Type-I error control for implicit confounder adjustment procedure with marginal resampling

Gene

October 1, 2022

1 Type-I error control statement

Let $(X_i, Y_i, Z_i) \overset{\text{i.i.d.}}{\sim} \mathcal{L}$ for some joint law, such that

$$\mathcal{L}(\boldsymbol{Y}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, \sigma_Y^2) \tag{1}$$

and therefore $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$. Given data $(X, Y, Z) = \{(X_i, Y_i, Z_i)\}_{1\leq i\leq n}$ (the dependence on n is left implicit), consider the test statistic

$$T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} X^T(Y - Z\widehat{\beta}), \tag{2}$$

where $\widehat{\beta}$ is the OLS estimate of β. Consider the critical value defined by marginal resampling:

$$C_n(Y, Z) \equiv \mathbb{Q}_{1-\alpha}[T_n(\widetilde{X}, Y, Z) \mid Y, Z], \quad \text{where } \widetilde{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{L}(\boldsymbol{X}). \tag{3}$$

Combining the test statistic (2) with the critical value (3), we arrive at the test

$$\phi_n(X, Y, Z) \equiv \mathbb{1}(T_n(X, Y, Z) > C_n(Y, Z)). \tag{4}$$

Proposition 1. *For $\mathcal{L}$ satisfying assumption* (1) *and moment conditions, the test ϕ_n has asymptotic Type-I error control, i.e.*

$$\limsup_{n\to\infty}\ \mathbb{E}_{\mathcal{L}}[\phi_n(X, Y, Z)] \leq \alpha. \tag{5}$$

The proof sketch of Proposition 1 (Section 2), depends on extensions of classical convergence results to the conditional case (Appendix A).

2 Proof sketch of Proposition 1

Denote

$$\sigma_X^2 \equiv \mathbb{E}_{\mathcal{L}}[\boldsymbol{X}^2]. \tag{6}$$

It suffices to show that

$$T_n(X,Y,Z) \xrightarrow{d} N(0,\sigma_X^2\sigma_Y^2) \tag{7}$$

and

$$T_n(\widetilde{X},Y,Z) \mid Y,Z \xrightarrow{d,p} N(0,\sigma_X^2\sigma_Y^2), \tag{8}$$

where the latter statement denotes conditional convergence in distribution (see Definition 1). Indeed, by Lemma 2, the conditional convergence (8) implies the convergence of the critical value in probability to a scaled normal quantile:

$$C_n(Y,Z) \equiv \mathbb{Q}_{1-\alpha}[T_n(\widetilde{X},Y,Z) \mid Y,Z] \xrightarrow{p} \mathbb{Q}_{1-\alpha}[N(0,\sigma_X^2\sigma_Y^2)] = \sigma_X\sigma_Y z_{1-\alpha}. \tag{9}$$

Taken together, the convergence of the test statistic (7) and the convergence of the critical value (9) imply Type-I error control (5).

To verify the convergence statements (7) and (8), we claim first that

$$T_n(X,Y,Z) = \frac{1}{\sqrt{n}}X^T(Y-Z\beta)+o_p(1), \quad T_n(\widetilde{X},Y,Z) = \frac{1}{\sqrt{n}}\widetilde{X}^T(Y-Z\beta)+o_p(1) \tag{10}$$

Indeed, by Cauchy-Schwarz,

$$\left|T_n(X,Y,Z) - \frac{1}{\sqrt{n}}X^T(Y-Z\beta)\right| = \left|\frac{1}{\sqrt{n}}X^T(Z\widehat{\beta}-Z\beta)\right| \leq \left\|\frac{1}{\sqrt{n}}X\right\| \cdot \|Z\widehat{\beta}-Z\beta\|. \tag{11}$$

The first part of conclusion (10) follows by observing that the LLN implies that $\left\|\frac{1}{\sqrt{n}}X\right\| \xrightarrow{p} \sigma_X$ and $Z\widehat{\beta}-Z\beta \xrightarrow{p} 0$ by standard OLS theory. The second part of conclusion (10) holds by the same logic.

Using the convergence statements (10), Slutsky and its conditional variant (Lemma 1), we see that to verify the convergences (7) and (8) it suffices to show that

$$\frac{1}{\sqrt{n}}X^T(Y-Z\beta) \xrightarrow{d} N(0,\sigma_X^2\sigma_Y^2) \tag{12}$$

and

$$\frac{1}{\sqrt{n}}\widetilde{X}^T(Y-Z\beta) \mid Y,Z \xrightarrow{d,p} N(0,\sigma_X^2\sigma_Y^2). \tag{13}$$

By assumption (1), we have $Y - Z\beta = \epsilon \sim N(0,\sigma_Y^2) \perp\!\!\!\perp X$. Hence, the first of the above statements follows by the CLT, while the second follows from the LLN, conditional Slutsky (Lemma 1), and the conditional CLT (Theorem 1).

A Conditional convergence results

Several classical convergence results have conditional analogs. Let $\mathcal{F}_n$ be a sequences of σ-algebras to condition on. These results are formulated in terms of the following conditional notion of convergence in distribution.

Definition 1 (Conditional convergence in distribution)**.** Given a sequence of random variables T_n, we say T_n converges conditionally on $\mathcal{F}_n$ in distribution to a random variable T if

$$\mathbb{P}_n[T_n \leq t \mid \mathcal{F}_n] \xrightarrow{p} \mathbb{P}[T \leq t] \quad \text{for each } t \in \mathbb{R}. \tag{14}$$

We denote this relation via $T_n|\mathcal{F}_n \xrightarrow{d,p} T$.

Theorem 1 (Conditional central limit theorem)**.** *Let W_{in} be a triangular array of random variables, such that for each n, W_{in} are mean zero and independent conditionally on $\mathcal{F}_n$. Define*

$$\overline{W}_n \equiv \frac{1}{n}\sum_{i=1}^{n} W_{in} \quad \textit{and} \quad s_n^2 \equiv \sum_{i=1}^{n} \mathrm{Var}[W_{in} \mid \mathcal{F}_n]. \tag{15}$$

If for some $\delta > 0$ we have

$$\frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{2+\delta} \mid \mathcal{F}_n] \xrightarrow{p} 0, \tag{16}$$

then

$$\frac{\sqrt{n}}{s_n}\overline{W}_n \mid \mathcal{F}_n \xrightarrow{d,p} N(0,1). \tag{17}$$

Lemma 1 (Conditional Slutsky's Theorem)**.** *Suppose a_n and b_n are sequences of random variables such that $a_n \xrightarrow{p} 1$ and $b_n \xrightarrow{p} 0$. If $T_n \mid \mathcal{F}_n \xrightarrow{d,p} T$ and T has continuous CDF, then*

$$a_n T_n + b_n \mid \mathcal{F}_n \xrightarrow{d,p} T. \tag{18}$$

Lemma 2 (Conditional convergence implies quantile convergence)**.** *If $T_n \mid \mathcal{F}_n \xrightarrow{d,p} T$ and T has continuous CDF, then for any $\alpha \in (0,1)$,*

$$\mathbb{Q}_\alpha[T_n \mid \mathcal{F}_n] \xrightarrow{p} \mathbb{Q}_\alpha[T]. \tag{19}$$