Introduction to Debiased Machine Learning in Big Data
International Iranian Economic Association Webinar

Whitney K. Newey MIT and NBER

6 September 2023

1 Introduction

Many interesting objects depend on a regression or other first step.

–OLS regression with many covariates.

–IV regression with many covariates.

–Average treatment effect (ATE) depends on average outcome given covariates and treatment and/or propensity score.

–Discrete dynamic structural economic models depend on conditional choice probabilities.

–Average equivalent variation bounds depend on average demand.

The regression (or other first step) may be high dimensional, e.g. there may be many covariates.

Machine learning provides good prediction with many regressors.

Methods include Lasso, Neural Nets, random forests, boosting.

Give excellent predictions but biased by regularization and/or model selection.

Regularization/model selection biases come from reducing variance.

For Lasso we penalize the sum of square residuals by the size of the absolute value of regression coefficients.

If "plug-in" machine learner into formula for parameter of interest the regularization biases "pass through" and give poorly centered confidence intervals for the parameter of interest.

Also, if "plug-in" learner with model selection then local mistakes under root-n alternatives lead to invalid confidence intervals parameters in root-n neighborhood; Leeb and Potscher (2005).

A solution to regularization and model selection biases is Neyman orthogonal moment functions for GMM.

Orthogonality means first step has no effect, to first order, on average moment function.

Also use cross-fitting to reduce overfitting bias and allow for complicated machine learning first steps.

References:

For an introduction see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal* 21, C1-C68.

Theory given in Chernozhukov, Escanciano, Ichimura, Newey, Robins (2022): "Locally Robust Semiparametric Estimation," *Econometrica* 90, 1501-1535.

Automatic debiasing introduced here: Chernozhukov, Newey, Singh (2022): "Automatic Debiased Machine Learning of Structural and Causal Effects," *Econometrica 90, 967-1027.*

Only need to know the orthogonal scores to construct estimators.

Many more recent papers; Handbook of Econometrics chapter under development.

Today work through introductory notes from 1st year graduate econometrics course at MIT including empirical examples.

Consider GMM estimation, where θ is the main target structural parameter, and we have other (very high dimensional) nuisance parameters η, that can identified and estimated without knowing θ.

We outline a general "debiased" GMM approach that provides high quality estimation and inference of θ.

This is not the most general setting, but it is a very practical one.

We consider whether the preliminary estimation of nuisance parameters by modern high-dimensional methods, such as lasso, has a spill-over effect on the estimation of the main parameter.

In general the answer is yes, posing difficulties for inference about θ.

However, the spill-over can be eliminated by working with scores (i.e. moment functions for GMM) that have "Neyman orthogonal structure", where the moment equations pinning down the target parameters have zero sensitivity to small perturbations in the nuisance parameters.

In fact "partialling out", used to estimate regression parameters of interest (like average treatment effects) in the presence of nuisance parameters (like covariate coefficients), produces moments that have this "orthogonality" structure.

Orthogonality turns out to be the most crucial property where the nuisance parameters η are high-dimensional, necessitating the use of regularized (and hence biased estimators) of nuisance parameters.

We outline a general "debiased" GMM approach that provides high quality estimation and inference of θ based on the use of orthogonal scores.

Having orthogonality structure is also convenient even when the nuisance parameters are low-dimensional, because it leads to simple variance calculations of the resulting estimators in low-dimensional settings, that are widely used in applied econometrics.

2 Inference in Linear Regression with Many Controls

Consider the regression model,

$$Y = \alpha D + \beta' W + \epsilon, \tag{1}$$

where D is the target regressor and W consists of p controls, and controls are high dimensional (not small compared to the available sample size n).

The key to setting up high quality inference on α will be to carry out partialling-out.

After partialling-out,

$$\tilde{Y} = \alpha \tilde{D} + \epsilon, \quad E[\epsilon \tilde{D}] = 0, \tag{2}$$

where the variables with tilde are residuals from taking out the linear effect of W

$$\tilde{D} = D - \gamma_{DW}' W, \quad \gamma_{DW} = \arg\min_{\gamma \in \mathbb{R}^p} E[(D - \gamma' W)^2],$$

$$\tilde{Y} = Y - \gamma_{YW}' W, \quad \gamma_{YW} = \arg\min_{\gamma \in \mathbb{R}^p} E[(Y - \gamma' W)^2],$$

Then α can be recovered from population linear regression of $\tilde{Y}$ on $\tilde{D}$:

$$\alpha = \arg\min_{a \in \mathbb{R}} E[(\tilde{Y} - a\tilde{D})^2] = (E[\tilde{D}^2])^{-1} E[\tilde{D}\tilde{Y}].$$

Note also that $a = \alpha$ solves the moment equation:

$$E[(\tilde{Y} - a\tilde{D})\tilde{D}] = 0].$$

This is the idea behind the Frisch-Waugh-Lovell partialling-out.

We now consider estimation of α in high-dimensional setting; for estimation purposes we have a random sample $(Y_i, X_i)_{i=1}^n$.

We will mimic in the sample the partialling-out procedure in the population.

Previously, when p/n was small, we employed ordinary least squares as the prediction method in the partialling-out steps.

Here p/n is not small, and we employ instead Lasso-based methods in the partialling-out steps; Lasso

The estimation procedure can be summarized as follows:

Double Lasso:

1. We run the Lasso regressions of Y_i on W_i and D_i on W_i:

$$\hat{\gamma}_{YW} = \arg\min_{\gamma \in \mathbb{R}^p} \quad \sum_i (Y_i - \gamma' W_i)^2 + \lambda_1 \sum_j |\gamma_j|,$$

$$\hat{\gamma}_{DW} = \arg\min_{\gamma \in \mathbb{R}^p} \quad \sum_i (D_i - \gamma' W_i)^2 + \lambda_2 \sum_j |\gamma_j|,$$

   and obtain the resulting residuals:

$$\check{Y}_i = Y_i - \hat{\gamma}'_{YW} W_i,$$

$$\check{D}_i = D_i - \hat{\gamma}'_{DW} W_i.$$

   In place of Lasso we can also use Post-Lasso.

2. We run the least squares of $\check{Y}_i$ on $\check{D}_i$ to obtain the estimator $\check{\alpha}$ as the root of:

$$\frac{1}{n} \sum_{i=1}^{n} (\check{Y}_i - a \check{D}_i) \check{D}_i = 0 \qquad (3)$$

   and use it for standard inference, proceeding as if the residuals were known.

Approximate sparsity of the population regression coefficients γ_{YW} and γ_{DW}, with a sufficiently large rate of decrease a in the sorted coefficients,

$$|\gamma_{YW}|_{(j)} \leq Aj^{-a}, \quad |\gamma_{DW}|_{(j)} \leq Aj^{-a} \quad a > 1, \quad j = 1, \ldots, p.$$

For this estimation procedure the following theorem can be shown:

Theorem 1 (Adaptive Inference with Double Lasso) *Under the stated approximate sparsity and additional regularity conditions, the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first order effect on $\check{\alpha}$, and*

$$\sqrt{n}(\check{\alpha} - \alpha) \approx \sqrt{n} E_n(\tilde{D}\epsilon)/E_n(\tilde{D}^2) a d N(0, V),$$

where

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

The above statement means that $\check{\alpha}$ concentrates in a $2\sqrt{V/n}$- neighborhood of α, with deviations controlled by the normal law.

Observe that the approximate behavior of the double lasso estimator is the same as the approximate behavior of the least squares estimator in low-dimensional models.

This result can be used for construction of confidence intervals.

Just like the low-dimensional case, we can use these results to construct a confidence interval for α.

Obtain the standard error of $\check{\alpha}$ as

$$\sqrt{\hat{V}/n},\ \hat{V} = n(\sum_{i=1}^{n} \check{D}_i^2)^{-1} \sum_{i=1}^{n} (\check{Y}_i - a\check{D}_i)^2 \check{D}_i^2 (\sum_{i=1}^{n} \check{D}_i^2)^{-1}$$

where $\hat{V}$ is the plug-in estimator of V. The result implies that the interval

$$[\check{\alpha} \pm 2\sqrt{\hat{V}/n}]$$

covers α in about 95% of the time.

2.1 Application to Testing Convergence Hypothesis

We provide an empirical example of partialling-out with Lasso to estimate the regression coefficient α in the high-dimensional linear regression model:

$$Y = \alpha D + \beta' W + \epsilon.$$

Specifically we are interested in how economic growth rates (Y) are related to the initial wealth levels in each country (D) controlling for country's institutional, educational, and other similar characteristics (W).

The relationship is captured by α, the "speed of convergence/divergence", which predicts the speed at which poor countries catch up $(\alpha < 0)$ or fall behind $(\alpha > 0)$ rich countries, after controlling for W.

The empirical question here is: do poor countries grow faster than rich countries, controlling for educational and other characteristics?

In other words, is the speed of convergence negative (this is the convergence hypothesis predicted by the several growth models, including the classical Solow model).

Under some strong assumptions that we won't state here, the predictive exercise we are doing here can be given causal interpretation.

The outcome (Y) is the realized annual growth rate of a country's wealth (Gross Domestic Product per capita).

The target regressor (D) is the initial level of the country's wealth.

The target parameter α is the speed of convergence, which measures the speed at which poor countries catch up with rich countries.

The controls (W) include measures of education levels, quality of institutions, trade openness, and political stability in the country.

The sample contains 90 countries and about 60 controls.

Thus $p \approx 60$, $n = 90$, and p/n is not small.

We expect the least squares method to provide a poor/ noisy estimate of α; we expect the method based on partialling-out with Lasso to provide a high quality estimate of α.

	Estimate	Std. Error	95% CI
OLS	-.009	.030	[-.071, .052]
Double Lasso	-.050	.014	[-.078, -.022]

As expected, least squares provides a rather noisy estimate of the speed of convergence, and does not allow us to answer the question about the convergence hypothesis.

In sharp contrast, double Lasso provides a more precise estimate.

The lasso based point estimate is -5% and the 95% confidence interval for the (annual) rate of convergence is -7.8% to -2.2%.

This empirical evidence does support the convergence hypothesis.

2.2 Why does Partialling-out work: Neyman Orthogonality

In the double lasso approach, α is the target parameter and η are the nuisance projection parameters with true value

$$\eta^o = (\gamma'_{DW}, \gamma'_{YW})'.$$

Note that the double lasso exploits the empirical analogue of the following moment condition for estimating α:

$$g(a, \eta) = E[\{\tilde{Y}(\eta_1) - a\tilde{D}(\eta_2)\}\tilde{D}(\eta_2)] = 0.$$

Here the true parameter value $a = \alpha$ solves this equation when

$$\eta := (\eta'_1, \eta'_2)'^o := (\gamma'_{DW}, \gamma'_{YW})',$$

where the notation

$$\tilde{Y}(\eta_1) = Y - \eta'_1 W, \quad \tilde{D}(\eta_2) = D - \eta'_2 W,$$

emphasizes the dependence on the nuisance parameters.

Here the true residualized quantities correspond to $\eta := \eta^o$, with

$$\tilde{Y} = Y - \gamma'_{YW} W, \quad \tilde{D} = D - \gamma'_{DW} W.$$

We have that

$$\partial_\eta g(\alpha, \eta^o)$$

consists of two components;

$$\partial_{\eta_1} g(\alpha, \eta^o) = E[-W\tilde{D}(\eta_2^0)] = -E[W\tilde{D}] = 0$$

and

$$\partial_{\eta_2} g(\alpha, \eta^o) = -E[\alpha W\tilde{D}] - E[(\tilde{Y} - \alpha\tilde{D})W] = 0.$$

We will call this condition Neyman orthogonality.

This condition states that the moment condition is formulated in such a way that small perturbations in terms of biased estimation of these parameters will translate to a negligible effect on estimating the target parameter.

In high-dimensional problems we estimate the nuisance parameters through solving prediction problems, where we have to regularize in order to do these predictions well, which results in biases.

Neyman orthogonality ensures that biases do not transmit to the estimation of the target parameter, at least to the first order.

Another way to look at this property is through the formulation of the estimand α.

Here α is formulated in such a way that it is not dependent on the nuisance parameters.

That is, $\alpha(\eta)$ solving $g(\alpha(\eta), \eta) = 0$,is first-order insensitive to local perturbations of these parameters:

$$\mathrm{D} = \partial_\eta \alpha(\eta^o) = 0,$$

which is tied to Neyman orthogonality condition via the implicit function theorem that gives

$$\mathrm{D} = -\partial_a g(\alpha, \eta^o)^{-1} \partial_\eta g(\alpha, \eta^o),$$

We summarize the discussion as follows:

Neyman Orthogonality. If the parameter α is defined as a root in a of the equation $g(a, \eta) = 0$, which depends on the nuisance parameters η with true value η^o, then the equation is Neyman orthogonal if

$$\partial_\eta g(\alpha, \eta^o) = 0.$$

The principle is applicable to other problems.

2.3 What happens if we don't have Neyman Orthogonality

If we don't have Neyman orthogonality, we should not expect to get high-quality estimates of the target parameters.

For example, a standard approach that is still widely used in statistical inference is as follows.

(Invalid) Single Selection/Naive Method.

In this invalid method one applies Lasso regression of Y on D and W to select relevant covariates W_Y, in addition to the covariate of interest, then refit the model by least squares of Y on D and W_Y, and carry out conventional inference.

This is post Lasso estimation except we always include D in the second regression.

Despite its wide use and proliferation, this is not a valid approach to perform inference on α, as it can result in very misleading conclusions, as we demonstrate below.

It is a fine approach to use if our goal is solely a prediction problem, and not inference on α.

This approach relies on the moment condition

$$g(a, b) = 0, \; g(a, b) = E[D(Y - Da - X'b)].$$

which is also satisfied by the true value $a = \alpha$ when $b = \beta$; this is the classical moment condition.

We don't have Neyman Orthogonality here, since

$$\partial_b g(\alpha, \beta) = -E[DX] \neq 0,$$

unless D is orthogonal to X.

For Lasso the bias will be of size $\sqrt{\log(p)/n}$, which goes to zero slower than $1/\sqrt{n}$.

Consequently, this estimator is not root-consistent; in fact $\sqrt{n}(\hat{\alpha} - \alpha)$ has a bias constant that goes to ∞ or $-\infty$ as n grows; see online Appendix to "Locally Robust Semiparametric Estimation," EMA 2022.

Consequently, while the procedure provides an estimator of α that will approach the true value in large samples, but at a slower than $\sqrt{n}$-rate, the biases in the estimator make it a poor estimator and preclude the use of standard inference.

We can set up a simulation experiment to verify that this approach would get us a low quality estimate for α.

We compare the performance of the naive and orthogonal methods in a computational experiment, where

$$p = n = 100,$$

where $\beta_j = 1/j^2$ and $(\gamma_{DW})_j = 1/j^2$, and where

$$Y = 1 \cdot D + \beta' W + \varepsilon_Y, \quad W \sim N(0, I), \quad \varepsilon_Y \sim N(0, 1)$$

$$D = \gamma'_{DW} W + \tilde{D}, \quad \tilde{D} \sim N(0, 1)/4.$$

for this estimation strategy, and the orthogonal estimator, based on partialling out, is approximately unbiased and approximately Gaussian.

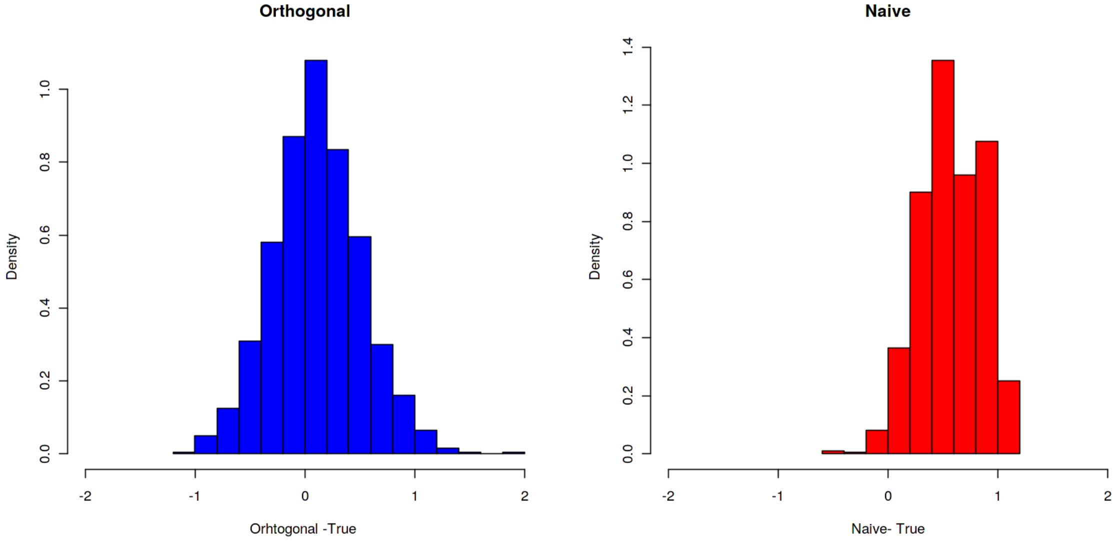

Figure 12.1: Left Panel: Simulated distribution of the centered orthogonal estimator minus the true value. **Right Panel:** Simulated distribution of the naive (single-selection) non-orthogonal estimator minus the true value

The reason, that the naive estimator does not perform well, is that it only selects controls W_j' s that are strong predictors of outcome, thereby omitting weak predictors, but the very same predictors could be strong predictors of D, and dropping these controls results in a strong omitted variable bias. In contrast, the orthogonal approach solves two prediction problems: one to predict Y and another predict D – and finds controls that are relevant for either. The resulting residuals are therefore approximately “de-counfounded".

12.3 General Debiased GMM

The general construction relies on a method-of-moments estimator for some low-dimensional target parameter θ_0 based upon the empirical analog of the moment condition

$$\mathrm{E}g(W;\theta_0,\eta_0) = 0, \tag{12.4}$$

where we call g the score function, the W denotes a data vector, θ_0 denotes the true value of a low-dimensional parameter of interest, and η denotes parameters or (nuisance functions, more generally) with true value η_0, estimated by penalized regression such as lasso.

From the histograms the naive estimator is heavily biased, as expected from the failure of the Neyman orthogonality for this estimation strategy.

In contrast, the orthogonal estimator, based on partialling out, is approximately unbiased and approximately Gaussian.

The reason, that the naive estimator does not perform well, is that it only selects controls W_j' s that are strong predictors of outcome, thereby omitting weak predictors, but the very same predictors could be strong predictors of D, and dropping these controls results in a strong omitted variable bias.

In contrast, the orthogonal approach solves two prediction problems: one to predict Y and another predict D – and finds controls that are relevant for either.

The resulting residuals are therefore approximately “de-confounded".

3 General Debiased GMM

The general construction relies on a method-of-moments estimator for some low-dimensional target parameter θ_0 based upon the empirical analog of the moment condition

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

where we call g the score function, the W denotes a data vector, θ_0 denotes the true value of a low-dimensional parameter of interest, and η denotes parameters or (nuisance functions, more generally) with true value η_0, estimated by penalized regression such as lasso.

The key requirement of a score function $g(W; \theta, \eta)$ such that

$$g(\theta, \eta) = E[g(W; \theta, \eta)]$$

identifies θ_0 when $\eta = \eta_0$, namely

$$g(\theta, \eta_0) = 0 \text{ if and only if } \theta = \theta_0,$$

is that the Neyman orthogonality condition is satisfied:

$$\partial_\eta g(\theta_0, \eta)|_{\eta=\eta_0} = 0.$$

$$\partial_\eta g(\theta_0, \eta)|_{\eta=\eta_0} = 0.$$

The Neyman orthogonality condition ensures that the moment condition used to identify and estimate θ_0 is insensitive to small perturbations of the nuisance function η around η_0.

[The name] The orthogonality condition is named after Neyman (1959), because he was the first to propose it in the context of parametric models with nuisance parameters that are estimated.

Using a Neyman-orthogonal score eliminates the first order biases arising from the replacement of η_0 with a regularized estimator $\hat{\eta}$.

Eliminating this bias is important because estimators $\hat{\eta}$ must be heavily regularized in high dimensional settings, and so these estimators will be biased in general.

The Neyman orthogonality property is responsible for the adaptivity of these estimators – namely, their approximate distribution will not depend on the fact that the estimate $\hat{\eta}$ contains error, if the latter is mild.

We focused so far on lasso methods; the inferential methods outlined below apply much more generally.

Another key input is to use a form of sample splitting at the stage of producing the estimator of the main parameter θ_0, which allows us to avoid *biases* arising from overfitting.

With proper choices of tuning parameters sample splitting is not needed for lasso, but we nonetheless give a general algorithm that is robust enough to be used with other modern high-dimensional regression methods.

3.1 The Debiased Inference Method

We assume that we have a sample $(W_i)_{i=1}^n$, modeled as i.i.d. copies of data vector W, whose law is determined by the probability measure P.

Let $\mathbb{E}_n$ and $\mathbb{V}_n$ denote the empirical expectation and variance:

$$\begin{aligned} \mathbb{E}_n[g(W_i)] \ &: \ = \frac{1}{n}\sum_{i=1}^{n} g(W_i), \\ \mathbb{V}_n[g(W_i)] \ &: \ = \mathbb{E}_n g(W_i) g(W_i)' - \mathbb{E}_n[g(W_i)]\mathbb{E}_n[g(W_i)]'. \end{aligned}$$

Generic Debiased GMM

1. **Inputs:** Provide data frame $(W_i)_{i=1}^n$, the Neyman-orthogonal score/moment function $g(W, \theta, \eta)$, which identifies the statistical parameter of interest, and the name and model for ML estimation method(s) for η.

2. **Obtain Nuisance Parameter Estimates on Folds:** Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $\{1, ..., n\}$ such that the size of each fold is about the same; for each $k \in \{1, \ldots, K\}$, construct a high-quality machine learning estimator $\hat{\eta}_{[k]}$ that depends only on subset of data $(X_i)_{i \notin I_k}$ that excludes the k-th fold.

3. **Estimate Moments:** Letting $k(i) = \{k : i \in I_k\}$, construct the moment estimate

$$\hat{g}(\theta) = \mathbb{E}_n[g(W_i; \theta, \hat{\eta}_{[k(i)]})]$$

4. **Compute Estimator:** Set the estimator $\hat{\theta}$ as the minimizer of

$$\hat{g}(\theta)' \hat{A} \hat{g}(\theta), \qquad (4)$$

where $\hat{A}$ is the weighting matrix consistent for a positive definite matrix A.

The cross-fitting eliminates overfitting bias terms like the many instrument bias of 2SLS.

Cross-fitting is essential for some first step machine learners, like overfit neural nets that can have zero residuals in sample.

The cross-fitting also makes it possible to allow for $\hat{\eta}$ that are only known to satisfy relatively weak convergence conditions.

Machine learners where only weak convergence conditions are known include neural nets and random forests.

Cross-fitting also leads to large simplifcations in regularity conditions.

4 Linear IV Models with Many Controls

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 where the variables with tilde are residuals from taking out the linear effect of W

$$\tilde{D} = D - \gamma_{DW}' W, \quad \gamma_{DW} = \arg\min_{\gamma \in \mathbb{R}^p} E[(D - \gamma' W)^2],$$

$$\tilde{Y} = Y - \gamma_{YW}' W, \quad \gamma_{YW} = \arg\min_{\gamma \in \mathbb{R}^p} \arg\min_{\gamma \in \mathbb{R}^p} E[(Y - \gamma' W)^2],$$

$$\tilde{Z} = Z - \gamma_{ZW}' W, \quad \gamma_{ZW} = \arg\min_{\gamma \in \mathbb{R}^p} E[(Z - \gamma' W)^2].$$

Here we take Z and D to be scalars, with obvious notational extensions to the case where both are vectors.

We are in the settings when the number of controls W is not small compared to the sample size.

Therefore we can impose approximate sparsity assumptions analogous to those in the first section:

$$|\gamma_{VW}|_{(j)} \leq Aj^{-a}, \quad a > 1, \quad j = 1, \ldots, p,$$

for each $V \in \{Y, D, Z\}$.

Under this structure we can estimate the projection parameters by Lasso-based methods, and apply the DGMM algorithm with the score function: with the score

$$g(W; \theta, \eta) := (Y - \gamma_1' W - \theta(D - \gamma_2' W))(Z - \gamma_3' W), \tag{5}$$

where $W = (Y, D, X, Z)$ and $\eta = (\gamma_1', \gamma_2', \gamma_3')'$ and $\eta_0 = (\gamma_{DW}', \gamma_{YW}', \gamma_{ZW}')'$.

The properties of the resulting Debiased IV estimator follow from the properties of DGMM estimator.

4.1 The Effect of Institutions on Economic Growth Revisited

This is for data from Acemoglu, Johnson, Robinson (2001, AER).

Here we consider a much richer specification for geography controls using a polynomial in latitude interacted with continent dummies.

The standard IV approach in this case gives extremely wide confidence bands, failing to provide support for the base findings in Acemoglu-Johnson-Robinson.

We can apply the approach outlined above, enabling a systematic (non-ad-hoc) way of finding sensible controls via the "double lasso" approach.

More precisely, we are using lasso to estimate the residualized variables more precisely in the setting where p is not small compared to n ($p = 16, n = 62$).

Application of the DGMM approach using the score function above gives the point estimate

$$[.77 \pm 2 \cdot .17] = [.43, 1.14].$$

Regression of residualzied D on residualized Z reveals a potential weak instrument problem, so we go ahead an implement the weak-identification robust inference, of Anderson-Rubin type.

The resulting robust confidence region is:

$$[.44, 1.74].$$

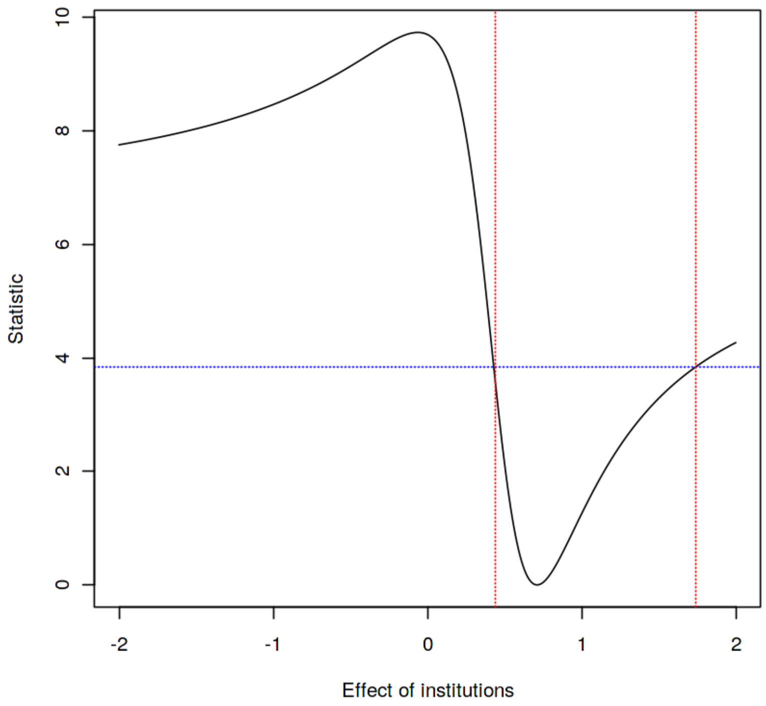

Figure 12.2: Construction of weak IV robust confidence regions for the effect of institutions on output using DGMM. Values of the $C(\theta)$ statistic are shown on the vertical axis; values of θ tested on the horizontal axis. The 90% confidence region is given by the red vertical bars.

Regression of residualzied D on residualized Z reveals a potential weak instrument problem, so we go ahead an implement the weak-identification robust inference, of Anderson-Rubin type. The resulting robust confidence region is:

$$[.44, 1.74].$$

The results on the effect of institutions on growth appear to hold up.

Code

For the code for the first empirical example, see this link. For the code for the second empirical example, see this link.

Notes

For a detailed literature review and technical regularity conditions needed for each of theorems, see [128], which also gives an overview of various analytical methods for generating Neyman-orthogonal scores in a wide variety of problems.