Tim

Limiting variance of a regression coefficient in under random-X

1 Model with intercept

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

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

We estimate β using the standard OLS estimator:

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

where

$$\overline{(x_n^2)} = \frac{1}{n}\sum_{i=1}^n x_i^2$$

and

$$\overline{x_n} = \frac{1}{n}\sum_{i=1}^n x_i.$$

Our goal is to compute $\lim_{n\to\infty} \mathbb{V}(\sqrt{n}\hat{\beta})$. We have by law of total variance that

$$\mathbb{V}(\sqrt{n}\hat{\beta}) = \mathbb{E}\left[\mathbb{V}(\sqrt{n}\hat{\beta}|X)\right] + \mathbb{V}\left[\mathbb{E}(\sqrt{n}\hat{\beta}|X)\right],$$

where $X = [x_1, \ldots, x_n]$ is the vector of $x's$. It is a well-known fact that

$$\mathbb{V}(\hat{\beta}|X) = \frac{1}{n\left(\overline{(x_n^2)} - (\overline{x_n})^2\right)}.$$

Multiplying the above equality by n yields

$$\mathbb{V}(\sqrt{n}\hat{\beta}|X) = \frac{1}{\overline{(x_n^2)} - (\overline{x_n})^2}.$$

Next, because $\hat{\beta}$ is an unbiased estimator of β, we have that

$$\mathbb{E}(\sqrt{n}\hat{\beta}|X) = \sqrt{n}\mathbb{E}(\hat{\beta}|X) = \sqrt{n}\beta.$$

Applying law of total variance,

$$\mathbb{V}(\sqrt{n}\hat{\beta}) = \mathbb{E}\left(\frac{1}{\overline{(x_n^2)} - (\overline{x_n})^2}\right) + n\mathbb{V}(\beta) = \mathbb{E}\left(\frac{1}{\overline{(x_n^2)} - (\overline{x_n})^2}\right).$$

Let the random variable T_n be defined by

$$T_n = \overline{(x_n^2)} - (\overline{x_n})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2.$$

We have by LLN that

$$\{T_n\}_{n=1}^{\infty} \xrightarrow{a.s.} \mathbb{V}(x_i) = \pi(1-\pi).$$

Moreover,

$$T_n \leq \frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq \frac{1}{n}\sum_{i=1}^{n} 1 = 1.$$

Thus, T_n is bounded for all n. By the continuous mapping theorem and bounded convergence theorem,

$$\lim_{n\to\infty} \mathbb{V}(\sqrt{n}\hat{\beta}) \xrightarrow{a.s.} \mathbb{E}\left[\lim_{n\to\infty}\frac{1}{T_n}\right] = \mathbb{E}\left(\frac{1}{\pi(1-\pi)}\right) = \frac{1}{\pi(1-\pi)}.$$

The function $\pi \to 1/(\pi(1-\pi))$ is strictly decreasing over $\pi \in [0, 1/2]$. Moreover, this function blows up at $\pi = 0$.

2 Model without intercept

Consider now the no-intercept model, i.e. $\beta_0 = 0$. Let

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

We have that

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}\left[\mathbb{V}(\hat{\beta}|X)\right] + \mathbb{V}\left[\mathbb{E}(\hat{\beta}|X)\right].$$

Now,

$$\mathbb{V}(\hat{\beta}|X) = \frac{\sum_{i=1}^n \mathbb{V}(x_i y_i | x_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{1}{\sum_{i=1}^n x_i^2}.$$

Next,

$$\mathbb{E}(\hat{\beta}|X) = \frac{\sum_{i=1}^n x_i x_i \beta}{\sum_{i=1}^n x_i^2} = \beta \left(\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} \right) = \beta.$$

Therefore, applying law of total variance,

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}\left(\frac{1}{\sum_{i=1}^n x_i^2} \right),$$

and

$$\mathbb{V}(\sqrt{n}\hat{\beta}) = \mathbb{E}\left(\frac{1}{(1/n)\sum_{i=1}^n x_i^2} \right).$$

By WLLN,

$$(1/n)\sum_{i=1}^n x_i^2 \xrightarrow{a.s.} \mathbb{E}(x_i^2) = \mathbb{E}(x_i) = \pi.$$

Therefore,

$$\lim_{n\to\infty} \mathbb{V}(\sqrt{n}\hat{\beta}) \xrightarrow{a.s.} \mathbb{E}\left[\lim_{n\to\infty} \frac{1}{\sum_{i=1}^n x_i^2} \right] = \frac{1}{\pi}.$$

Note that this variance is different than the variance for the intercept model. (If x_i were a variable such that $\mathbb{E}(x_i) = 0$, then the variances would coincide.)

2.1 Errors-in-variables model without intercept

We derive the limiting variance of the thresholding method in the no-intercept model. Consider the following model:

$$\begin{cases} m_i = \beta_m p_i + \epsilon_i \\ g_i = \beta_g p_i + \tau_i \\ p_i \sim \text{Bern}(\pi) \\ \epsilon_i, \tau_i \sim N(0,1) \\ p_i \perp\!\!\!\perp \tau_i \perp\!\!\!\perp \epsilon_i \end{cases}$$

Define

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

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

$$\hat{\beta} = \frac{\sum_{i=1}^n \hat{p}_i m_i}{\sum_{i=1}^n \hat{p}_i^2}.$$

We have that

$$\mathbb{V}(\hat{\beta}|p, \tau) = \frac{\sum_{i=1}^n \hat{p}_i^2 \mathbb{V}(m_i|\tau_i, p_i)}{\left(\sum_{i=1}^n \hat{p}_i^2\right)^2} = \frac{1}{\sum_{i=1}^n \hat{p}_i^2}.$$

Furthermore,

$$\mathbb{E}(\hat{\beta}|p, \tau) = \frac{\sum_{i=1}^n \hat{p}_i m_i}{\sum_{i=1}^n \hat{p}_i^2} = \frac{\sum_{i=1}^n \hat{p}_i \beta_m p_i}{\sum_{i=1}^n \hat{p}_i^2} = \beta_m \left(\frac{\sum_{i=1}^n \hat{p}_i p_i}{\sum_{i=1}^n \hat{p}_i^2}\right).$$

Thus, by law of total variance,

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}\left(\frac{1}{\sum_{i=1}^n \hat{p}_i^2}\right) + \beta_m^2 \mathbb{V}\left(\frac{\sum_{i=1}^n \hat{p}_i p_i}{\sum_{i=1}^n \hat{p}_i^2}\right).$$

Now,

$$\beta_m^2 \mathbb{V}\left(\frac{\sum_{i=1}^n \hat{p}_i p_i}{\sum_{i=1}^n \hat{p}_i^2}\right) = \frac{\beta_m^2}{\left(\sum_{i=1}^n \hat{p}_i^2\right)^2} \mathbb{V}\left(\sum_{i=1}^n \hat{p}_i p_i\right).$$

Next, $\hat{p}_i p_i$ is a Bernoulli random variable. We therefore can calculate its variance by calculating its mean:

$$\mathbb{E}\left[\hat{p}_i p_i\right] = \mathbb{E}\left(\mathbb{E}\left[\hat{p}_i p_i | p_i\right]\right) = \mathbb{E}\left(p_i \mathbb{P}(\tau_i \geq c - \beta_g p_i)\right) = \pi \mathbb{P}(\tau_i \geq c - \beta_g) = \omega\pi.$$

Thus,

$$\mathbb{V}[\hat{p}_i p_i] = \omega\pi(1 - \omega\pi).$$

Because the p_is are independent, we have

$$\mathbb{V}\left(\sum_{i=1}^n \hat{p}_i p_i\right) = n\omega\pi(1 - \omega\pi).$$

We can rewrite $\mathbb{V}(\hat{\beta})$ as

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}\left(\frac{1}{\sum_{i=1}^n \hat{p}_i^2}\right) + \frac{n\beta_m \omega\pi(1 - \omega\pi)}{\left(\sum_{i=1}^n \hat{p}_i\right)^2}.$$

Multiplying the above by n,

$$\mathbb{V}(\sqrt{n}\hat{\beta}) = \mathbb{E}\left(\frac{1}{(1/n)\sum_{i=1}^{n}\hat{p}_i^2}\right) + \frac{\beta_m \omega\pi(1-\omega\pi)}{\left((1/n)\sum_{i=1}^{n}\hat{p}_i^2\right)^2}.$$

Taking the limit,

$$\lim_{n\to\infty}\mathbb{V}(\sqrt{n}\hat{\beta}) = \frac{1}{\mathbb{E}[\hat{p}_i^2]} + \frac{\beta_m\omega\pi(1-\omega\pi)}{\mathbb{E}[\hat{p}_i^2]^2} = \frac{1}{\mathbb{E}[\hat{p}_i]} + \frac{\beta_m^2\omega\pi(1-\omega\pi)}{\mathbb{E}[\hat{p}_i]^2}.$$

Next, we derive the limit of $\hat{\beta}$ in the no-intercept model. We have that

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

Now,

$$\lim_{n\to\infty}(1/n)\sum_{i=1}^{n}\hat{p}_i^2 = \mathbb{E}\left[\hat{p}_i^2\right] = \mathbb{E}[\hat{p}_i] = \zeta(1-\pi)+\omega\pi.$$

Next,

$$\lim_{n\to\infty}(1/n)\sum_{i=1}\hat{p}_i m_i = \mathbb{E}[\hat{p}_i m_i].$$

We have

$$\mathbb{E}[\hat{p}_i m_i] = \mathbb{E}[\hat{p}_i(\beta_m p_i + \epsilon_i)] = \mathbb{E}[\beta_m\hat{p}_i p_i + \hat{p}_i\epsilon_i] = \beta_m\mathbb{E}[\hat{p}_i p_i] = \beta_m\omega\pi.$$

Thus,

$$\hat{\beta} \xrightarrow{P} \frac{\beta_m\omega\pi}{\zeta(1-\pi)+\omega\pi},$$

where $\zeta = \mathbb{P}(\tau_i \geq c)$ and $\omega = \mathbb{P}(\tau_i \geq c - \beta_g)$.