Review of double ML for casual parameter

by V Chernozhukov - 2016

Ding Chih, Lin Yi Ting, Lin

National Taipei University, Dept of Economics

April, 2020

Introduction

Main goal:
Provide a framework for estimating and doing inference about low-dimensional parameter (θ_0) in the presence of high-dimensional nuisance parameter (η_0) which may be estimated using ML method, such as

- lasso
- ridge
- random forest
- boosting method
- NN method

國立臺北大學
National Taipei University

Pros and cons of ML method

- Pros: remarkably efficient in prediction
- Cons: performance of inference can be poor

Good performance on prediction of y does not necessarily translate into good performance of estimation or inference of parameter θ.

Method developed in this paper

Two critical ingradients:

1. Doing Neyman-orthogonal moment / score
2. Sample-splitting and cross-fitting

Construct high-quality (unbiased) estimator of treatment parameter, θ_0.
Then, we can use DML estimator to make statistical inference

國立臺北大學
National Taipei University

Partially linear model

Partially linear model (PLR):

$$Y = D\theta_0 + g_0(\mathrm{X}) + U,\ E[U|\mathrm{X}, D] = 0$$

$$D = m_0(\mathrm{X}) + V,\ E[V|\mathrm{X}] = 0$$

- Y: outcome variable
- D: treatment variable of interest
- $\mathrm{X} = (\mathrm{X}_1, ..., \mathrm{X}_p)$: high-dimensional confounding factor
- θ_0 is main/target parameter
- m_0 and g_0 are nuisance parameter in infinite-dimensional space

國立臺北大學
National Taipei University

Partially linear model (cont.)

- $m_0 = 0$: in experimental studies
- $m_0 \neq 0$: in observational studies, which is typically the case in econometrics

Naive approach is BAD in PLR

Alternating minimization approach:

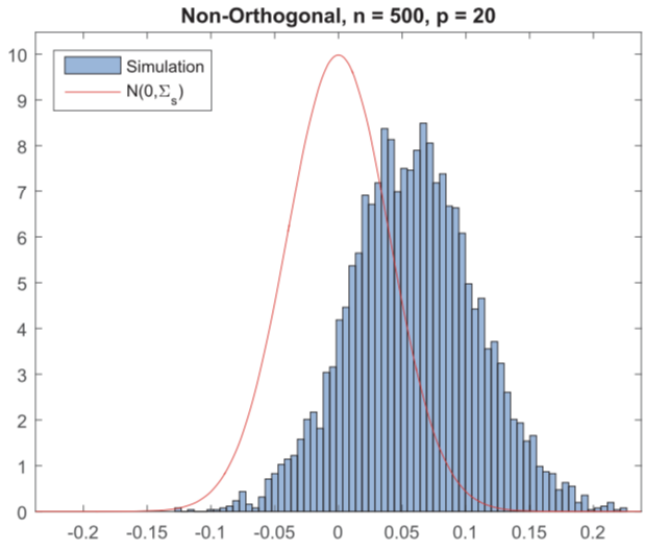
1. set initial value $\hat{\theta}_0^{(0)}$
2. build gradient boosting of $Y - D\theta_0$ on $X \implies \hat{g}_0^{(1)}$
3. build regression model of $Y - \hat{g}_0^{(1)}$ on $D \implies \hat{\theta}_{0,OLS}^{(1)}$
4. Repeat until the convergence of $|\hat{\theta}_{0,OLS}^{(i)} - \hat{\theta}_{0,OLS}^{(i-1)}| \implies \hat{\theta}_0$

國立臺北大學
National Taipei University

Naive approach is BAD in PLR (cont.)

Naive plugging estimator of $\hat{\theta}_0$

- excellent performance in prediction context
- fails to be a root-N estimator, i.e. $\hat{\theta}_0 \xrightarrow{p} \theta_0$ but $\sqrt{n}(\hat{\theta}_0 - \theta_0) \xrightarrow{p} \infty$
- regularized bias is the culprit!!!



國立臺北大學
National Taipei University

Scaled estimation error explodes in PLR

Suppose $\hat{g}_0$ is obtained using auxiliary sample and that, given this $\hat{g}_0$, the final estimate of θ_0 is obtained using the main sample and least square estimate:

$$\hat{\theta}_0 = (\frac{1}{n}\sum_{i\in I} D_i^2)^{-1}\frac{1}{n}\sum_{i\in I} D_i(Y_i - \hat{g}_0(X_i))$$

Scaled estimation error explodes in PLR conti.

Decompose the scaled estimation error in $\hat{\theta}_0$:

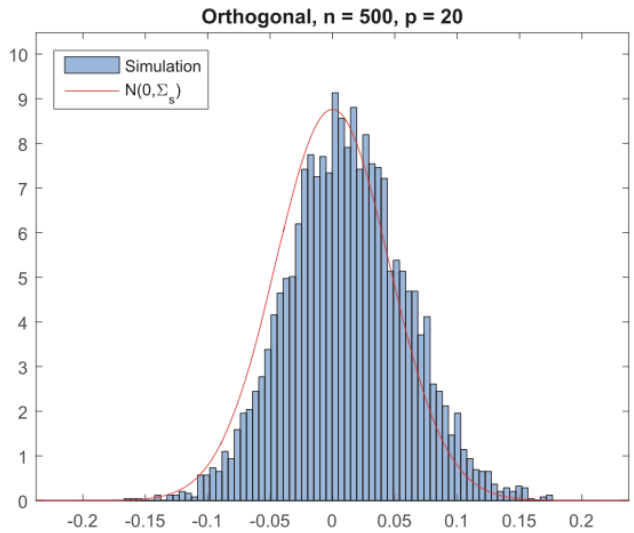
$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{(\frac{1}{n}\sum_{i\in I} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{:=\text{a}} + \underbrace{(\frac{1}{n}\sum_{i\in I} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}_0(X_i))}_{:=\text{b}}$$

- term a is normally distributed
- **Since $E[\hat{g}_0] \neq g_0$, term b tends to explode.**

Double ML approach is GOOD in PLR

1. Predict Y and D using X obtained by ML method
2. Get residual $\widehat{U} = Y - \widehat{Y}$ and $\widehat{V} = D - \widehat{D}$
3. Build regression model of $\widehat{U}$ on $\widehat{V} \implies \breve{\theta}_0$

The "filter-out" concept is based on Frisch-Waugh-Lovell therom.



國立臺北大學
National Taipei University

Scaled estimation error is bounded

With both orthogonalized estimator $\hat{V}$ and preliminary ML estimator $\hat{g}_0$ from the auxiliary sample, we can formulate the least square debiased ML (DML) estimator for θ_0 using main sample:

$$\breve{\theta}_0 = (\frac{1}{n}\sum_{i \in I} \hat{V}_i D_i)^{-1} \frac{1}{n}\sum_{i \in I} \hat{V}_i (Y_i - \hat{g}_0(X_i))$$

Scaled estimation is bounded cont.

By the same token, decompose the scaled estimation error in $\breve{\theta}_0$

$$\sqrt{n}(\breve{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

- $a^* = (E[V^2])^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} V_i U_i \xrightarrow{d} N(0, \Sigma)$
- $b^* = (E[V^2])^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} (\hat{m}_0(X_i) - m_0(X_i))(\hat{g}_0(X_i) - g_0(X_i))$
- $c^* = \frac{1}{\sqrt{n}} \sum_{i \in I} V_i (\hat{g}_0(X_i) - g_0(X_i)) = o_p(1)$

Properties of DML estimator of $\breve{\theta}_0$

- Rely on the sample-splitting to make the c^* to be $o_p(1)$
- DML estimator is now root-N consistent and approximately normally distributed

Moment conditions in PRL model

$$E[\{Y - D\theta_0 - g_0(X)\}D] = 0 \qquad \text{(1) Regression setting}$$

$$E[\{Y - D\theta_0 - g_0(X)\}\{D - m_0(X)\}] = 0 \qquad \text{(2) Frisch-Waugh-Lovell's style}$$

Use ML estimator of g_0 and m_0 as a plug-in to solve empirical analog of the moment condition, then get the estimation of θ_0.

國立臺北大學
National Taipei University

Difference between (1) and (2)

We can easily figure out the construction of moment condition in (2) from Frisch-Waugh-Lovell's style and moment condition in (1) by the concept of inner product space and orthogonality.

The key difference between these two moment condition is that (2) satisfies Neyman orthogonality condition.

Surprising properties of moment condition in (2)

Score function (moment equation):

$$\psi(W;\theta_0,\eta_0) = \{Y - D\theta_0 - g_0(X)\}\{D - m_0(X)\}$$

If score ψ satisfies

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

- ψ is an Neyman orthogonal score function
- refer this property as Neyman orthogonality
- moment condition used to identify θ_0 are locally insensitive to the value of η

The DML estimator $\breve{\theta}_0$ solves

$$\frac{1}{n}\sum_{i\in I}\psi(W;\breve{\theta}_0,\hat{\eta}_0) = 0$$

國立臺北大學
National Taipei University

transformation of moment condition

The main problem here is that how can we transform the original moment condition into Neyman orthogonal moment condition?

i.e. transform moment condition defined in (1) into (2) in PLR model

Discussed in detail later :)

Generalize PLR to moment condition model

Moment conditions model:

$$E[\psi_j(W, \theta_0, \eta_0)] = 0, \quad j = 1, 2, ..., d_\theta$$

- $\psi = (\psi, ...\psi_{d_\theta})'$ is a vector of known score function
- W is a random variable and $(W)_{i=1}^N$ is a realization set from it
- θ_0 is low-dimensional parameter of interest
- η_0 is high-dimensional nuisance parameter

Generalize PLR to moment condition model (cont.)

In ordinal linear regression setting, the score function is

$$\psi_j(W, \theta_0, \eta_0) = x_j(y - X\beta), \quad j = 1, 2, ..., k$$

In partially linear regression setting, the score function is

$$\psi_j(W, \theta_0, \eta_0) = D\{Y - D\theta_0 - g_0(X)\}$$

The score ψ_j is not unique given a data generating process.

國立臺北大學
National Taipei University

Ingredient I: Neyman orthogonality condition

We require the score ψ to be Neyman orthogonality condition in order to make $\hat{\theta}_0$ a $\sqrt{N}$-estimator

Ingredient I: Neyman orthogonality condition (conti.)

For $\widetilde{T} = \eta - \eta_0 \ : \ \eta \in T$, define pathwise derivative map D_r: $\widetilde{T} \to \mathbb{R}^{\mathrm{d}_\theta}$:

$$D_r[\eta - \eta_0] := \partial_r E[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))], \quad \eta \in T, \ \ \forall r \in [0, 1)$$

In addition, denote

$$\partial_\eta E[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] := D_0[\eta - \eta_0], \quad \eta \in T$$

Ingredient I: Neyman orthogonality condition (conti.)

Definition 1 (NEYMAN ORTHOGONALITY):

$\psi = (\psi_1, ..., \psi_{d_\theta})'$ *obeys the orthogonality condition at* (θ_0, η_0) *if*

$$\partial_\eta E[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0, \quad \forall \eta \in T_N$$

- Small deviation in nuisance function do not invalidate moment conditions.
- Practically speaking, it will be helpful to use approximate condition called *NEYMAN NEAR-ORTHOGONALITY*.

國立臺北大學
National Taipei University

Estimating equations in parametric likelihood setting

Describe the construction of orthogonal score in maximum likelihood setting.

- $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$: target parameter
- $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$: nuisance parameter
- Assume (θ_0, β_0) solves optimization problem

$$\max_{\theta \in \Theta,\, \beta \in \mathcal{B}} E[\ell(W; \theta, \beta)]$$

- $\ell(W; \theta, \beta)$: quasi-log-likelihood function

Estimating equations in parametric likelihood setting (conti.)

Under regularity conditions, (θ_0, β_0) satisfy

- $\partial_\theta E[\ell(W, \theta_0, \beta_0)] = E[\partial_\theta \ell(W; \theta_0, \beta_0)] = 0$
- $\partial_\beta E[\ell(W, \theta_0, \beta_0)] = E[\partial_\beta \ell(W; \theta_0, \beta_0)] = 0$

Generally, original score function $\varphi(W; \theta, \beta) = \partial_\theta \ell(W, \theta_0, \beta_0)$ doesn't satisfy orthogonality condition.

國立臺北大學
National Taipei University

Orthogonal estimating equations in parametric likelihood setting

Consider new score function called Neyman orthogonal score:

$$\psi(W;\theta,\eta) = \partial_\theta \ell(W;\theta_0,\beta_0) - \mu \partial_\beta \ell(W;\theta_0,\beta_0)$$

- correct the original score φ
- nuisance parameter $\eta = (\beta', vec(\mu)')'$
- μ is the $d_\theta x d_\beta$ orthogonalization parameter matrix
  - true value μ_0 solves the equation $J_{\theta\beta} - \mu J_{\beta\beta} = 0 \implies \mu_0 = J_{\theta\beta} J_{\beta\beta}^{-1}$

    information matrix $J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \sigma^2(\varphi)$

- $E[\varphi(W;\theta_0,\eta_0)] = 0$, where $\eta_0 = (\beta_0', vec(\mu_0)')'$
- φ satisfy Neyman orthogonality condition.

國立臺北大學 National Taipei University

Parametric example: High-dimensional linear regression

Consider high-dimensional linear regression model:

$$Y = D\theta_0 + X'\beta_0 + U, \quad E[U(X', D)'] = 0$$

$$D = X'\gamma_0 + V, \quad E[VX] = 0$$

- (θ_0, β_0) solves optimization problem

$$\max_{\theta\in\Theta,\, \beta\in\mathcal{B}} E[\ell(W;\theta,\beta)] = \max_{\theta\in\Theta,\, \beta\in\mathcal{B}} E[-\frac{1}{2}(Y - D\theta - X'\beta)^2]$$

- $E[\partial_\theta \ell(W;\theta_0,\beta_0)] = 0 \implies E[(Y - D\theta - X'\beta)D] = 0$
- $E[\partial_\beta \ell(W;\theta_0,\beta_0)] = 0 \implies E[(Y - D\theta - X'\beta)X] = 0$

國立臺北大學
National Taipei University

Parametric example: High-dimensional linear regression (conti.)

Neyman orthogonal score is given by

$$\psi(W;\theta,\eta) = (Y - D\theta - X'\beta)(D - \mu X)$$

- $\eta = (\beta', vec(\mu)')'$
- $\psi(W;\theta_0,\eta_0) = U(D - \mu_0 X) \implies E[U(D - \mu_0 X)] = 0$
- $\mu_0 = E[DX'](E[XX'])^{-1} = \gamma_0'$

Estimation of β_0, $\gamma = \mu_0'$ and $\mu_0 = (\beta_0', vec(\mu_0)')'$

- Penalized least square
- OCMT

Extend the result

Extend the results obtained above:

- Concentrating-out approach used in Neywey (1994)
- Neyman orthogoanl scores for likelihood problems with infinite-dimensional nuisance paramter
- Neyman orthogoanl scores for GMM setting

Estimating equations in conditional moment restriction problems

Conditional moment restrictions framework assumes $(\theta_0,\ h_0)$ satisfy

$$E[m(W;\theta_0, h_0(Z) \mid R)] = 0$$

- $W \subset \mathbb{R}^{\mathrm{d_w}}$, $\mathrm{R} \subset \mathbb{R}^{\mathrm{d_r}}$ and $Z \subset \mathbb{R}^{\mathrm{d_z}}$: *random vector*
- $\theta \in \Theta \subset \mathbb{R}^{\mathrm{d}_\theta}$: *finite-dimensional parameter*
- h: *vector-valued functional nuisance parameter*
- $m : W\ x\ \Theta\ x\ \mathbb{R}^{\mathrm{d_h}}$: *known vector-valued function*

It covers a rich variety of models without relying on likelihood setting.

國立臺北大學
National Taipei University

Orthogonal estimating equations with conditional moment restrictions

Neyman orthogonal score:

$$\psi(W;\theta,\eta) = \mu(R)m(W;\theta,h(Z))$$

- matrix-valued functional parameter $\mu : R \to \mathbb{R}^{\mathrm{d}_\theta \mathrm{x} \mathrm{d}_\mathrm{m}}$

$$\mu_0(R) = A(R)'\Omega(R)^{-1} - G(Z)\Gamma(R)'\Omega(R)^{-1}$$

  - A: moment selection matrix-valued function
  - Ω: Weighting positive definite matrix-valued function
  - $\Gamma(R)$, $G(R)$: defined explicitly

- nuisance parameter $\eta = (\mu, h)$

國立臺北大學
National Taipei University

Semi-parametric example: PLR

Neyman orthogonal score in PLR:

- Previous: Frisch-Waugh-Lovell's style moment condition
- Now: using this framework

Semi-parametric example: PLR (conti.)

Original conditional moment condition:

$$E[m(W;\theta_0, h_0(Z) \mid R)] = E[Y - D\theta_0 - g_0(X) \mid X,\ D] = 0$$

Neyman orthogonal score:

$$\psi(W;\theta,\eta_0) = (D - m_0(X))x(Y - D\theta - g_0(X))$$

Solve the sample analog of this score, $\breve{\theta}$ is the DML estimator!

Ingredient II: Sample splitting

Sample splitting:

- $\{1, ..., N\}$: index set of all observations
- I main sample: estimate θ_0
- I^c auxiliary sample: estimate η_0

Simplify the analysis - issue about convergence rate

DML and Properties

< **DML1** >

1. K-fold random partition on the sample W
2. for each k, build ML estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$$

3. for each k, construct an estimator for $\check{\theta}_{0,k}$, which solves:

$$\mathbb{E}_{n,k}\{\psi(W; \check{\theta}_{0,k}; \hat{\eta}_{0,k})\} = n^{-1} \sum_{i \in I_k} \psi(W_i) = 0$$

4. Aggregated estimator of θ_0:

$$\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^{K} \check{\theta}_{0,k}$$

DML and Properties cont.

< DML2 >

1. K-fold random partition on the sample W
2. for each k, build an ML estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$$

3. construct the estimator $\tilde{\theta}_0$ as the solution of:

$$\frac{1}{K} \sum_{k=1}^{K} E_{n,k}\{\psi(W; \tilde{\theta}_0; \hat{\eta}_{0,k})\} = 0$$

Preference

Prefer DML2 over DML1, pooled jacobian is more stable than the separate jacobian of DML1

Moment Condition Models with Linear Scores

$$\psi(w, \theta; \eta) = \psi^a(w; \eta)\theta + \psi^b(w; \eta) \quad \text{for all } w \in \mathcal{W}, \quad \theta \in \Theta, \quad \eta \in \mathbf{T}$$

< Assumptions >

1. The score ψ fits the Neyman near -orthogonality conditions
2. estimator of η_0 belongs to the realization set $\mathcal{T}_N \subset T$
3. Variance of the score ψ is non-degenerate

國立臺北大學
National Taipei University

Moment Condition Models with Linear Scores cont.

< Properties >

1. The estimator reaches root-N rate of convergence with P varying over an expanding class of probability measures $\mathcal{P}_N$

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0, I_d)$$

2. Either styles of DML, the estimator concentrates around the true variance matrix σ^2

$$\hat{\sigma}^2 = \sigma^2 + O_p(\varrho_N)$$

國立臺北大學
National Taipei University

Models with Non-linear Scores

< Assumptions >

Assumptions of non-linear case are similar to linear cases.
Additionally, the function class

$$\mathcal{F}_{1,\eta} = \{\psi_j(\cdot, \theta, \eta) : j = 1, \cdots, d_\theta, \theta \in \Theta\}$$

is suitably measurable.

< Properties >

Root-N convergence and variance of estimator $\tilde{\theta}_0$ also holds.

國立臺北大學
National Taipei University

Fixing the Impact of Sample Splitting

Splitting auxiliary sample asymptotically no impact.
Finite samples might cause $\tilde{theta}_0$ to vacillate.

< Two methods >

Repeat the DML estimation of $\tilde{\theta}_0$ s times

$$\tilde{\theta}_0^{mean} = \frac{1}{S}\sum_{s=1}^{S}\tilde{\theta}_0^s \quad \text{or} \quad \tilde{\theta}_0^{median} = median\{\tilde{\theta}_0^s\}_{s=1}^{S}$$

Median method is favored, as $\tilde{\theta}_0^{median}$ and $\hat{\sigma}^{2,median}$ is more robust to outliers.

Partially Linear Regression

$$Y = D\theta_0 + g_0(X) + U, \quad E_P[U|X, D] = 0$$

$$D = m_0(X) + V, \quad E_P[V|X] = 0$$

< Two score functions approach >

Conventional

$\psi(W; \theta; \eta) := \{Y - D\theta - g(X)\}(D - m(X)), \quad \eta = (g, m)$

Robinson Style

$\psi(W; \theta; \eta) := \{Y - l(x) - \theta(D - m(X))\}(D - m(X)), \quad \eta = (l, m)$

國立臺北大學
National Taipei University

Contrast of the Scores

Ordinary scores and Robinson style scores are first-order equivalent

Both θ_0 satisfies moment condition

$$E_P \psi(W; \theta_0; \eta_0) = 0$$

Both satisfies orthogonality conditions:

$$\partial_\eta E_P \psi(W; \theta_0; \eta_0)[\eta - \eta_0] = 0$$

where

Ordinary Score Function $\eta_0 = (g_0, m_0)$

Robinson Style $\eta_0 = (l_0, m_0)$

國立臺北大學
National Taipei University

Regularity Conditions Assumption

Let P be the collection of probability laws P for $W = (Y, D, X)$

let c, C, q be fixed positive constants

1. $\|Y\|_{P,q} + \|D\|_{P,q} \leq C$
2. $\|UV\|_{P,2} \geq c^2 \quad \text{and} \quad E_P[V^2] \geq c$
3. $\|E_P[U^2|X]\|_{P,\infty} \quad \text{and} \quad \|E_P[V^2|X]_{P,\infty} \leq C$

η estimators rate conditions of convergence are achievable for most ML methods and are case specific

PLR Coefficients Inference

Regardless of utilizing DML1 or DML2, and styles of score functions

estimators $\tilde{\theta}_0$ obeys:

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0,1)$$

replacing σ^2 with $\hat{\sigma}^2$, the CI of $\tilde{\theta}_0$ have uniform asymptotic validty:

$$\lim_{N\to\infty} \sup_{P\in P} |\mathrm{Pr}_p(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}]) - (1-\alpha)| = 0$$

Partially Linear IV Models

$$Y = D\theta_0 + g_0(X) + U \quad E_P[U|X, Z] = 0$$

$$Z = m_0(X) + V, \quad E_P[V|X] = 0$$

Z is the instrument variable

< Two score functions approach >

Conventional

$\psi(W; \theta; \eta) := \{Y - D\theta - g(X)\}(Z - m(X)), \quad \eta = (g, m)$

Robinson Style

$\psi(W; \theta; \eta) := \{Y - l(x) - \theta(D - r(X))\}(Z - m(X)), \quad \eta = (l, m, r)$

Contrast of Scores for IV model

Both score functions satisfies:

1. Moment Condition $E_P[\psi(W;\theta_0;\eta_0)] = 0$
2. Orthogonality Condition $\partial_\eta E_P[\psi(W;\theta_0;\eta_0)][\eta - \eta_0] = 0$

where:

Ordinary $\eta_0 = (g_0, m_0)$

Robinson Style $\eta_0 = (l_0, m_0, r_0)$

$l_0(X) = E_p[Y|X]$ and $r_0(X) = E_p[D|X]$

國立臺北大學
National Taipei University

Regularity Conditions Assumptions for IV Model

Let P be the collection of probability laws P for $W = (Y, D, X, Z)$

let c, C, q be fixed positive constants

1. $\|Y\|_{P,q} + \|D\|_{P,q} + \|Z\|_{P,q} \leq C$
2. $\|UV\|_{P,2} \geq c^2$ and $|E_P[DV]| \geq c$
3. $\|E_P[U^2|X]\|_{P,\infty}$ and $\|E_P[V^2|X]_{P,\infty}\| \leq C$

Partial Linear IV Model Coefficient Inference

Simalar to PLR, as the IV model is an expansion case DML1 or DML2, regardless of score function styles

estimator $\tilde{\theta}_0$ are first order equivalent and obeys:

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0,1)$$

replacing σ^2 with $\hat{\sigma}^2$, the CI of $\tilde{\theta}_0$ have uniform asymptotic validty:

$$\lim_{N\to\infty} \sup_{P\in P} |\mathrm{Pr}_P(\theta_0 \in [\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}]) - (1-\alpha)| = 0$$

國立臺北大學
National Taipei University