximation after seeing the result of a test for the strength of association between $\tilde{D}$ and $\tilde{Z}$ can introduce substantial pre-test bias that further distorts inference. See Andrews, Stock, and Sun (2019) [9].

mation may not be safe if we are worried that our instruments are not strongly related to the endogenous variables. If we have such worries, is there anything else we can do?

Of course there is. There are a variety of alternative inferential procedures whose behavior does not hinge on the well-separation of $\mathbb{E}_n[\tilde{Z}\tilde{D}]$ from zero. Here, we consider one specific approach based upon the statistic

$$C(\theta) = \frac{|\mathbb{E}_n[(\tilde{Y} - \theta\tilde{D})\tilde{Z}]|^2}{\mathbb{V}_n[(\tilde{Y} - \theta\tilde{D})\tilde{Z}]/n}.$$

The statistic $C(\theta)$ is Neyman's $C(\alpha)$ statistic [10]. In the case we consider here, the statistic is essentially the same as considered in Anderson and Rubin (1949) [11] without imposing homoskedasticity as in Stock and Wright (2000) [12].

If $\theta_0 = \theta$, then $C(\theta) \overset{a}{\sim} N(0,1)^2 = \chi^2(1)$. Therefore, we can reject the hypothesis $\theta_0 = \theta$ at level a (for example $\mathsf{a} = .05$ for a 5% level test) if $C(\theta) > c(1-\mathsf{a})$ where $c(1-\mathsf{a})$ is the $(1-\mathsf{a})$-quantile of a $\chi^2(1)$ variable. The probability of falsely rejecting the true hypothesis is approximately $\mathsf{a} \times 100\%$. To construct a $(1-\mathsf{a}) \times 100\%$ confidence region for θ, we can then simply invert the test by collecting all parameter values that are not rejected at the a level:

$$CR(\theta) = \{\theta \in \Theta : C(\theta) \leq c(1-\mathsf{a})\}.$$

In more complex settings with many controls or controls that enter with unknown functional form, we can simply replace the residuals $\tilde{Y}$, $\tilde{D}$, and $\tilde{Z}$ by machine learned cross-fitted residuals $\check{Y}$, $\check{D}$, and $\check{Z}$. Thanks to the orthogonality of the IV moment condition underlying the formulation outlined above, we can formally assert that the properties of $C(\theta)$ and the subsequent testing procedure and confidence region for θ continue to hold when using cross-fitted residuals. We will further be able to apply the general procedure to cases where D is a vector, with a suitable adjustment of the statistic $C(\theta)$.

DML Inference Robust to Weak-IV in PLMs

Here, we present a more general version of weak identification robust inference, including implementation and theoretical details, in settings where we want to use machine learning to aid in controlling for confounding variables X.

DML Weak-IV-Robust Inference for PLIV Model

1. **Initialize:** Let Θ be a known parameter space that contains the true value θ_0. Using the DML-PLIV

algorithm, produce the cross-fitted residuals: $\check{Y}_i$, $\check{D}_i$, and $\check{Z}_i$. Using the cross-fitted residuals and for $\theta \in \Theta$, compute the moment function

$$\check{\mathsf{M}}(\theta) := \mathbb{E}_n[(\check{Y}_i - \theta' \check{D}_i)\check{Z}_i],$$

the empirical covariance function

$$\check{\Omega}(\theta) := \mathbb{V}_n[(\check{Y} - \theta' \check{D})\check{Z}],$$

and the score statistic

$$C(\theta) := n\check{\mathsf{M}}(\theta)'\check{\Omega}^{-1}(\theta)\check{\mathsf{M}}(\theta).$$

2. **Robust Confidence Region:** Construct the approximate $(1 - \mathsf{a}) \times 100\%$ confidence region as

$$CR(\theta_0) = \{\theta \in \Theta : C(\theta) \leq c(1 - \mathsf{a})\},$$

where $c(1 - \mathsf{a}) :=$ $(1 - \mathsf{a})$-quantile of a $\chi^2(m)$ variable, where $m = \dim(Z_i)$.

In order to state the next result, define the oracle version of the moment and covariance functions given in Step 1 of the DML Weak-IV-Robust Inference algorithm,

$$\hat{\mathsf{M}}(\theta) = \mathbb{E}_n[(\tilde{Y} - \theta' \tilde{D})\tilde{Z}]$$

and

$$\hat{\Omega}(\theta) = \mathbb{V}_n[(\tilde{Y} - \theta' \tilde{D})\tilde{Z}],$$

which are defined in terms of the true residuals $\tilde{Y}_i$, $\tilde{D}_i$, and $\tilde{Z}_i$.

Theorem 13.3.1 *Under regularity conditions, estimation of the nuisance parameters does not affect the behavior of the $C(\theta)$ statistic in the sense that*

$$C(\theta_0) \approx n\hat{\mathsf{M}}(\theta_0)'\hat{\Omega}^{-1}(\theta_0)\hat{\mathsf{M}}(\theta_0) \overset{a}{\sim} \chi^2(m).$$

Consequently, the test rejects the true value with approximate probability a*,*

$$\mathrm{P}(C(\theta) \geq c(1 - \mathsf{a})) \approx \mathsf{a},$$

and the confidence region $CR(\theta_0)$ contains θ_0 with approximate

probability $(1 - \mathsf{a})$,

$$P(\theta_0 \in CR(\theta_0)) \approx (1 - \mathsf{a}).$$

The Effect of Institutions on Economic Growth Revisited

We illustrate the use of DML weak identification robust inference by revisiting the AJR example from Section 13.1. Recall that Random Forests performed best in all auxiliary predictive steps in our original exercise in this example, so we only consider the use of Random Forests to form residuals in this section.

After partialling out controls using Random Forests, we run the regression of $\check{D}$ on $\check{Z}$ to assess the strength of the instruments. The resulting t-statistic is approximately 2, much lower than any rule-of-thumb "safety" threshold that appears in the literature. As such, we conclude that we have a weak instrument and proceed with weak identification robust inference.

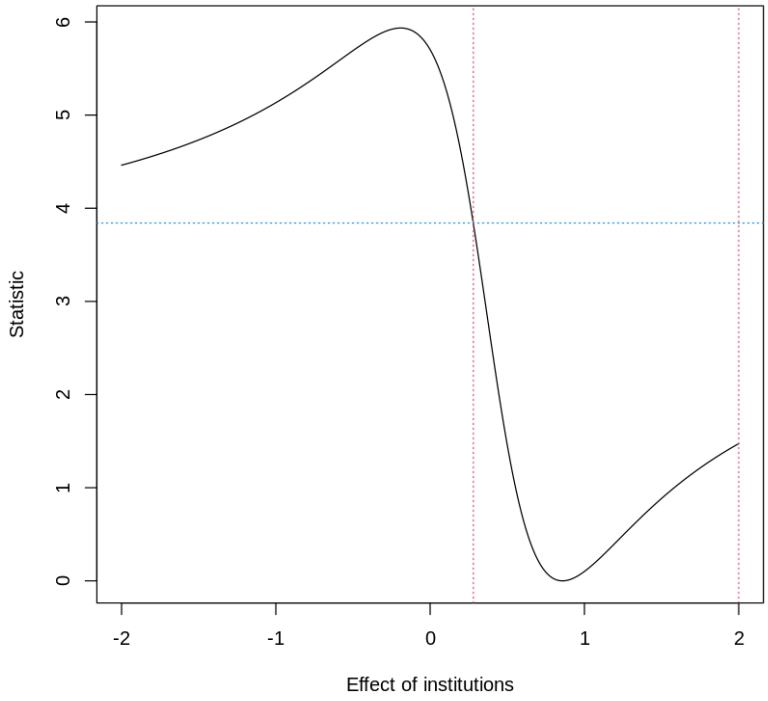

Figure 13.3: Construction of weak IV robust confidence regions for the effect of institutions on output using DML. Values of the $C(\theta)$ statistic are shown on the vertical axis; values of θ tested on the horizontal axis. The 90% confidence region is given by the red vertical bars.

We implement the robust inferential approach from the previous subsection considering $\Theta = [-2, 2]$ as our parameter space for the causal effect of institutions on wealth. We note that, because the outcome we consider is the logarithm of GDP per capita, the range [-2,2] includes extremely (likely implausibly) large

negative and positive effects, so restricting attention to this range *a priori* seems reasonable. We illustrate the procedure in Figure 13.3 which plots the value of the test statistic $C(\theta)$ for $\theta \in [-2, 2]$.

The resulting 95% confidence region is

$$[.28, 2].$$

We can compare this region to the confidence region produced by the usual Gaussian asymptotic approximation which is not robust to weak identification:

$$[.86 \pm 2 \cdot 0.33] = [.20, 1.52].$$

Both the usual and robust confidence regions are consistent with relatively large positive effects of institutions on wealth. However, it is interesting that the lower end of the robust confidence region is larger than the lower end of the usual region and that this difference is economically meaningful. That is, we could not rule out that a one unit increase in quality of institutions causes an approximately a 20% increase in GDP per capita looking at the usual interval, while we could rule out all effect sizes smaller than 28% with the robust interval. The difference between a 20% and 28% increase in GDP per capita is small but certainly economically relevant. Given that the instruments are weak, we should, of course, rely on the robust confidence interval.

13.4 Generic DML Inference under Weak Identification

We now present a generally applicable formulation of weak identification robust inference. This formulation covers the problem of weak instruments in the context of LATE estimation as well as other problems where Neyman orthogonal scores are available.

The initialization and first two steps to our approach to weak identification robust inference are the same as in the Generic DML Algorithm. We then use these estimates of the nuisance parameters in conjunction with the score function at a fixed value of θ to construct a score test statistic analogous to $C(\theta)$ from the previous section which can be used to test the hypothesis that $\theta_0 = \theta$ and to form confidence regions. We collect this procedure in the following algorithm:

Generic DML Robust to Weak Identification

1. **Initialize:** Provide the data frame $(W_i)_{i=1}^n$, the Neyman orthogonal score/moment function $\psi(W, \theta, \eta)$ and the name and model for ML estimation method(s) for learning nuisance parameters η. Specify Θ to be a known parameter space that contains the true value θ_0. We then take a K-fold random partition $(\mathcal{I}_k)_{k=1}^K$ of observation indices $\{1, ..., n\}$ such that the size of each fold is about the same. For each $k \in \{1, \ldots, K\}$, we construct a machine learning estimator $\hat{\eta}_{[k]}$ using data $(W_i)_{i \notin \mathcal{I}_k}$, that is, all the data *except* the data from the k^{th} fold.

2. **Estimate Moments and Their Variance:** Letting $k(i) = \{k : i \in I_k\}$, construct the moment function

$$\check{\mathsf{M}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi(W_i; \theta, \hat{\eta}_{[k(i)]})$$

covariance function,

$$\check{\Omega}(\theta) = \frac{1}{n} \sum_{i=1}^{n} [\psi(W_i; \theta, \hat{\eta}_{[k(i)]}) \psi(W_i; \theta, \hat{\eta}_{[k(i)]})'] - \frac{1}{n} \sum_{i=1}^{n} [\hat{\psi}(W_i; \theta, \hat{\eta}_{[k(i)]})] \frac{1}{n} \sum_{i=1}^{n} [\psi(W_i; \theta, \hat{\eta}_{[k(i)]})]',$$

and score statistic

$$C(\theta) = n \check{\mathsf{M}}(\theta)' \check{\Omega}^{-1}(\theta) \check{\mathsf{M}}(\theta).$$

3. **Confidence Region:** Construct the approximate $(1 - \mathsf{a}) \times 100\%$ confidence region as

$$CR(\theta_0) = \{\theta \in \Theta : C(\theta) \leq c(1 - \mathsf{a})\}$$

where $c(1 - \mathsf{a})$ is the $(1 - \mathsf{a})-$quantile of a $\chi^2(m)$ variable where $m = \dim(\check{\mathsf{M}}(\theta))$.

Note that this confidence region simply collects all values $\theta \in \Theta$ that are not rejected by testing $\theta_0 = \theta$ using test statistic $C(\theta)$ at the a-level.

As in the previous section, we define oracle versions of the moment and covariance functions from the preceding algorithm

for use in stating formal results:

$$\hat{\mathsf{M}}(\theta) = \mathbb{E}_n[\psi(W;\theta,\eta_0)],$$

$$\hat{\Omega}(\theta) = \mathbb{V}_n[\psi(W;\theta,\eta_0)].$$

Theorem 13.4.1 *Under regularity conditions, estimation of nuisance parameters does not affect the behavior of the $C(\theta)$ statistic in the sense that*

$$C(\theta_0) \approx n\hat{\mathsf{M}}(\theta_0)\hat{\Omega}^{-1}(\theta_0)\hat{\mathsf{M}}(\theta_0) \overset{a}{\sim} \chi^2(m).$$

Consequently, a test that rejects when $C(\theta) \geq c(1-\mathsf{a})$, for $c(1-\mathsf{a})$ the $(1-\mathsf{a})-$quantile of a $\chi^2(m)$ variable, rejects the true value with approximate probability a:

$$\mathrm{P}(C(\theta_0) \geq c(1-\mathsf{a})) \approx \mathsf{a}.$$

Similarly, the confidence region corresponding to this test, $CR(\theta_0)$, contains θ_0 with approximate probability $(1-\mathsf{a})$:

$$\mathrm{P}(\theta_0 \in CR(\theta_0)) \approx (1-\mathsf{a}).$$

13.5 Notebooks

Notebook 13.5.1 (Weak IV) R Notebook on Weak IV and Python Notebook on Weak IV provide a simulation experiment illustrating the weak instrument problem with IV estimators.

Notebook 13.5.2 (DML for Partially Linear IV) DML for Partially Linear IV R Notebook and DML for Partially Linear IV Python Notebook carry out the DML IV analysis of the Acemoglu-Johnson-Robinson example, which considers the impact of the quality of institutions on economic growth, instrumenting quality of institutions with settler mortality. The notebook explores the partially linear IV model and tests for the presence of weak instruments.

Notebook 13.5.3 (DML for LATE Models) DML for LATE Models R Notebook and DML for LATE Models Python Notebook estimate the Local Average Treatment Effects of 401(K) participation on net financial wealth.

13.6 Exercises

Exercise 13.6.1 (Weak IV) Experiment with Notebook 13.5.1, varying the strength of the instrument. How strong should the instrument be in order for the conventional normal approximation based on strong identification to provide accurate inference? Based on your experiments, provide a brief explanation of the weak IV problem to a friend.

Exercise 13.6.2 (DML for Partially Linear IV) Experiment with Notebook 13.5.2. Try to extend the analysis by including other control variables (e.g. religion, other measures of geography, or measures of natural resources) or consider another empirical application to another IV example. (See some potential applications at the the Angrist data archive). In the case of a new application, don't forget to draw your DAGs!

Exercise 13.6.3 (DML for LATE Models) Experiment with the Notebook 13.5.3. Apply the analysis to another dataset. For example, try the JTPA data from Joshua Angrist's data archive. Don't forget to draw your DAGs!

Exercise 13.6.4 ((Theoretical) Neyman Orthogonality of the Partially Linear IV Methods) Verify that the scores for the partially linear IV methods are Neyman orthogonal.

Bibliography

[1] Guido W. Imbens. 'Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)'. In: *Journal of Economic Literature* 48.2 (2010), pp. 399–423. DOI: 10.1257/jel.48.2.399 (cited on page 346).

[2] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 'Double/debiased machine learning for treatment and structural parameters'. In: *Econometrics Journal* 21.1 (2018), pp. C1–C68 (cited on pages 348, 353).

[3] Daron Acemoglu, Simon Johnson, and James A. Robinson. 'The colonial origins of comparative development: An empirical investigation'. In: *American Economic Review* 91.5 (2001), pp. 1369–1401 (cited on pages 349, 351).

[4] James M. Poterba, Steven F. Venti, and David A. Wise. '401(k) Plans and Tax-Deferred savings'. In: *Studies in the Economics of Aging*. Ed. by D. A. Wise. Chicago, IL: University of Chicago Press, 1994, pp. 105–142 (cited on page 354).

[5] James M. Poterba, Steven F. Venti, and David A. Wise. 'Do 401(k) Contributions Crowd Out Other Personal Saving?' In: *Journal of Public Economics* 58.1 (1995), pp. 1–32 (cited on page 354).

[6] D. Staiger and J. H. Stock. 'Instrumental Variables Regression with Weak Instruments'. In: *Econometrica* 65 (1997), pp. 557–586 (cited on page 356).

[7] Christian Hansen, Jerry Hausman, and Whitney K. Newey. 'Estimation with Many Instrumental Variables'. In: *Journal of Business and Economic Statistics* 26 (4 2008), pp. 398–422 (cited on page 356).

[8] Isaiah Andrews. 'Valid Two-Step Identification-Robust Confidence Sets for GMM'. In: *The Review of Economics and Statistics* 100.2 (May 2018), pp. 337–348. DOI: 10.1162/REST_a_00682 (cited on page 356).

[9] Isaiah Andrews, James Stock, and Liyang Sun. In: *Annual Review of Economics* 11 (2019), pp. 727–753 (cited on page 356).

[10] Jerzy Neyman. 'Optimal asymptotic tests of composite hypotheses'. In: *Probability and Statistics* (1959), pp. 213–234 (cited on page 357).

[11] T. W. Anderson and H. Rubin. 'Estimation of the Parameters of Single Equation in a Complete System of Stochastic Equations'. In: *Annals of Mathematical Statistics* 20 (1949), pp. 46–63 (cited on page 357).

[12] James H. Stock and Jonathan H. Wright. 'GMM with Weak Identification'. In: *Econometrica* 68 (2000), pp. 1055–1096 (cited on page 357).