Econometrics **January 24, 2023**

Topic 3: *Moving the Goalposts* Approach

by Sai Zhang

Key points:

-

Disclaimer: *These notes are written by Sai Zhang (email me or check my Github page). The main references for this topic are Armstrong et al. (2020); Armstrong and Kolesár (2018), I thank Prof. Armstrong for his valuable advice.*

3.1 Finite Sample Bias-Variance Tradeoffs

3.1.1 Setup

Consider the fixed design regression model

$$y_i = w_i \beta\left(z_i\right) + h\left(z_i\right) + \epsilon_i \tag{3.1}$$

where

- w_i, z_i are treated as **fixed**
- ϵ_i is **independent**, with $\mathbb{E}\left[\epsilon_i\right] = 0, \mathbb{E}\left[\epsilon_i^2\right] = \sigma_i^2$
- observation: $\left\{\left(y_i, w_i, z_i'\right)'\right\}_{i=1}^n$

one example is the case where w_i is **binary**, then

$$\beta(z) = f(1, z) - f(0, z)$$

which is just the ATE conditional on z under the unconfoundedness assumption. This includes the RD design, where z_i is the running variable and w_i is the treatment assignment.

Now, consider for the weighted average treatment effect

$$L_\mu\left[\beta(\cdot)\right] = \int \beta(z) \mathrm{d}\mu(z)$$

where $\int \mu(z) = 1$ is a **signed** measure (weight, allowing **negative** weights), construct a linear estimator

$$\hat{L}_a = \sum_{i=1}^n a_i y_i$$

where the estimation weights a_i can depend on $\left\{z_i, w_i, \sigma_i^2\right\}_{i=1}^n$, but **not** on y_i. Together, the bias of $\hat{L}_a$ for $L_\mu\left[\beta(\cdot)\right]$, given the regression function $\beta(\cdot), h(\cdot)$, is

$$\mathbb{E}_{\beta(\cdot),h(\cdot)}\left[\hat{L}_a\right] - L_\mu\left[\beta(\cdot)\right] = \sum_{i=1}^n a_i\left[w_i\beta(z_i) + h(z_i)\right] - \int \beta(z)\mathrm{d}\mu(z)$$

and its variance, given the regression function $\beta(\cdot), h(\cdot)$, is just

$$\mathrm{Var}_{\beta(\cdot),h(\cdot)}\left[\hat{L}_a\right] = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

To bound the bias, assume $h(\cdot)$ is known to belong in a class of functions $\mathcal{H}$, then two approaches can be adopted, for the regularity of $\beta(\cdot)$ and the choice of $\mu(\cdot)$:

1 arbitrary $\beta(\cdot)$, optimizing weights μ by *moving the goalposts*, s.t. $L_\mu\left[\beta(\cdot)\right]$ is easy to estimate (Crump et al., 2006; Imbens and Wager, 2019) which gives the worst-case bias

$$\inf_{\mu} \sup_{\beta(\cdot),h(\cdot)} \left| \sum_{i=1}^{n} a_i \left[w_i \beta(z_i) + h(z_i)\right] - \int \beta(z) \mathrm{d}\mu(z) \right| \qquad \text{s.t. } h(\cdot) \in \mathcal{H}, \int \mathrm{d}\mu(z) = 1 \tag{3.2}$$

2 assume constant treatment effects, i.e., $\beta(z) = \beta, \forall z$, which means that $L_\mu\left[\beta(\cdot)\right] = \beta$ regardless of μ (Armstrong et al., 2020), and the worst-case bias is

$$\sup_{\beta,h(\cdot)} \left| \sum_{i=1}^{n} a_i \left[w_i \beta + h(z_i)\right] - \beta \right| \qquad \text{s.t. } h(\cdot) \in \mathcal{H} \tag{3.3}$$

And, the two approaches can be linked as such:

- If $\sum_{i=1}^{n} a_i w_i = 1$, 3.2 and 3.3 are both equal to

$$\sup_{h(\cdot)} \left| \sum_{i=1}^{n} a_i h(z_i) \right| \text{ s.t. } h(\cdot) \in \mathcal{H} \tag{3.4}$$

  - 3.2 automatically equals 3.4
  - 3.3 is optimized (w.r.t. μ) by setting μ to place weight $a_i w_i$ on observation i, i.e., $\mu(\mathcal{Z}) = \sum_{i:z_i \in \mathcal{Z}} a_i w_i$, which implies $\sum_{i=1}^{n} a_i w_i \beta(z_i) - \int \beta(z) \mathrm{d}\mu(z) = 0$, hence the equality.
- Otherwise, 3.2 and 3.3 are both infinite:
  - 3.3 can be made arbitrarily large by choosing large enough β
  - 3.2 can be made arbitrarily large by making $\beta(\cdot)$ constant (as in 3.3) and large enough

3.1.2 Moving-the-goalpost Approach

3.1.3 Constant-treatment-effect Approach

Armstrong et al. (2020) adopt this approach, focusing on the case where $h(\cdot)$ is a high dimensional linear function, and the penalty function is an l_p norm of the coefficients.

Basic setting: Homoskedastic Gaussian errors

First, consider

$$Y = w\beta + Z\gamma + \epsilon \tag{3.5}$$

where

- $\beta \in \mathbb{R}$ is the constant treatment effect to be estimated
- $\gamma \in \Gamma$ is the control coefficients, subject to the restriction (i.e., the function class $\mathcal{H}$)

$$\Gamma = \Gamma(C) = \left\{\gamma \in \mathcal{G} : \mathrm{Pen}(\gamma) \leq C\right\} \tag{3.6}$$

  where $\mathrm{Pen}(\cdot)$ is a seminorm[1] on some linear subspace $\mathcal{G}$ of $\mathbb{R}^k$.
- $w = (w_1, \cdots, w_n)' \in \mathbb{R}^n$ and $Z = \left(z_1', \cdots, z_n'\right)' \in \mathbb{R}^{n\times k}$ are defined as before
- $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$ is assumed **normal and homoskedastic**, with σ^2 known

For estimation, the goal is to construct estimators and CIs for β:

- estimator $\hat{\beta}$: consider the worst-case performance over the parameter space $\mathbb{R}\times\Gamma$ under the **MSE** criterion

$$R_{MSE}\left(\hat{\beta};\Gamma\right) = \sup_{\beta\in\mathbb{R},\gamma\in\Gamma} \mathbb{E}_{\beta,\gamma}\left[\left(\hat{\beta}-\beta\right)^2\right]$$

- for CIs, we have 2 requirements:

  A <u>**coverage**</u>: A $100\cdot(1-\alpha)\%$ CI with half-length $\hat{\chi} = \hat{\chi}(Y, X)$ is an interval $\left\{\hat{\beta} \pm \hat{\chi}\right\}$ s.t.

$$\inf_{\beta\in\mathbb{R},\gamma\in\Gamma} \mathrm{P}_{\beta,\gamma}\left(\beta \in \left\{\hat{\beta} \pm \hat{\chi}\right\}\right) \geq 1-\alpha$$

  B <u>**length**</u>: the exepcted length of a CI $\mathbb{E}_{\beta,\gamma}\left[2\hat{\chi}\right]$ should be as short as possible

  notice that length-optimized CIs are **not** necessarily centered at an MSE-centered $\hat{\beta}$.

Linear estimators and CIs

Again, consider estimators that are **linear** in the outcomes Y, $\hat{\beta} = a'Y$, where a is the $n-$vector weights. In the vector form, the worst-case bias (as 3.3) is

$$\overline{\mathrm{bias}}_\Gamma\left(\hat{\beta}\right) = \sup_{\beta\in\mathbb{R},\gamma\in\Gamma} a'\left(w\beta + Z\gamma\right) - \beta \tag{3.7}$$

and the variance, under the assumption of homoskedasticity, is

$$\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2 a'a$$

Then the MSE is

$$R_{MSE}\left(\hat{\beta};\Gamma\right) = \sup_{\beta\in\mathbb{R},\gamma\in\Gamma} \mathbb{E}_{\beta,\gamma}\left[\left(\hat{\beta}-\beta\right)^2\right] = \overline{\mathrm{bias}}_\Gamma\left(\hat{\beta}\right)^2 + \mathrm{Var}\left(\hat{\beta}\right)$$

The $t-$statistic is

$$\frac{\hat{\beta}-\beta}{\sqrt{\mathrm{Var}\left(\hat{\beta}\right)}} \sim \mathcal{N}(b, 1),\ |b| \leq \frac{\overline{\mathrm{bias}}_\Gamma\left(\hat{\beta}\right)}{\sqrt{\mathrm{Var}\left(\hat{\beta}\right)}}$$

and a two-sided CI can then be formed as

$$\hat{\beta} \pm \chi, \qquad \text{where } \chi = \sqrt{\mathrm{Var}\left(\hat{\beta}\right)}\cdot \mathrm{cv}_\alpha\left(\frac{\overline{\mathrm{bias}}_\Gamma\left(\hat{\beta}\right)}{\sqrt{\mathrm{Var}\left(\hat{\beta}\right)}}\right) \tag{3.8}$$

and the $\mathrm{cv}_\alpha(B)$ denotes the $1-\alpha$ quantile of a $|\mathcal{N}(B, 1)|$. This is a **fixed-length confidence interval** (FLCI), with a fixed length of 2χ. It depends on X and σ^2, but not on Y or $(\beta, \gamma)'$.

[1]Seminorm satisfies **triangle inequality** $\mathrm{Pen}\left(\gamma + \tilde{\gamma}\right) \leq \mathrm{Pen}(\gamma)$ and **homogeneity** $\mathrm{Pen}\left(c\gamma\right) = |c|\mathrm{Pen}(\gamma), \forall c$, but **NOT** necessarily positive definite ($\mathrm{Pen}(\gamma) = 0$ does not imply $\gamma = 0$). Essentially, any convex set Γ that is symmetric satisfies this definition.

Optimal weights

We have two optimization goals

- minimizing MSE: $R_{MSE}\left(\hat{\beta};\Gamma\right) = \overline{\text{bias}}_{\Gamma}\left(\hat{\beta}\right)^2 + \text{Var}\left(\hat{\beta}\right)$
- minimizing CI length: $\chi = \sqrt{\text{Var}\left(\hat{\beta}\right)} \cdot \text{cv}_{\alpha}\left(\overline{\text{bias}}_{\Gamma}\left(\hat{\beta}\right) / \sqrt{\text{Var}\left(\hat{\beta}\right)}\right)$

They both increasing in $\text{Var}\left(\hat{\beta}\right)$ and $\overline{\text{bias}}_{\Gamma}\left(\hat{\beta}\right)$, hence to find the optimal weights, it suffices to minimize variance subject to a bound B on worst-case bias, which can be written as:

$$\min_{a\in\mathbb{R}} a'a \text{ s.t. } \sup_{\beta\in\mathbb{R},\gamma\in\Gamma} a'\left(w\beta + Z\gamma\right) - \beta \leq B \tag{3.9}$$

The optimal weight is then given by:

Theorem 3.1.1: Optimal Weight

Let π^*_{λ} be a solution to[a]

$$\min_{\pi} \|w - Z\pi\|_2^2 \;\text{ s.t. }\; \text{Pen}(\pi) \leq t_{\lambda} \tag{3.10}$$

and suppose that $\|w - Z\pi\|_2 > 0$, $\text{Pen}(\cdot)$ is continuous, then the optimal weight solving 3.9 is

$$a^*_{\lambda} = \frac{w - Z\pi^*_{\lambda}}{\left(w - Z\pi^*_{\lambda}\right)' w}$$

with the bound

$$B = \frac{C}{t_{\lambda}} \cdot \frac{\left(w - Z\pi^*_{\lambda}\right)' Z\pi^*_{\lambda}}{\left(w - Z\pi^*_{\lambda}\right)' w}$$

Consequently, we have

- estimator

$$\hat{\beta}_{\lambda} = a^{*\prime}_{\lambda} Y = \frac{\left(w - Z\pi^*_{\lambda}\right)' Y}{\left(w - Z\pi^*_{\lambda}\right)' w}$$

- worst-case bias

$$\overline{\text{bias}}_{\Gamma}\left(\hat{\beta}_{\lambda}\right) = C\overline{B}_{\lambda} = \frac{C}{\text{Pen}\left(\pi^*_{\lambda}\right)} \frac{\left(w - Z\pi^*_{\lambda}\right)' Z\pi^*_{\lambda}}{\left(w - Z\pi^*_{\lambda}\right)' w}$$

- variance of estimator

$$V_{\lambda} = \frac{\sigma^2 \|w - Z\pi^*_{\lambda}\|_2^2}{\left[\left(w - Z\pi^*_{\lambda}\right)' w\right]^2}$$

[a] This regression can be refered to as a regularized propensity score regression (but w_i need not be binary) with penalty $\text{Pen}(\pi)$

This result follows by applying Donoho (1994), Low (1995) and Armstrong and Kolesár (2018), rewriting 3.9

as a convex optimization problem.

A Proofs

A.1 Proof of Theorem 3.1.1

Following Armstrong and Kolesár (2018, Equation (25)), the modulus of continuity is given by

$$\omega(\delta) = \sup_{\beta,\gamma} 2\beta \qquad \text{s.t. } \|w\beta + Z\gamma\|_2 \leq \frac{\delta}{2}, \qquad \text{Pen}(\gamma) \leq C$$

Introducing a substitution (rescaling γ by β) $\pi = -\frac{\gamma}{\beta}$, get

$$\omega(\delta) = \sup_{\beta,\pi} 2\beta \qquad \text{s.t. } \beta\|w - Z\pi\|_2 \leq \frac{\delta}{2}, \qquad \beta\text{Pen}(\pi) \leq C \tag{3.11}$$

recall the optimization problem in Theorem 3.1.1:

$$\min_{\pi} \|w - Z\pi\|_2^2 \ \text{ s.t. } \ \text{Pen}(\pi) \leq t_\lambda$$

We can relate the two problems via the following logic: we want to make $\|w - Z\pi\|_2$ and $\text{Pen}(\pi)$ small so that large values of β satisfy the constraint of 3.11. Formally:

Lemma A.1

- If $\exists \pi \in \mathcal{G}$ s.t. $w = Z\pi$ and $\text{Pen}(\pi) = 0$, then $w(\delta) = \infty, \forall \delta \geq 0$ (automatic)
- **Otherwise**:

(i) $\forall \delta > 0$, the problem 3.11 has a solution $\beta_\delta^{mod}, \pi_\delta^{mod}$ with $\beta_\delta^{mod} > 0$. For $t_\lambda = \frac{C}{\beta_\delta^{mod}} = \frac{2C}{w(\delta)}$, π_δ^{mod} is also a solution to the penalized regression (3.10)

$$\min_{\pi} \|w - Z\pi\|_2^2 \ \text{ s.t. } \ \text{Pen}(\pi) \leq t_\lambda$$

with optimized objective

$$\|w - Z\pi\|_2 = \frac{\delta}{2\beta_\delta^{mod}} = \frac{\delta}{w(\delta)} > 0$$

(ii) $\forall t_\lambda > 0$, the penalized regression above has a solution $\pi_\lambda^\star$. Setting

$$\beta_\lambda^* = \frac{C}{t_\lambda}$$
$$\delta_\lambda = 2\beta_\lambda^* \|w - Z\pi\|_2 = \frac{2C}{t_\lambda}\|w - Z\pi\|_2$$

the pair $\left(\beta_\lambda^*, \pi_\lambda^*\right)$ solves the modulus problem 3.11 at $\delta = \delta_\lambda$, with optimized objective $w(\delta_\lambda) = \frac{2C}{t_\lambda}$, as long as $\|w - Z\pi\|_2 > 0$

Proof of Lemma A.1: we prove the lemma with the following steps:

A. the penalized problem 3.10 has a solution:

Let $\mathcal{G}^{(0)}$ denote the linear subspace of vectors $\pi \in \mathcal{G}$ s.t. $Z\pi = 0$, $\text{Pen}(\pi) = 0$; let $\mathcal{G}^{(1)}$ be a subspace s.t. $\mathcal{G} = \mathcal{G}^{(0)} \oplus \mathcal{G}^{(1)}$. Then, we can write $\pi \in \mathcal{G}$ uniquely as $\pi = \pi^{(0)} + \pi^{(1)}$ where $\pi^{(0)} \in \mathcal{G}^{(0)}, \pi^{(1)} \in \mathcal{G}^{(1)}$.

Therefore, we have $Z\pi = Z\pi^{(1)}$, and

$$\begin{rcases} \mathrm{Pen}(\pi^{(1)}) = \mathrm{Pen}(\pi^{(1)}) - \mathrm{Pen}(-\pi^{(0)}) \leq \mathrm{Pen}(\pi) \\ \mathrm{Pen}(\pi^{(1)}) = \mathrm{Pen}(\pi^{(1)}) + \mathrm{Pen}(\pi^{(0)}) \geq \mathrm{Pen}(\pi) \end{rcases} \Rightarrow \mathrm{Pen}(\pi^{(1)}) = \mathrm{Pen}(\pi)$$

THen, 3.10 can be wriiten in terms of $\pi^{(1)} \in \mathcal{G}^{(1)}$ only. The level sets of this optimization problem are bounded and closed (by continuity of the seminorm $\mathrm{Pen}(\cdot)$), so it has a solution, which is also the solution to the original problem.

B <u>the modulus problem 3.11 has a solution</u>: for the problem 3.11, feasible values of β are bounded as:

$$\beta \leq \frac{\delta}{2} \cdot \frac{1}{\|w - Z\pi\|_2} \qquad\qquad \beta \leq C \cdot \frac{1}{\mathrm{Pen}(\pi)}$$

i.e., β is bounded by the inverse of the minimum of $\max\{\|w - Z\pi\|_2, \mathrm{Pen}(\pi)\}$ over π, and it is strictly positive. Hence, $\beta, \tilde{\pi}^{(1)}$ can be restricted to a compact set without changing the optimization problem.

C <u>proof of statement (i)</u>: Proof by contradiction, if it's not true, then $\exists \tilde{\pi}$ s.t.

$$\mathrm{Pen}(\tilde{\pi}) \leq \frac{C}{\beta_\delta^{mod}} \equiv t_\lambda, \qquad\qquad \|w - Z\tilde{\pi}\|_2 \leq \|w - Z\pi_\delta^{mod}\|_2 - \nu$$

then for some η, let $\tilde{\pi}_\eta = (1-\eta)\tilde{\pi}$, we have

$$\begin{aligned} \|w - Z\tilde{\pi}_\eta\|_2 &= \|w - Z(1-\eta)\tilde{\pi}\|_2 \\ &\leq \|w - X\tilde{\pi}\|_2 + \eta\|Z\tilde{\pi}\|_2 \\ &\leq \|w - X\pi_\delta^{mod}\|_2 - v + \eta\|Z\tilde{\pi}\|_2 \\ &\leq \frac{\delta}{2\beta_\delta^{mod}} - v + \eta\|Z\tilde{\pi}\|_2 \end{aligned}$$

Hence, $\exists\eta$ small enough, s.t.

$$\|w - Z\tilde{\pi}_\eta\|_2 < \frac{\delta}{2\beta_\delta^{mod}}, \qquad\qquad \mathrm{Pen}(\tilde{\pi}_\eta) \leq (1-\eta)\frac{C}{\beta_\delta^{mod}} < \frac{C}{\beta_\delta^{mod}}$$

therefore, by setting $\pi = \tilde{\pi}_\eta$, we can allow a strictly bigger β, which is a contradiction.

D <u>proof of statement (ii)</u>: This result follows immediately.

Next, use Lemma A.1 to prove Theorem 3.1.1. Following Armstrong and Kolesár (2018), the class of bias-variance optimizing estimators is

$$\underline{\left(w\beta_{delta}^{mod} + Z\gamma_\delta^{mod}\right)}$$

References

Timothy B Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

Timothy B Armstrong, Michal Kolesár, and Soonwoo Kwon. Bias-aware inference in regularized regression models. *arXiv preprint arXiv:2012.14823*, 2020.

Richard K Crump, V Joseph Hotz, Guido Imbens, and Oscar Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, 2006.

David L Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270, 1994.

Guido Imbens and Stefan Wager. Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278, 2019.

Mark G Low. Bias-variance tradeoffs in functional estimation problems. *The Annals of Statistics*, 23(3): 824–835, 1995.