

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 reference for this topic is [Armstrong, Kolesár, and Kwon \(2020\)](#), 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

$$y_i = w_i \beta(z_i) + h(z_i) + \epsilon_i \quad (3.1)$$

where

- w_i, z_i are treated as **fixed**
- ϵ_i is **independent**, with $\mathbb{E}[\epsilon_i] = 0, \mathbb{E}[\epsilon_i^2] = \sigma_i^2$
- 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

$$L_\mu [\beta(\cdot)] = \int \beta(z) d\mu(z)$$

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

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

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

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

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. $L_\mu [\beta(\cdot)]$ is easy to estimate (Crump et al., 2006; Imbens and Wager, 2019) which gives the worst-case bias

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

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

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

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

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

- 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 $\sum_{i=1}^n a_i w_i \beta(z_i) - \int \beta(z) d\mu(z) = 0$, 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

$$Y = w\beta + Z\gamma + \epsilon \quad (3.5)$$

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})

$$\Gamma = \Gamma(C) = \{\gamma \in \mathcal{G} : \text{Pen}(\gamma) \leq C\} \quad (3.6)$$

where $\text{Pen}(\cdot)$ is a seminorm¹ on some linear subspace \mathcal{G} of \mathbb{R}^k .

- $w = (w_1, \dots, w_n)' \in \mathbb{R}^n$ and $Z = (z'_1, \dots, z'_n)' \in \mathbb{R}^{n \times k}$ are defined as before
- $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ 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

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

- for CIs, we have 2 requirements:

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

$$\inf_{\beta \in \mathbb{R}, \gamma \in \Gamma} P_{\beta, \gamma}(\beta \in \{\hat{\beta} \pm \hat{\chi}\}) \geq 1 - \alpha$$

B **length**: the expected length of a CI $\mathbb{E}_{\beta, \gamma}[2\hat{\chi}]$ 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

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

and the variance, under the assumption of homoskedasticity, is

$$\text{Var}(\hat{\beta}) = \sigma^2 a'a$$

Then the MSE is

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

The t -statistic is

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

and a two-sided CI can then be formed as

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

and the $\text{cv}_{\alpha}(B)$ 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 $(\beta, \gamma)'$.

¹Seminorm satisfies **triangle inequality** $\text{Pen}(\gamma + \tilde{\gamma}) \leq \text{Pen}(\gamma) + \text{Pen}(\tilde{\gamma})$ and **homogeneity** $\text{Pen}(c\gamma) = |c| \text{Pen}(\gamma), \forall c$, but **NOT** necessarily positive definite ($\text{Pen}(\gamma) = 0$ 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: $R_{MSE}(\hat{\beta}; \Gamma) = \overline{\text{bias}}_{\Gamma}(\hat{\beta})^2 + \text{Var}(\hat{\beta})$
- minimizing CI length: $\chi = \sqrt{\text{Var}(\hat{\beta})} \cdot \text{cv}_{\alpha} \left(\overline{\text{bias}}_{\Gamma}(\hat{\beta}) / \sqrt{\text{Var}(\hat{\beta})} \right)$

They both increasing in $\text{Var}(\hat{\beta})$ and $\overline{\text{bias}}_{\Gamma}(\hat{\beta})$, 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:

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

The optimal weight is then given by:

Theorem 3.1.1: Optimal Weight

Let π_{λ}^* be a solution to^a

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

and suppose that $\|w - Z\pi\|_2 > 0$, then the optimal weight solving 3.9 is

$$a_{\lambda}^* = \frac{w - Z\pi_{\lambda}^*}{(w - Z\pi_{\lambda}^*)' w}$$

with the bound

$$B = \frac{C}{t_{\lambda}} \cdot \frac{(w - Z\pi_{\lambda}^*)' Z\pi_{\lambda}^*}{(w - Z\pi_{\lambda}^*)' w}$$

Consequently, we have

- estimator

$$\hat{\beta}_{\lambda} = a_{\lambda}^* Y = \frac{(w - Z\pi_{\lambda}^*)' Y}{(w - Z\pi_{\lambda}^*)' w}$$

- worst-case bias:

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

- variance of estimator:

^aThis regression can be referred to as a regularized propensity score regression (but w_i need not be binary) with penalty $\text{Pen}(\pi)$

References

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