

# Applied Causal Inference Powered by ML and AI

Victor Chernozhukov\*

Christian Hansen<sup>†</sup>

Nathan Kallus<sup>‡</sup>

Martin Spindler<sup>§</sup>

Vasilis Syrgkanis<sup>¶</sup>

February 28, 2024

Publisher: Online

Version 0.1.1

\* MIT

<sup>†</sup> Chicago Booth

<sup>‡</sup> Cornell University

<sup>§</sup> Hamburg University

<sup>¶</sup> 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 8. 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 . . . . .	345
The Effect of Institutions on Economic Growth . . . . .	347
13.2 DML Inference in the Interactive IV Regression Model (IRM) . . . . .	350
DML Inference on LATE	350
The Effect of 401(k) Participation on Net Financial Assets	351
13.3 DML Inference with Weak Instruments . . . . .	352
Motivation . . . . .	352
DML Inference Robust to Weak-IV in PLMs . . . . .	355
The Effect of Institutions on Economic Growth Revisited . . . . .	356
13.4 Generic DML Inference under Weak Identification .	358

## 13.1 DML Inference in Partially Linear IV Models

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

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

where

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

and

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

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 8.

To estimate  $\theta_0$  and to perform inference on it 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  $\ell$ ,  $m$ , and  $r$  being  $P$ -square-integrable functions mapping the support of  $X$  to  $\mathbb{R}$ . By definition, we have that

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

and it is not difficult to check (via homework) that the Neyman orthogonality condition,

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

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

**DML for Partially Linear IV and Proxy Models**

1. Partition data indices into  $k$  folds of approximately equal size:  $\{1, \dots, n\} = \cup_{k=1}^K I_k$ . For each fold  $k = 1, \dots, K$ , 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  $i \in I_k$ :

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

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  $\theta$  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

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

where, as before,  $\mathbb{E}_X$  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  $\mathbb{E}[\tilde{D}\tilde{Z}]$  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  $\hat{\theta}$ :

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

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

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

where

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

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

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

covers  $\theta$  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

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 [3] (AJR).

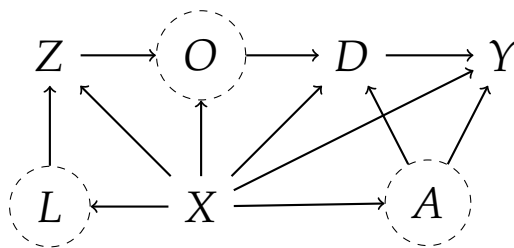
We use the same set of 64 country-level observations as AJR. The data set contains measurements of GDP, settler mortality, an index which measures protection against expropriation risk and geographic information. The outcome variable,  $Y$ , is the logarithm of GDP per capita and the endogenous explanatory variable,  $D$ , is a measure of the strength of individual property rights 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  $Z \rightarrow D$  and  $Z \rightarrow Y$ . From the DAG, we see that  $X$  blocks the backdoor paths from  $Y$  to  $Z$  and from  $D \rightarrow Z$ . 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  $L$  as additional controls.  $L$  could represent many different latent

Lasso	Forest	Best
0.77	0.88	0.88
(0.17)	(0.32)	(0.32)

**Note:** Estimated coefficient from DML based estimation of a linear instrumental variables model based on orthogonal estimating equations in the AJR example. Column labels denote the method used to estimate nuisance functions. We used  $K = 20$  folds for cross-fitting.

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

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.<sup>1</sup>

We report results from applying DML following the procedure outlined in Section 10.4 in Table 13.1. For cross-fitting, we use 20 folds given how small the data set is. Here we just tried two successful methods, Lasso and Random Forests, for learning the nuisance functions  $\eta$ . As predictors in the Lasso estimates, we used a dictionary formed by taking latitude and latitude<sup>2</sup> 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.

<sup>1</sup>: It is good to revisit their analysis using ML tools. See their [Data archive](#) to get started.

## 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:

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

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 8, this statistical parameter is a causal parameter – the average treatment effect for compliers (LATE).

To set up estimation, define the regression functions:

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

Define the nuisance parameter  $\eta = (\mu, m, p)$  to denote square-integrable functions  $\mu$ ,  $m$ , and  $p$ , with  $\mu$  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  $\theta_0$  employs the orthogonal score

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

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

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

It is easy to verify (for homework) that this score satisfies the moment condition

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



and also the Neyman orthogonality condition

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

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

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  $\mu_0$  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,*

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

where

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

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

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

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

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

Here we continue to re-analyze the effects of 401(k)'s on household financial assets, picking up from Section 10.3. In this section, we report the LATE in this example where we take the endogenous treatment variable to be *participating* in a 401(k)

plan using 401(k) *eligibility* 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 10.3.

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 10.3, so we omit the details for brevity. Looking at the results, we see that the estimated causal effect of 401(k) participation on net financial assets is uniformly positive and statistically significant across all of the considered methods. As when looking at the ATE of 401(k) eligibility, it is reassuring that the results obtained from the different flexible methods 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.

Lasso	Forest	Boosting	Neural Net.	Ensemble	Best
8944 (2259)	11764 (1788)	11133 (1661)	11186 (1795)	11173 (1641)	11113 (1645)

**Note:** Estimated LATE under a fully interactive IV model. Column labels denote the method used to estimate nuisance functions.

R Notebook on DML for Impact of 401(K) Participation on Financial Wealth

**Table 13.2:** DML Estimates of LATE on 401(k) Participation on Net Financial Assets

## 13.3 DML Inference with Weak Instruments

### Motivation

As a simple motivating example, consider a statistical model with instrumental moment conditions and 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

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

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.

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  $\mathbb{E}[\tilde{Z}\tilde{D}]$  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  $\mathbb{E}[\tilde{Z}\tilde{D}]$  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

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

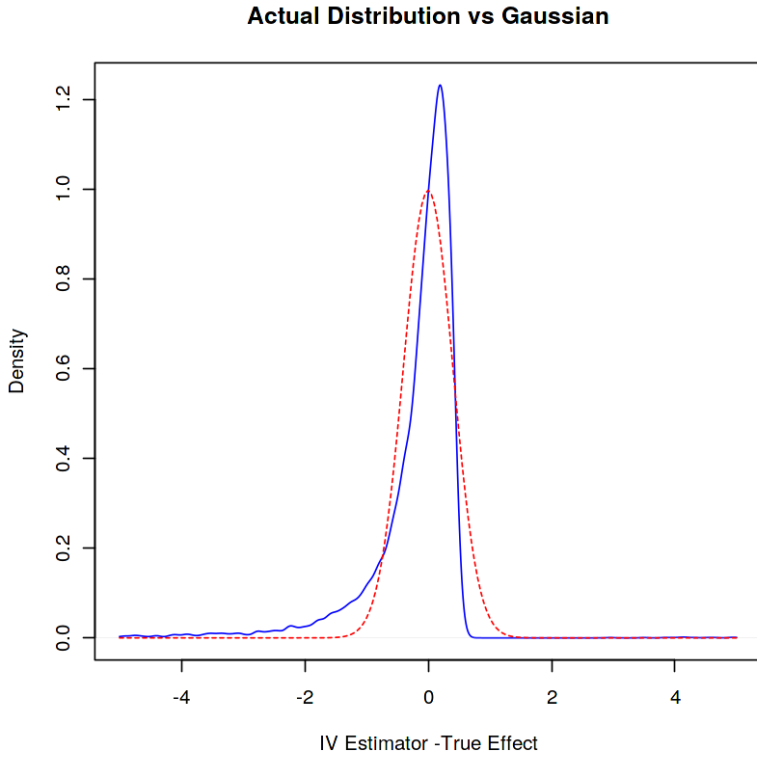
Econometricians have found that the normal approximation above is adequate for inferential properties if the t-statistic for testing the null  $\beta = 0$  is bigger than 4:<sup>2</sup>

$$|\hat{\beta} - \beta|/\text{se}(\hat{\beta}) > 4.$$

If this happens, then we are said to have a "strong" instrument. If this test for the strong instrument is passed, then it is relatively safe to use the normal approximation for inference with the IV

"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  $\theta$  is not identified when  $\mathbb{E}[\tilde{Z}\tilde{D}] = 0$  as solving the population moment condition requires solving  $\mathbb{E}[\tilde{Z}\tilde{D}]\theta = \mathbb{E}[\tilde{Z}\tilde{Y}]$ .

2: These are called "rules of thumb" and are based on simulation experiments.



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

estimator. If not, using the usual asymptotic approximation is not safe, but is there anything else that 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<sup>3</sup>

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

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

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

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

3: The empirical variance  $\mathbb{V}_n$  is defined as:

$$\mathbb{V}_n[g(W_i)] := \mathbb{E}_n g(W_i)g(W_i)' - \mathbb{E}_n[g(W_i)]\mathbb{E}_n[g(W_i)]'.$$

$\check{Y}$ ,  $\check{D}$ , 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  $\theta$  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  $\Theta$  be a known parameter space that contains the true value  $\theta_0$ . 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

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

the empirical covariance function

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

and the score statistic

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

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

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

where  $c(1 - a) := (1 - a)$ -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,

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

and

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

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

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

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

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

$$P(C(\theta) \geq c(1 - a)) \approx a,$$

and the confidence region  $CR(\theta_0)$  contains  $\theta_0$  with approximate probability  $(1 - a)$ ,

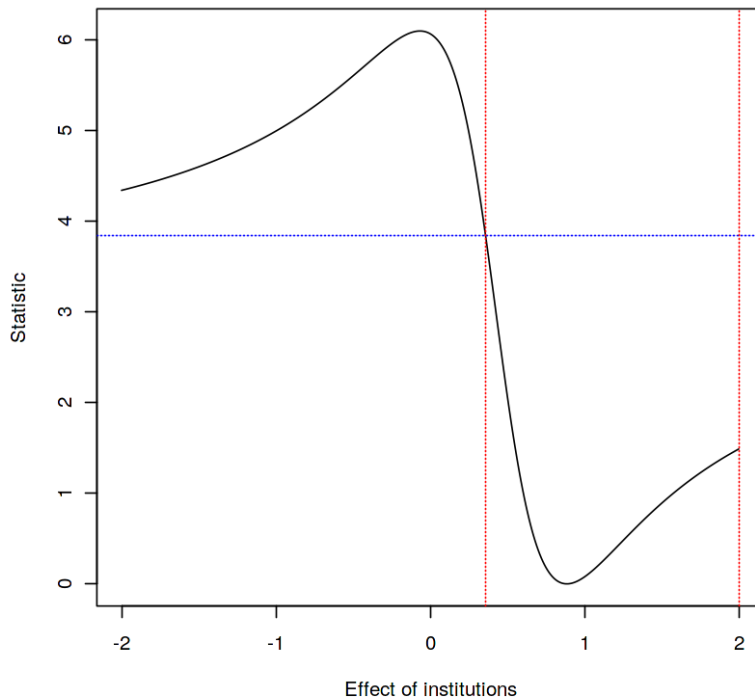
$$P(\theta_0 \in CR(\theta_0)) \approx (1 - a).$$

## 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 the "safety" threshold of 4. As such, we conclude that we have a weak instrument and proceed with weak identification robust inference.

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  $\theta \in [-2, 2]$ .



**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  $\theta$  tested on the horizontal axis. The 90% confidence region is given by the red vertical bars.

The resulting 95% confidence region is

$$[.35, 2].$$

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

$$[.88 \pm 2 \cdot 0.32] = [.24, 1.52].$$

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 27% increase in GDP per capita looking at the usual interval, while we could rule out all effect sizes smaller than 42% with the robust interval. The difference between a 27% and 42% increase in GDP per capita is 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  $\theta$  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  $\eta$ . Specify  $\Theta$  to be a known parameter space that contains the true value  $\theta_0$ . We then take a  $K$ -fold random partition  $(I_k)_{k=1}^K$  of observation indices  $\{1, \dots, n\}$  such that the size of each fold is about the same, and for each  $k \in \{1, \dots, K\}$ , we construct a machine learning estimator  $\hat{\eta}_{[k]}$  using data  $(W_i)_{i \notin I_k}$ , that is, all the data *except* the data from the  $k^{\text{th}}$  fold.
2. **Estimate Moments and Their Variance:** Letting  $k(i) = \{k : i \in I_k\}$ , construct the moment function

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

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

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

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

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

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

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:

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

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

**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*

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

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

$$P(C(\theta_0) \geq c(1 - a)) \approx a.$$

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

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

## Notebooks

- [The R Notebook on Weak IV](#) provides a simulation experiment illustrating the weak instrument problem with IV

estimators.

- ▶ [R Notebook on DML for Impact of Institutions on Output](#) provides DML analysis of the impact of institutions on a country's wealth following AJR. The notebook first proceeds with the analysis assuming strong identification. It then notes the weak instrument problem and performs DML analysis that is robust to weak identification.
- ▶ [R Notebook on DML for Impact of 401\(K\) Eligibility on Financial Wealth](#) provides application of DML inference to learn predictive/causal effects of 401(K) eligibility on net financial wealth. (Note: The results produced in this notebook and provided in the text are slightly different than those in the original paper [2]. The replication files for [2] are given at the following [Github repository](#). The difference is due to our use of a single split of the sample in producing the results for this text while the results in [2] are based on a method that aggregates results across multiple data splits.)

## Notes

The statistic  $C(\theta)$  is Neyman's  $C(\alpha)$  statistic.

## Study Problems

1. Experiment with [The R Notebook on Weak IV](#), 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.
2. Experiment with [R Notebook on DML for Impact of 401\(K\) Eligibility on Financial Wealth](#). Apply the analysis to another data-set. For example, try the JTPA data from [Joshua Angrist's data archive](#). Don't forget to draw your DAGs!
3. Experiment with [R Notebook on DML for Impact of Institutions on Output](#). 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!

4. (Theoretical). 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](https://doi.org/10.1257/jel.48.2.399) (cited on page 344).
- [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 346, 351, 360).
- [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 347, 349).
- [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 352).
- [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 352).