Double Machine Learning

1 Introduction

1.1 Motivation

The goal is finding the root-N consistent estimation, where N is the sample size.[1] We adopt the semi-parametric setting, with θ_0 as the parameter of interest and η_0 as the infinite-dimensional nuisance parameter. The paper's main contribution is to offer a simple procedure for estimating and make inference on θ_0.

Consider the simple example of a partially linear regression (PLR)

$$Y = D\theta_0 + g_0(X) + U, \quad D = m_0(X) + V$$

with $E[U|X, D] = 0$, and $E[V|X] = 0$. Here, Y is the outcome variable, D is the policy variable of interest, X is a vector of other controls, and U, V are disturbances.

Notice that the second equation keeps track of confounding. The confounding factors X affect the policy variable D via the function $m_0(X)$ and the outcome variable via the function $g_0(X)$. If $dim(X) = p$ increases with N, traditional assumptions that limit the complexity of the parameter space for the nuisance parameters $\eta_0 = (m_0, g_0)$ to fail.

1.2 Regularization bias

Consider a simple approach, we estimate θ_0 by first constructing a complicated ML estimator $D\hat{\theta}_0 + \hat{g}_0(X)$ for learning the regression function. We adopt the two-part sample-splitting strategy, with each part a sample size of $n = N/2$. We index the main sample by $i \in I$, the auxiliary sample as $i \in I^c$. Suppose $\hat{g}_0$ is obtained using the **auxiliary sample**, the final estimate of θ_0 is obtained using the **main sample**:

$$\begin{aligned} \frac{\partial \frac{1}{2} \sum_{i \in I} \hat{\epsilon}_i^2}{\partial \theta_0} &= 0 \\ \sum_{i \in I} (y_i - D_i\hat{\theta}_0 - \hat{g}_0(X_i))D_i &= 0 \\ \sum_{i \in I} D_i(Y_i - \hat{g}_0(X_i)) - \sum_{i \in I} D_i^2\hat{\theta}_0 &= 0 \end{aligned}$$

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

In general, $\hat{\theta}_0$ does not have the root-N consistency, namely,

$$|\sqrt{n}(\hat{\theta}_0 - \theta_0)| \xrightarrow{p} \infty$$

Heuristically, this is due to the bias in learning g_0. Observe that we can plug in $Y_i = D_i\theta_0 + g_0(X_i) + U_i$ into $\hat{\theta}_0$ and obtain:

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i + \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}_0(X_i))$$

With some regularity conditions, the first part of the decomposition will converge in distribution to normal. The second part (call it b) is the **regularization bias** term. Indeed, we have

$$b = (E[D_i^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i)) + o_p(1)$$

to the first order. Notice that b is the sum of n terms that do not have mean zero $m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i))$, divided by $\sqrt{n}$. These terms have non-zero mean because, in high-dimensional settings, we must **employ regularized estimators** – such as lasso, ridge, boosting or penalized neural nets – for informative learning to be feasible. We need the regularization terms to **keep the variance from exploding** but due to the bias-variance trade-off, we induce substantive biases in the estimator $\hat{g}_0$ of g_0.

Specifically, the rate of convergence of the bias in the root-mean-squared error sense will typically be $n^{-\varphi_g}$ with $\varphi_g < 1/2$. Hence, we expect b to be of stochastic order $\sqrt{n}n^{-\varphi_g} \to \infty$ as D_i is centered at $m_0(X_i) \neq 0$, so we can't get the root-N consistency

1.3 Using orthogonalization to overcome regularization biases

Now, consider the second construction by directly partialling out the effect of X from D to obtain the orthogonalized regressor $V = D - m_0(X)$. That means we get $\hat{V} = D - \hat{m}_0(X)$ using ML and the auxiliary samples.

We can now formulate the DML estimator for θ_0 using the main sample of observations:

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

1.3.1 Derivation

- **One approach**

$$\begin{aligned} Y - g_0(X) &= D\theta_0 + U \\ \hat{V}(Y - g_0(X)) &= \hat{V}(D\theta) + VU \\ E[\hat{V}(Y - g_0(X))] &= E[\hat{V}D]\theta \end{aligned}$$

We can also obtain the estimator using the score function in the latter section.

- **IV and 2SLS**

  I had a hard time trying to figure out how did he come up with $\check{\theta}_0$. As hinted in the paper, it's related to IV in econometrics. So we shall probably start with IV first. [2]

  Problem: Figuring out the causal link between schooling and wages, we write the potential outcomes as:

  $$Y_{si} \equiv f_i(s)$$

  and $f_i(s) = \pi_0 + \pi_1 s + \eta_i$ Suppose there is a vector of control variables, A_i, called "ability",that:

  $$\eta_i = A_i^T + v_i$$

  s.t. v_i, A_i are uncorrelated by construction. For now, the variables A_i, are assumed to be the only reason why η_i and s_i are correlated, so that $E[s_i v_i] = 0$. In other words, we are happy to write

  $$Y_i = \alpha + \rho S_i + A_i^T \gamma + v_i$$

  The problem we want to tackle is how to estimate the coefficient ρ, when A_i is unobserved. Suppose we have an instrument Z_i that is correlated with S_i but uncorrelated with any other determinants of the dependent variable.

  Since we have $Cov(Y_i, Z_i) = Cov(\alpha + \rho S_i + A_i^T \gamma + v_i, Z_i) = \rho Cov(S_i, Z_i)$, that implies

  $$\rho = \frac{Cov(Y_i, Z_i)}{Cov(S_i, Z_i)} = \frac{Cov(Y_i, Z_i)/V(Z_i)}{Cov(S_i, Z_i)/V(Z_i)}$$

  So it can be viewed as two regression problems. Consider the first-stage and reduced-form regression equations:

  $$S_i = X_i^T \pi_{10} + \pi_{11} Z_i + \xi_{1i}$$
  $$Y_i = X_i^T \pi_{20} + \pi_{21} Z_i + \xi_{2i}$$

  π_{11} captures the first-stage effect of Z_i on S_i, adjusting for covariates, X_i. π_{21} captures the reduced-form effect of Z_i on Y_i, adjusting for these same covariates.

  Note that the denominators of the reduced form and first stage effects are the same. Hence the ratio is

  $$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{Cov(Y_i, \tilde{z}_i)}{Cov(S_i, \tilde{Z}_i)}$$

  where $\tilde{Z}_i$ is the residual from a regression of Z_i on the exogenous covariates X_i. Econometricians call the sample analog of the left hand side of the equation an Indirect Least Squares estimator of ρ in the causal model with covariates,

  $$Y_i = \alpha^T X_i + \rho S_i + \eta_i$$

  where η_i is the compound error term, $A_i^T \gamma + v_i$. We can use it to confirm directly that $Cov(Y_i, \tilde{z}_i) = \rho Cov(S_i, \tilde{z}_i)$ since $\tilde{z}_i$ is uncorrelated with X_i by construction and with η_i by assumption. (Some form of orthogonalization)

  We can plug in the first stage equation into the reduced-form equation.

  $$\begin{aligned} Y_i &= \alpha^T X_i + \rho[X_i^T \pi_{10} + \pi_{11} Z_i + \xi_{1i}] + \eta_i \\ &= X_i^T[\alpha + \rho\pi_{10}] + \rho\pi_{11} Z_i + [\rho\xi_{1i} + \eta_i] \\ &= X_i^T \pi_{20} + \pi_{21} Z_i + \xi_{2i} \end{aligned}$$

We can slightly rearrange the above equation as

$$Y_i = \alpha^T X_i + \rho[X_i^T \pi_{10} + \pi_{11} Z_i] + \xi_{2i}$$

The term in the bracket is the population fitted value from the first stage regression of S_i on X_i and Z_i. Because Z_i and X_i are uncorrelated with the reduced form error ξ_{2i}, ρ is obtained by the population regression of Y_i on X_i and $[X_i^T \pi_{10} + \pi_{11} Z_i]$

In practice, we only have samples, so we can obtain

$$\hat{S}_i = X_i^T \hat{\pi}_{10} + \hat{\pi}_{11} Z_i$$

The 2SLS estimator of ρ is obtained by

$$Y_i = \alpha^T X_i + \rho \hat{S}_i + [\eta_i + \rho(S_i - \hat{S}_i)]$$

- **Frisch-Waugh-Lovell Theorem**
In a multiple linear regression setting. Suppose $y = X\beta + \epsilon$, denote $P = X(X^T X)^{-1} X^T$, $M = I - P$, then $\hat{y} = X\hat{\beta} = Py$, while $e = y - \hat{y} = My$. Suppose we partition the design matrix by $X = (X_1, X_2)$, and $\beta = (\beta_1, \beta_2)^T$, that implies $y = X_1\beta_1 + X_2\beta_2 + \epsilon$. By the first order condition $X^T X\hat{\beta} = X^T y$ We have

$$X^T X = \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix}$$

plugging into the FOC, we obtain

$$\begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1^T y \\ X_2^T y \end{pmatrix}$$

$$X_1^T X_1 \hat{\beta}_1 + X_1^T X_2 \hat{\beta}_2 = X_1^T y$$
$$X_2^T X_1 \hat{\beta}_1 + X_2^T X_2 \hat{\beta}_2 = X_2^T y$$

From the first equation, we have

$$\hat{\beta}_1 = (X_1^T X_1)^{-1}(X_1^T y - X_1^T X_2 \hat{\beta}_2) = (X_1^T X_1)^{-1} X_1^T (y - X_2 \hat{\beta}_2)$$

If we plug the estimate into the second equation, we have

$$X_2^T X_1 (X_1^T X_1)^{-1} X_1^T (y - X_2 \hat{\beta}_2) + X_2^T X_2 \hat{\beta}_2 = X_1^T y$$

Notice that we can denote $P_1 = X_1(X_1^T X_1)^{-1} X_1^T$, so

$$X_2^T P_1 y - X_2^T P_1 X_2 \hat{\beta}_2 + X_2^T X_2 \hat{\beta}_2 = X_2^T y$$
$$(X_2^T X_2 - X_2^T P_1 X_2)\hat{\beta}_2 = X_2^T y - X_2^T P_1 y$$
$$X_2^T (I - P_1) X_2 \hat{\beta}_2 = X_2^T (I - P_1) y$$

So we finally get $\hat{\beta}_2 = (X_2^T M_1 X_2)^{-1} X_2^T M_1 y$. Notice that M_1 is an idempotent and symmetric matrix, we can rewrite $\hat{\beta}_2 = [(M_1 X_2)^T M_1 X_2]^{-1} (M_1 x_2)^T M_1 y$. $\hat{\beta}_2$ is obtained by regressing $M_1 y$ (the residual obtained by regressing y on X_1), on $M_1 X_2$ (the residual obtained by regressing X_2 on X_1)

1.3.2 Removing Regularization Bias

To illustrate the detail, we decompose the scaled estimation error as:

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

The leading term, a^*, will satisfy

$$a^* = (E[V^2])^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} V_i U_i \xrightarrow{d} N(0, \Sigma)$$

under mild conditions. The second term, b^*, captures the impact of regularization bias in estimating g_0 and m_0. Specifically, we will have

$$b^* = (E[V^2])^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} (\hat{m}_0(X_i) - m_0(X_i))(\hat{g}_0(X_i) - g_0(X_i))$$

Because this term depends only on the product of the estimation errors, it can vanish under a broad range of data-generating processes. Indeed, the term is upper bounded by $\sqrt{n} n^{-(\varphi_m + \varphi_s)}$ where n^{φ_m} and n^{φ_s} are the rates of convergence of $\hat{m}_0$ to m_0 and $\hat{g}_0$ to g_0 This upper bound can clearly vanish even though both m_0 and g_0 are estimated at relatively slow rates.

By adopting sample splitting, it allows us to guarantee that $c^* = o_p(1)$ under weak conditions.

1.4 Role of Sample Splitting

In the partially linear model, c^* contains terms such as

$$\frac{1}{\sqrt{n}} \sum_{i \in I} V_i (\hat{g}_0(X_i) - g_0(X_i))$$

which involve $1/\sqrt{n}$ normalized sums of products of structural unobservables from the PLR model. The use of sample splitting allows simple and tight control of such terms. To see this, assume that observations are independent and recall that $\hat{g}_0$ is estimated using only observations in the auxiliary sample.

Conditioning on the auxiliary sample and recalling that $E[V_i|X_i] = 0$, it is easy to verify that $\frac{1}{\sqrt{n}} \sum_{i \in I} V_i (\hat{g}_0(X_i) - g_0(X_i))$ has mean zero and variance of order

$$\frac{1}{n} \sum_{i \in I} (\hat{g}_0(X_i) - g_0(X_i))^2 \xrightarrow{p} 0$$

By Chebyshev's inequality, the term vanishes.

Using sample splitting can result in a substantial loss of efficiency, but we can flip the role of the main and auxiliary samples to obtain a second version of the estimator of the parameter of interest. By averaging the two estimators, we can regain full efficiency.

Without sample splitting, the term above might not vanish. Because V_i and $\hat{g}_0(X_i) - g_0(X_i)$ are generally related because the data for observation i are used in forming the estimator $\hat{g}_0$. The association can then lead to poor performance of an estimator of θ_0 that makes use of g_0 as a plug-in estimator for g0 even when this estimator converges at a very favorable rate, say $N^{-1/2+\epsilon}$.

1.4.1 Illustrative example of the potential overfitting problem

Let $\hat{g}_0(X_i) = g_0(X_i) + (Y_i - g_0(X_i))/N^{1/2-\epsilon}$ for any i in the sample used to form estimator $\hat{g}_0$, the second term captures overfitting of the outcome variable within the estimation sample. This estimator is excellent in terms of rates. If U_i and D_i are bounded, $\hat{g}_0$ converges uniformly to g_0 at the nearly parametric rate $N^{-1/2+\epsilon}$. But c^* explodes because

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} V_i(\hat{g}_0(X_i) - g_0(X_i)) \propto N^{\epsilon} \to \infty$$

1.5 Neyman orthogonality and moment conditions

1.5.1 Conventional Estimator

The first estimator $\hat{\theta}_0$ can be viewed as a solution to estimating equations

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

where φ is a known score function and $\hat{g}_0$ is the estimator of the nuisance parameter g_0. In a partially linear model, the score function is $\varphi(W; \theta, g) = (Y - \theta D - g(X))D$. The score function φ is sensitive to biased estimation of g. Specifically, the Gateaux derivative operator w.r.t. g does not vanish:

$$\partial_g E[\varphi(W; \theta_0, g_0)][g - g_0] \neq 0$$

1.5.2 DML

We consider $\check{\theta}_0$ as solving

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

where $\hat{\eta}_0$ is the estimator of the nuisance parameter η_0 and ψ is an orthogonalized score function that satisfies:

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

In the partially linear model, $\check{\theta}_0$ uses the score function $\psi(W; \theta, \eta) = (Y - D\theta - g(X))(D - m(X))$, with the nuisance parameter being $\eta = (m, g)$.

Question: I noticed that the above FOC will result in the DML estimator, but how did they get this FOC condition?

2 Notation

The symbols Pr and E denote probability and expectation operators with respect to a generic probability measure that describes the law of the data. To signify the importance of the measure

P, we use Pr_P and E_p. Capital letters are random variables, and lowercase letters denote the realizations.

We use $||\cdot||_{P,q}$ to denote the $L^q(P)$ norm; for example, we denote $||f||_{P,q} \equiv ||f(W)||_{P,q} \equiv (\int |f(w)|^q dP(w))^{1/q}$

3 Construction of Neyman Orthogonal Score/Moment Functions

3.1 Moment condition/estimating equation framework

We are interested in the true value θ_0 of the low-dimensional target parameter $\theta \in \Theta$ where Θ is a non-empty measurable subset of $\mathbb{R}^{d_\theta}$. We assume that θ_0 satisfies the moment conditions

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

where $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ is a vector of known score functions, W is a random element taking values in a measurable space $(\mathcal{W}, \mathcal{A}_\mathcal{W})$ with law determined by a probability measure $P \in \mathcal{P}_N$, and η_0 is the true value of the nuisance parameter $\eta \in T$, where T is a convex subset of some normed vector space with the norm denoted by $\|\cdot\|_T$. We assume that the score functions $\psi_j : \mathcal{W} \times \Theta \times T \to \mathbb{R}$ are measurable once we equip Θ and T with their Borel σ-fields, and we assume that a random sample $(W_i)_{i=1}^N$ from the distribution of W is available for estimation and inference.

To introduce the Neyman orthogonality condition, for $\tilde{T} = \{\eta - \eta_0 : \eta \in T\}$ we define the Gateaux derivative map $D_r : \tilde{T} \to \mathbb{R}^{d_\theta}$

$$D_r[\eta - \eta_0] \equiv \partial_r\{E_P[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))]\}$$

where $\eta \in T$ and for all $r \in [0, 1)$. We also denote

$$\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] \equiv D_0[\eta - \eta_0]$$

Note that $\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))$ here is well defined because for all $r \in [0, 1)$ and $\eta \in T$,

$$\eta_0 + r(\eta - \eta_0) = (1 - r)\eta_0 + r\eta \in T$$

as T is a convex set. In addition, let $\mathcal{T}_N \subset T$ be a nuisance realization set such that the estimators $\hat{\eta}_0$ of η_0 specified below take values in this set with high probability. In practice, we typically assume that $\mathcal{T}_N$ is a properly shrinking neighbourhood of η_0. Note that $\mathcal{T}_N - \eta_0$ is the nuisance deviation set, which contains deviations of $\hat{\eta}_0$ from η_0, $\hat{\eta}_0 - \eta_0$, with high probability. The Neyman orthogonality condition requires that the derivative in (2.2) vanishes for all $\eta \in \mathcal{T}_N$.

DEFINITION 2.1. (NEYMAN ORTHOGONALITY) *The score* $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ *obeys the orthogonality condition at* (θ_0, η_0) *with respect to the nuisance realization set* $\mathcal{T}_N \subset T$ *if (2.1) holds and the pathwise derivative map* $\mathrm{D}_r[\eta - \eta_0]$ *exists for all* $r \in [0, 1)$ *and* $\eta \in \mathcal{T}_N$ *and vanishes at* $r = 0$*; namely,*

$$\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0, \quad \text{for all } \eta \in \mathcal{T}_N. \tag{2.3}$$

We remark here that condition (2.3) holds with $\mathcal{T}_N = T$ when η is a finite-dimensional vector as long as $\partial_\eta E_P[\psi_j(W; \theta_0, \eta_0)] = 0$ for all $j = 1, \ldots, d_\theta$, where $\partial_\eta E_P[\psi_j(W; \theta_0, \eta_0)]$ denotes the vector of partial derivatives of the function $\eta \mapsto E_P[\psi_j(W; \theta_0, \eta)]$ for $\eta = \eta_0$.

Sometimes it will also be helpful to use an approximate Neyman orthogonality condition as opposed to the exact one given in Definition 2.1.

DEFINITION 2.2. (NEYMAN NEAR-ORTHOGONALITY) *The score* $\psi = (\psi_1, \ldots, \psi_{d_\theta})'$ *obeys the* λ_N *near-orthogonality condition at* (θ_0, η_0) *with respect to the nuisance realization set* $\mathcal{T}_N \subset T$ *if (2.1) holds and the pathwise derivative map* $\mathrm{D}_r[\eta - \eta_0]$ *exists for all* $r \in [0, 1)$ *and* $\eta \in \mathcal{T}_N$ *and is small at* $r = 0$*; namely,*

$$\|\partial_\eta E_P[\psi(W; \theta_0, \eta_0)][\eta - \eta_0]\| \leq \lambda_N, \quad \text{for all } \eta \in \mathcal{T}_N, \tag{2.4}$$

where $\{\lambda_N\}_{N \geq 1}$ *is a sequence of positive constants such that* $\lambda_N = o(N^{-1/2})$.

Notice that the Gateaux derivative of our score function with respect to our nuisance parameter is 0. That implies, when the derivative is 0, our score function is robust to small change in η_0.

3.2 Construction of Neyman orthogonal scores

3.2.1 Neyman orthogonal scores for likelihood and other M-estimation problems with finite-dimensional nuisance parameters

Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$, $\mathcal{B}$ is a convex set, θ, β are target and nuisance parameters respectively. Further, suppose the true parameter θ_0, β_0 solve the optimization problem

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

$l(W; \theta, \beta)$ can be the log-likelihood function associated with observation W. Under mild regularity conditions,

$$E_p[\partial_\theta l(W; \theta_0, \beta_0)] = 0, \quad E_P[\partial_\beta l(W; \theta_0, \beta_0)] = 0$$

Note that if we take $\phi(W; \theta, \beta) = \partial_\theta l(W; \theta, \beta)$ for estimating θ_0 will not generally satisfy the orthogonality condition. Now consider the Neyman orthogonal score,

$$\psi(W; \theta, \eta) = \partial_\theta l(W; \theta, \beta) - \mu \partial_\beta l(W; \theta, \beta)$$

where the nuisance parameter is

$$\eta = (\beta^T, vec(\mu)^T)^T \in T = \mathcal{B} \times \mathbb{R}^{d_\theta d_\beta} \subset \mathbb{R}^p, \quad p = d_\beta + d_\theta d_\beta$$

and μ is the $d_\theta \times d_\beta$ orthogonalization parameter matrix whose true value μ_0 solves the equation

$$J_{\theta\beta} - \mu J_{\beta\beta} = 0$$

for

$$J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \partial_{(\theta^T,\beta^T)} E_P[\partial_{(\theta^T,\beta^T)^T} l(W;\theta,\beta)]|_{\theta=\theta_0;\beta=\beta_0}$$

The true value of the nuisance parameter η is

$$\eta_0 = (\beta_0^T, vec(\mu_0)^T)^T$$

and when $J_{\beta\beta}$ is invertible, the equation has a unique solution,

$$\mu_0 = J_{\theta\beta} J_{\beta\beta}^{-1}$$

3.2.2 High-Dimensional Linear Regression

Consider the following model,

$$Y = D\theta_0 + X^T\beta_0 + U, \quad D = X^T\gamma_0 + V$$

where $E_P[U(X^T, D)^T] = 0$, $E_P[VX] = 0$ and θ_0 is a scalar. We know that θ_0, β_0 solve the following optimization problem

$$l(W;\theta,\beta) = -\frac{(Y - D\theta - X^T\beta)^2}{2}, \ \ \theta \in \Theta = \mathbb{R}, \ \ \beta \in \mathcal{B} = \mathbb{R}^{d_\beta}$$

where we denote $W = (Y, D, X^T)^T$. The following hold

$$E_p[\partial_\theta l(W;\theta_0,\beta_0)] = 0, \quad E_P[\partial_\beta l(W;\theta_0,\beta_0)] = 0$$

with

$$\partial l_\theta(W;\theta,\beta) = (Y - D\theta - X^T\beta)D, \quad \partial l_\beta(W;\theta,\beta) = (Y - D\theta - X^T\beta)X$$

and the matrix J satisfies

$$J_{\theta\beta} = -E_P[DX^T], \quad J_{\beta\beta} = -E_P[XX^T]$$

The Neyman orthogonal score is then given by

$$\begin{aligned} \psi(W;\theta,\eta) &= \partial_\theta l(W;\theta,\beta) - \mu\partial_\beta l(W;\theta,\beta) \\ &= (Y - D\theta - X^T\beta)D - \mu(Y - D\theta - X^T\beta)X \\ &= (Y - D\theta - X^T\beta)(D - \mu X) \end{aligned}$$

with $\eta = (\beta^T, vec(\mu)^T)^T$ Therefore, by plugging in the true parameter θ_0, η_0, we obtain

$$\psi(W;\theta_0,\eta_0) = U(D - \mu_0 X), \ \text{ where } \ \mu_0 = E_P[DX^T](E_P[XX^T])^{-1} = \gamma_0^T$$

3.2.3 Concentrating-out Approach

We skip the Neyman near-orthogonal scores part first for simplicity
If it's hard to understand, we shall refer to Newey(1994)[3]

For all $\theta \in \Theta$, let β_θ be the solution of the following optimization problem:

$$\max_{\beta \in \mathcal{B}} E_P[l(W; \theta, \beta)]$$

Under mild regularity conditions, β_θ satisfies

$$\partial_\beta E_P[l(W; \theta, \beta_\theta)] = 0, \quad \text{for all } \theta \in \Theta$$

We differentiate the above equation with respect to θ and interchange the order of differentiation gives

$$\begin{aligned} 0 &= \partial_\theta \partial_\beta E_P[l(W; \theta, \beta_\theta)] \\ &= \partial_\beta \partial_\theta E_P[l(W; \theta, \beta_\theta)] \\ &= \partial_\beta E_P[\partial_\theta l(W; \theta, \beta_\theta) + [\partial_\theta \beta_\theta]^T \partial_\beta l(W; \theta, \beta_\theta)] \\ &= \partial_\beta E_P[\psi(W; \theta, \beta, \partial_\theta \beta_\theta)]|_{\beta = \beta_\theta} \end{aligned}$$

where we denote $\psi(W; \theta, \beta, \partial_\theta \beta_\theta) \equiv \partial_\theta l(W; \theta, \beta) + [\partial_\theta \beta_\theta]^T \partial_\beta l(W; \theta, \beta)$

Re-read this paragraph after studying 2.2.3

The **vector** of functions is a score with nuisance parameters $\eta = (\beta^T, vec(\partial_\theta \beta_\theta))^T$. The additional nuisance parameters, $\partial_\theta \beta_\theta$ in this case, are introduced when the orthogonal score is formed. Evaluating these equations at θ_0 and β_0, it follows that $\psi(W; \theta, \beta, \partial_\theta \beta_\theta)$ is orthogonal with respect to β and from $E_P[\partial_\beta l(W; \theta_0, \beta_0)] = 0$ that we have orthogonality with respect to $\partial_\theta \beta_\theta$. Thus, maximizing the expected objective function with respect to the nuisance parameters, plugging that maximum back in, and differentiating with respect to the parameters of interest produces an orthogonal moment condition. See the next section.

3.2.4 Neyman orthogonal scores for likelihood and other M-estimation problems with infinite-dimensional nuisance parameters

We try to extend the concentrating out approach to infinite-dimensional nuisance parameters. Let $l(W; \theta, \beta)$ be a known criterion function, where θ and β are the target and the nuisance parameters. θ_0 and β_0 are the true values for solving the maximizing log-likelihood optimization problem. We assume that $\mathcal{B}$ is some convex set of functions, so that β is the functional nuisance parameter. For example, $l(W; \theta, \beta)$ could be a semi-parametric log-likelihood where β is the non-parametric part of the model. More generally, $l(W; \theta, \beta)$ could be some other criterion function such as the negative of a squared residual. Let

$$\beta_\theta = arg \max_{\beta \in \mathcal{B}} E_P[l(W; \theta, \beta)]$$

be the concentrated-out non-parametric part of the part of the model. Note that β_θ is a function-valued function. Now consider the score function

$$\psi(W;\theta,\eta) = \frac{dl(W;\theta,\eta(\theta))}{d\theta}$$

where the nuisance parameter is $\eta : \Theta \to \mathcal{B}$, and its true value η_0 is given by

$$\eta_0(\theta) = \beta_\theta, \quad \text{for all} \quad \theta \in \Theta$$

LEMMA 2.5. (NEYMAN ORTHOGONAL SCORES VIA CONCENTRATING-OUT APPROACH) *Suppose that* (2.5) *holds, and let* T *be a convex set of functions mapping* Θ *into* $\mathcal{B}$ *such that* $\eta_0 \in T$. *Also, suppose that for each* $\eta \in T$, *the function* $\theta \mapsto \ell(W;\theta,\eta(\theta))$ *is continuously differentiable almost surely. Then, under mild regularity conditions, the score* ψ *in* (2.21) *is Neyman orthogonal at* (θ_0, η_0) *with respect to the nuisance realization set* $\mathcal{T}_N = T$.

Proof of Lemma 2.5: Take any $\eta \in T$, and consider the function

$$Q(W;\theta,r) := \ell(W;\theta,\eta_0(\theta) + r(\eta(\theta) - \eta_0(\theta))), \quad \theta \in \Theta,\ r \in [0,1].$$

Then

$$\psi(W;\theta,\eta_0 + r(\eta - \eta_0)) = \partial_\theta Q(W;\theta,r),$$

and so

$$\begin{aligned}\partial_r E_P[\psi(W;\theta,\eta_0 + r(\eta-\eta_0))] &= \partial_r E_P[\partial_\theta Q(W;\theta,r)] \\ &= \partial_r \partial_\theta E_P[Q(W;\theta,r)] = \partial_\theta \partial_r E_P[Q(W;\theta,r)] \\ &= \partial_\theta \partial_r E_P[\ell(W;\theta,\eta_0(\theta) + r(\eta(\theta) - \eta_0(\theta)))]. \qquad \text{(A.2)}\end{aligned}$$

Hence,

$$\partial_r E_P[\psi(W;\theta,\eta_0 + r(\eta-\eta_0))]|_{r=0} = 0$$

because

$$\partial_r E_P[\ell(W;\theta,\eta_0(\theta) + r(\eta(\theta) - \eta_0(\theta)))]|_{r=0} = 0, \quad \text{for all } \theta \in \Theta,$$

as $\eta_0(\theta) = \beta_\theta$ solves the optimization problem

$$\max_{\beta \in \mathcal{B}} E_P[\ell(W;\theta,\beta)], \quad \text{for all } \theta \in \Theta.$$

Here, the regularity conditions are needed to make sure that we can interchange E_P and ∂_θ and also ∂_θ and ∂_r in (A.2). □

As an example, consider the partially linear model. Let

$$l(W;\theta,\beta) = -\frac{1}{2}(Y - D\theta - \beta(X))^2$$

and let $\mathcal{B}$ be the set of functions of X with finite mean square. Then

$$(\theta_0,\beta_0) = arg \max_{\theta\in\Theta,\beta\in\mathcal{B}} E_P[l(W;\theta,\beta)]$$

and

$$\beta_\theta(X) = E_P[Y - D\theta|X], \;\; \theta \in \Theta$$

Hence, applying the above formula

$$\begin{aligned}\psi(W;\theta,\beta_\theta) &= -\frac{1}{2}\frac{d\{Y - D\theta - E_P[Y - D\theta|X]\}^2}{d\theta} \\ &= (D - E_P[D|X]) \times (Y - E_P[Y|X] - (D - E_P[D|X])\theta) \\ &= (D - m_0(X)) \times (Y - D\theta - g_0(X))\end{aligned}$$

We also skip the conditional moment setting

3.2.5 Neyman orthogonal scores and influence functions

Consider the original score $\varphi(W;\theta,\beta)$, where β is some function, and let $\hat{\beta}_0$ be a non-parametric estimator of β_0, the true value of β. Here, β is implicitly allowed to depend on θ. The corresponding orthogonal score can be formed when there is $\phi(W;\theta,\eta)$ such that

$$\int \varphi(w;\theta_0,\hat{\beta}_0)dP(w) = \frac{1}{n}\sum_{i=1}^{n}\phi(W_i;\theta_0,\eta_0) + o_p(n^{-1/2})$$

where η is a vector of nuisance functions that includes β, and $\phi(W;\theta,\eta)$ is an adjustment for the presence of the presence of the estimated function $\hat{\beta}_0$ in the original score $\varphi(W;\theta,\beta)$. The decomposition typically holds when $\hat{\beta}$ is either a kernel or a series estimator with a suitably chosen tuning parameter. The Neyman orthogonal score is given by

$$\psi(W;\theta,\eta) = \varphi(W;\theta,\beta) + \phi(W;\theta,\eta)$$

Here $\psi(W;\theta_0,\eta_0)$ is the influence function of the limit of $\frac{1}{n}\varphi(W_i;\theta_0,\hat{\beta}_0)$ with the restriction $E_P[\psi(W;\theta_0,\eta_0)] = 0$ identifying θ_0

For example, consider the PLR with the original score

$$\varphi(W;\theta,\beta) = D(Y - D\theta - g_0(X))$$

Here $\hat{\beta}_0 = \hat{g}_0$ is a non-parametric regression estimator. The influence function adjustment is obtained as:

$$\phi(W;\theta,\eta) = -m_0(X)[Y - D\theta - \beta(X,\theta)]$$

The corresponding orthogonal score is then simply

$$\begin{aligned}\varphi(W;\theta,\eta) &= (D - m_0(X))(Y - D\theta - \beta(X,\theta)) \\ \beta_0(X,\theta) &= E_P[Y - D\theta|X], \;\; m_0(X) = E_P[D|X]\end{aligned}$$

illustrating that an orthogonal score for the PLR can be derived from an influence function adjustment.

References

[1] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," 2018.

[2] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[3] W. K. Newey, "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, pp. 1349–1382, 1994.