Simulation

In our simulation analysis, we examine the bias in coefficient estimates and t-statistics. We characterize the bias as a function of the signs and magnitudes (using R-squared) of the correlations between independent variables. We use the following procedure to generate simulation results.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where x_1, x_2, and x_3 are correlated with each other. The objective is to construct variables x_1, x_2, and x_3 that have desired R-squared (R_1^2, R_2^2, and R_{13}^2) as defined below.

- (i) $x_1 \sim N(0,1)$
- (ii) $x_2 = n_1 x_1 + n_3 x_3 + e_2$
- (iii) $x_3 = n_4 x_1 + e_3$

where $e_2 \sim N(0,1)$, $e_3 \sim N(0,1)$, and x_1, e_2, and e_3 are uncorrelated with each other. We normalize x_2 and x_3 to unit normal distribution after identifying n_1, n_3, and n_4.

First, we solve for n_4. R_{13}^2 is the R-squared of regressing x_3 on x_1. From equation (iii) above, we have:

$$R_{13}^2 = \frac{n_4^2}{1 + n_4^2}$$

which implies

$$n_4 = \pm\sqrt{\frac{R_{13}^2}{1 - R_{13}^2}}$$

Note a positive n_4 indicates x_1 and x_3 are positively correlated, and a negative n_4 implies x_1 and x_3 are negatively correlated. We run simulations for both cases.

We normalize $x_3 \sim N(0,1)$, and we get:

$$x_3 = \frac{n_4}{\sqrt{n_4^2 + 1}} x_1 + \frac{1}{\sqrt{n_4^2 + 1}} e_3 \tag{1}$$

Next, we solve for n_1 and n_3. According to the expression of x_3, we can see x_3 consists of two components: one component is

perfectly correlated with x_1, and the other component is orthogonal to x_1. We can then re-write x_2 (see eq (ii)) as

$$\begin{aligned} x_2 &= n_1 x_1 + n_3 (\frac{n_4}{\sqrt{n_4^2+1}} x_1 + \frac{1}{\sqrt{n_4^2+1}} e_3) + e_2 \\ &= (n_1 + n_3 \frac{n_4}{\sqrt{n_4^2+1}}) x_1 + n_3 \frac{1}{\sqrt{n_4^2+1}} e_3 + e_2 \\ &= (n_1 + n_3 b) x_1 + n_3 c e_3 + e_2 \end{aligned}$$

where $b = \frac{n_4}{\sqrt{n_4^2+1}}$ and $c = \frac{1}{\sqrt{n_4^2+1}}$.

We define R_2^2 in our paper as the R-squared of regressing x_2 on the x_3 component that is orthogonal to x_1. Specifically, R_2^2 is the R-squared of regressing x_2 on ce_3:

$$R_2^2 = \frac{(n_3 c)^2}{1 + (n_1 + n_3 b)^2 + (n_3 c)^2}$$

where x_1, e_2, and e_3 are uncorrelated.

We define R_1^2 in our paper as the R-squared of regressing x_2 on x_1:

$$R_1^2 = \frac{(n_1 + n_3 b)^2}{1 + (n_1 + n_3 b)^2 + (n_3 c)^2}$$

From the expressions for R_1^2 and R_2^2, we can solve for n_1 and n_3.

First we obtain

$$\begin{aligned} (n_1 + n_3 b)^2 &= \frac{R_1^2}{1 - R_1^2 - R_2^2} \\ (n_3 c)^2 &= \frac{R_2^2}{1 - R_1^2 - R_2^2} \end{aligned}$$

Then it follows that:

$$\begin{aligned} n_3 &= \pm \sqrt{\frac{R_2^2}{1 - R_1^2 - R_2^2}} / c \\ n_1 &= \pm \sqrt{\frac{R_1^2}{1 - R_1^2 - R_2^2}} - n_3 b \end{aligned}$$

where b and c are defined above.

Finally, we normalize $x_2 \sim N(0,1)$, and generate x_2 as follows:

$$x_2 = \frac{1}{1 + (n_1 + n_3 b)^2 + (n_3 c)^2}(n_1 x_1 + n_3 x_3 + e_2) \tag{2}$$