Tim

Limiting variance of $\hat{\beta}$ random design for both non-EIV and EIV models.

1 Non-EIV

Suppose we observe data $(x_1, y_1), \ldots, (x_n, y_n)$ from the following model:

$$\begin{cases} y_i = \beta_0 + \beta x_i + \epsilon_i \\ x_i \sim X \\ \epsilon_i \sim N(0,1) \\ x_i \perp\!\!\!\perp \epsilon_i. \end{cases}$$

Let the estimator $\hat{\beta}$ of β be given by

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}.$$

Our goal is to compute the limiting distribution

$$\sqrt{n}\left(\hat{\beta} - \beta\right).$$

Define

$$T = \sum_{i=1}^n (x_i - \bar{x}) y_i - \beta \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right].$$

Manipulating T: We first manipulate T. We have that

$$T_i = \sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i - \beta\sum_{i=1}^n x_i^2 + \beta(1/n)(n\bar{x})^2$$

$$= \left[\sum_{i=1}^n x_i y_i - \beta\sum_{i=1}^n x_i^2\right] + \left[\beta n(\bar{x})^2 - \bar{x}\sum_{i=1}^n y_i\right]$$

$$= \sum_{i=1}^n x_i(y_i - \beta x_i) + \bar{x}\left(\bar{x}n\beta - \sum_{i=1}^n y_i\right)$$

$$= \sum_{i=1}^n x_i(y_i - \beta x_i) + \bar{x}\left(\sum_{i=1}^n \beta x_i - \sum_{i=1}^n y_i\right)$$

$$\sum_{i=1}^n x_i(y_i - \beta x_i) - \bar{x}\sum_{i=1}^n (y_i - \beta x_i) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \beta x_i)$$

$$= \sum_{i=1}^n (x_i - \bar{x})(y_i - \beta_0 - \beta x_i) = \sum_{i=1}^n (x_i - \bar{x})\epsilon_i = \sum_{i=1}^n (x_i - \mathbb{E}[x_i] + \mathbb{E}[x_i] - \bar{x})\epsilon_i$$

$$= \sum_{i=1}^n (x_i - \mathbb{E}[x_i])\epsilon_i + \sum_{i=1}^n (\mathbb{E}[x_i] - \bar{x})\epsilon_i$$

$$= \sum_{i=1}^n (x_i - \mathbb{E}[x_i])\epsilon_i - (\bar{x} - \mathbb{E}[x_i])\sum_{i=1}^n \epsilon_i.$$

Writing down limiting distribution: We can express

$$\sqrt{n}(\hat{\beta} - \beta)$$

as

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2} - \frac{\beta\left[\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\right]}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}\right]$$

$$= \sqrt{n}\left[\frac{(1/n)\sum_{i=1}^n (x_i - \mathbb{E}[x_i])\epsilon_i - (\bar{x} - \mathbb{E}[x_i])(1/n)\sum_{i=1}^n \epsilon_i}{(1/n)\sum_{i=1}^n x_i^2 - \left((1/n)\sum_{i=1}^n x_i\right)^2}\right]$$

$$= \left[\frac{(1/\sqrt{n})\sum_{i=1}^n (x_i - \mathbb{E}[x_i])\epsilon_i - (\bar{x} - \mathbb{E}[x_i])(1/\sqrt{n})\sum_{i=1}^n \epsilon_i}{(1/n)\sum_{i=1}^n x_i^2 - \left((1/n)\sum_{i=1}^n x_i\right)^2}\right].$$

Applying theorems:

1. We have that
$$\mathbb{E}\left((x_i - \mathbb{E}[x_i])\epsilon_i\right] = 0$$
and
$$V\left[(x_i - \mathbb{E}[x_i])\epsilon_i\right] = \mathbb{E}\left[\mathbb{V}\left((x_i - \mathbb{E}[x_i])\epsilon_i\right)|\epsilon_i\right] = \mathbb{E}\left[(x_i - \mathbb{E}[x_i])^2\right] = \mathbb{V}(x_i).$$
Therefore, by CLT,
$$(1/\sqrt{n})\sum_{i=1}^{n}(x_i - \mathbb{E}[x_i])\epsilon_i \xrightarrow{d} N(0, \mathbb{V}[x_i]).$$

2. $$(\bar{x} - \mathbb{E}[x_i]) \xrightarrow{P} 0.$$

3. $$(1/\sqrt{n})\sum_{i=1}^{n}\epsilon_i \xrightarrow{d} N(0, 1).$$

4. Combining (2) and (3) yields
$$(\bar{x} - \mathbb{E}[x_i])(1/\sqrt{n})\sum_{i=1}^{n}\epsilon_i \xrightarrow{P} 0.$$

5. Combining (1) and (4) yields (by Slutsky's theorem)
$$(1/\sqrt{n})\sum_{i=1}^{n}(x_i - \mathbb{E}[x_i])\epsilon_i - (\bar{x} - \mathbb{E}[x_i])(1/\sqrt{n})\sum_{i=1}^{n}\epsilon_i \xrightarrow{d} N(0, V[x_i]).$$

6. LLN implies that
$$(1/n)\sum_{i=1}^{n}x_i^2 - \left((1/n)\sum_{i=1}^{n}x_i\right)^2 \xrightarrow{P} \mathbb{V}[x_i].$$

7. Therefore, Slutsky's theorem implies that
$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \frac{1}{V[x_i]}\right).$$

2 EIV

Next, consider the errors-in-variables model:

$$\begin{cases} m_i = \beta_0^m + \beta_1^m p_i + \epsilon_i \\ g_i = \beta_0^g + \beta_1^g p_i + \tau_i \\ p_i \sim \text{Bern}(\pi) \\ \epsilon_i, \tau_i \sim N(0,1) \\ p_i \perp\!\!\!\perp \tau_i \perp\!\!\!\perp \epsilon_i \end{cases}$$

Define

$$\hat{p}_i = \mathbb{I}\left(g_i \geq c\right)$$

for some $c > 0$. The thresholding estimator is

$$\hat{\beta} = \frac{\sum_{i=1}^n (\hat{p}_i - \bar{\hat{p}}) m_i}{\sum_{i=1}^n \hat{p}_i^2 - (1/n)\left(\sum_{i=1}^n \hat{p}_i\right)^2}.$$

Our goal is to compute the limiting distribution

$$\sqrt{n}\left(\hat{\beta}_1^m - l\right),$$

where l is the limit (in probability) of $\hat{\beta}_1^m$:

$$\hat{\beta}_1^m \xrightarrow{P} \frac{\beta_1^m \pi\left(\omega - \mathbb{E}[\hat{p}_i]\right)}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} := l.$$

We have that

$$\begin{aligned} \hat{\beta}_1^m - l = \hat{\beta}_1^m - l \frac{\sum_{i=1}^n \hat{p}_i^2 - (1/n)\left(\sum_{i=1}^n \hat{p}_i\right)^2}{\sum_{i=1}^n \hat{p}_i^2 - (1/n)\left(\sum_{i=1}^n \hat{p}_i\right)^2} \\ = \frac{\sum_{i=1}^n (\hat{p}_i - \bar{\hat{p}}) m_i - l\left[\sum_{i=1}^n \hat{p}_i^2 - (1/n)\left(\sum_{i=1}^n \hat{p}_i\right)^2\right]}{\sum_{i=1}^n \hat{p}_i^2 - (1/n)\left(\sum_{i=1}^n \hat{p}_i\right)^2}. \end{aligned}$$

After doing some algebra, the numerator of $\sqrt{n}(\hat{\beta}_1^m - l)$ becomes:

$$\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n}(\hat{p}_i - \bar{\hat{p}})m_i - l\left[\sum_{i=1}^{n}\hat{p}_i^2 - (1/n)\left(\sum_{i=1}^{n}\hat{p}_i\right)^2\right]\right]$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{p}_i - \bar{\hat{p}})(m_i - l\hat{p}_i)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}((\hat{p}_i - \mathbb{E}[\hat{p}_i]) - (\bar{\hat{p}}_i - \mathbb{E}[\hat{p}_i]))\,(m_i - l\hat{p}_i)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{p}_i - \mathbb{E}[\hat{p}_i]\right)(m_i - l\hat{p}_i) - \left(\bar{\hat{p}}_i - \mathbb{E}[\hat{p}_i]\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(m_i - l\hat{p}_i). \quad (1)$$

We evaluate the two sums of (1) separately. First, we have that

$$\left(\bar{\hat{p}} - \mathbb{E}[\hat{p}_i]\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(m_i - l\hat{p}_i)$$

$$= \left(\bar{\hat{p}} - \mathbb{E}[\hat{p}_i]\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left((m_i - l\hat{p}_i) - \mathbb{E}[m_i - l\hat{p}_i] + \mathbb{E}[m_i - l\hat{p}_i]\right)$$

$$= \left(\bar{\hat{p}} - \mathbb{E}[\hat{p}_i]\right)\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}(m_i - l\hat{p}_i) - \mathbb{E}[m_i - l\hat{p}_i]\right) + \mathbb{E}[m_i - l\hat{p}_i]\sqrt{n}(\bar{\hat{p}} - \mathbb{E}[\hat{p}_i]). \quad (2)$$

The first piece of (2) converges in probability to 0, because $\mathbb{V}(m_i - l\hat{p}_i)$ exists. The second piece of (2) converges in distribution to

$$N(0, \mathbb{V}(\hat{p}_i)\mathbb{E}[m_i - l\hat{p}_i]^2)$$

by CLT. Returning to (1), we evaluate the limit of

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{p}_i - \mathbb{E}[\hat{p}_i])(m_i - l\hat{p}_i) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i.$$

First, we compute $\mathbb{E}(X_i)$, which we do in four parts:

1. $$\mathbb{E}\left[\hat{p}_i m_i\right] = \beta_0^m\mathbb{E}[\hat{P}_i] + \beta_1^m\omega\pi = \frac{(1 - \mathbb{E}[\hat{p}_i])(\beta_0^m\mathbb{E}[\hat{p}_i] + \beta_1^m\omega\pi)}{(1 - \mathbb{E}[\hat{p}_i])}$$

2.

$$\mathbb{E}[l\hat{p}_i] = \frac{\beta_1^m \pi(\omega - \mathbb{E}[\hat{p}_i])}{1 - \mathbb{E}[\hat{p}_i]}$$

3.

$$\mathbb{E}[\hat{p}_i]\mathbb{E}[m_i] = \mathbb{E}[\hat{p}_i]\beta_0^m + \mathbb{E}[\hat{p}_i]\pi\beta_1^m = \frac{(1 - \mathbb{E}[\hat{p}_i])(\mathbb{E}[\hat{p}_i]\beta_0^m + \mathbb{E}[\hat{p}_i]\pi\beta_1^m)}{(1 - \mathbb{E}[\hat{p}_i])}$$

4.

$$l\mathbb{E}[\hat{p}_i]^2 = \frac{\mathbb{E}[\hat{p}_i]\beta_1^m \pi\left(\omega - \mathbb{E}[\hat{p}_i]\right)}{(1 - \mathbb{E}[\hat{p}_i])}$$

Adding the numerators of these fractions, we get

$$(1 - \mathbb{E}[\hat{p}_i])\mathbb{E}\left[X_i\right] = \beta_0^m \mathbb{E}[\hat{p}_i] + \beta_1^m \omega\pi - \mathbb{E}[\hat{p}_i]^2 \beta_0^m - \mathbb{E}[\hat{p}_i]\beta_1^m \omega\pi - \beta_1^m \pi\omega + \beta_1^m \pi \mathbb{E}[\hat{p}_i] \\ - \mathbb{E}[\hat{p}_i]\beta_0^m - \mathbb{E}[\hat{p}_i]\pi\beta_1^m + \mathbb{E}[\hat{p}_i]^2 \beta_0^m + \mathbb{E}[\hat{p}_i]^2 \pi\beta_1^m + \mathbb{E}[\hat{p}_i]\beta_1^m \pi\omega - \mathbb{E}[\hat{p}_i]^2 \beta_1^m \pi = 0,$$

and so $\mathbb{E}[X_i] = 0$. Because $\mathbb{E}[X_i] = 0$, we easily can apply CLT to the left side of (1) (after first computing the variance of X_i, which is not too hard).

In conclusion, we can evaluate the distributional limits of the two pieces of (1). However, we cannot compute the limit of the sum, because Slutsky's Theorem does not apply to sums of random variables.