Replication Exercise -Set Identification, Mres

Jordi Torres

December 8, 2025

1 Setting

Let $Y = \{1, 0\}$, where $Y = 1$ when the individual in my sample has graduated from STEM degree in university and 0 where graduated in non-STEM degree. Let $D = \{1, 0\}$ where $D = 1$ if the individual chooses a STEM track in high-school and $D = 0$ if the individual chooses a non-STEM track in high-school. Let X denote a vector of covariates at the individual level, such as gender, type of high-school attended, minority status.... Finally, let Z denote a set of instruments.
We will make the following assumptions:

1. **ASSUMPTION** (potential outcomes): $Y = Y_1 D + Y_0(1 - D)$, where (Y_1, Y_0) is the distribution of the potential outcomes -which we don't observe- and Y, D are what we observe in the data.
2. **ASSUMPTION**(selection): if $Y_1 > Y_0 \implies D = 1$
3. **ASSUMPTION** (expected outcomes): $\mathbf{E}(Y_1 - Y_0|\mathcal{I}_t) > 0 \implies D = 1$, where $\mathcal{I}_t$ denotes the information set that the individual has at time t.
4. **ASSUMPTION** (SMIV) for any pair $z_2 \geq z_1$ in the support of Z, the conditional distribution of (Y_1, Y_0) given $Z = z_2$ first order stochastically dominates the distribution of (Y_1, Y_0) given $Z = z_1$ [1].

Under ASSUMPTIONS 1,3,4 (here I omit the first theorem, should I test for it too?) , we can construct the following bounds on the joint distribution of potential outcomes:

1. $\mathbf{P}(Y_0 = 1, Y_1 = 0|Z = z) \leq \mathbf{P}(Y = 1, D = 0|Z = z) + \mathbf{P}(Y = 0, D = 1|Z = z)$
2. $\mathbf{P}(Y_0 = 0, Y_1 = 1|Z = z) \leq \mathbf{P}(Y = 1, D = 1|Z = z) + \mathbf{P}(Y = 0, D = 0|Z = z)$
3. $\mathbf{P}(Y_0 = 0, Y_1 = 0|Z = z) \leq \mathbf{P}(Y = 0|Z = z)$

   And, from the joint distribution, combining 1,2 and 3 we can derive conditions for the marginal distribution of outcomes:

4. $\sup_{\tilde{z} \leq z} \mathbf{P}(Y = 1, D = 0|Z = \tilde{z}) \leq \mathbf{E}(Y_0|Z = z)$
5. $\sup_{\tilde{z} \leq z} \mathbf{P}(Y = 1, D = 1|Z = \tilde{z}) \leq \mathbf{E}(Y_1|Z = z)$
6. $\max\{\mathbf{E}(Y_0|Z = z), \mathbf{E}(Y_1|Z = z)\} = \mathbf{E}(y|Z = z)$

These six inequalities defined the set for the joint distributions of outcomes, that are consistent with the model derived under the 3 previous assumptions. In the paper they show that we can test the validity of the assumptions by simply testing that the last equality holds (THEOREM 2), which then simply boils down to testing monothonicity of $\mathbf{E}(Y|Z = z)$ in Z.
We can test monothonicity of $\mathbf{E}(Y|Z = z)$ using the usual methods developed by Chetverikov (2013), Shi et al (2019)
Measure of departure from Roy The second contribution of the paper is to generate a measure of rejection of Roy model behavior, which they call efficiency loss. This measures the extend towards which there are other factors influencing Outcomes besides self-selection based on potential outcomes.

[1] This is very similar to the Manski and Pepper (2000) MIV with the difference that this assumption is stronger. M-P assumed monotonicity of expected potential outcome , i.e $\mathbf{E}(Y_1|Z = z)$, but the assumption that we are doing is in the joint distribution. What we are looking for is instruments that affect both potential outcomes at the same time. In my setting, parental education both affects the probability of being in stem graduation in uni and non-stem graduation, for example.

Definition 1. ***Efficiency loss.*** *For each z in the support of Z, define*

$$el(z) := \mathbb{P}(\max\{Y_0, Y_1\} = 1 \mid Z = z) - \mathbb{P}(Y = 1 \mid Z = z).$$

In the binary case $Y_d \in \{0, 1\}$, note that $\max\{Y_0, Y_1\} = 1$ except when $(Y_0, Y_1) = (0, 0)$. Hence

$$\mathbb{P}(\max\{Y_0, Y_1\} = 1 \mid Z = z) = 1 - \mathbb{P}(Y_0 = 0, Y_1 = 0 \mid Z = z),$$

and we can equivalently write

$$el(z) = \mathbb{P}(Y = 0 \mid Z = z) - \mathbb{P}(Y_0 = 0, Y_1 = 0 \mid Z = z).$$

By construction, $el(z) \geq 0$ for all z, since $\max\{Y_0, Y_1\} \geq Y$ almost surely.

Perfect foresight Roy. Under Assumption 2 (perfect foresight Roy selection), individuals always choose the track that yields the highest realization of Y_d, so

$$Y = \max\{Y_0, Y_1\}.$$

It follows that

$$\mathbb{P}(Y = 0 \mid Z = z) = \mathbb{P}(\max\{Y_0, Y_1\} = 0 \mid Z = z) = \mathbb{P}(Y_0 = 0, Y_1 = 0 \mid Z = z),$$

and therefore

$$el(z) = 0 \quad \text{for all } z.$$

Thus, any positive efficiency loss is evidence against perfect foresight.

Imperfect foresight Roy. Under Assumption 3 (expected outcome maximization) and Assumption 4 (SMIV), we no longer have $Y = \max\{Y_0, Y_1\}$, but the Roy-type selection rule and stochastic monotonicity still impose nontrivial constraints on the joint distribution of (Y_0, Y_1) given $Z = z$. Let

$$L_{\text{IF}}(z) \leq \mathbb{P}(Y_0 = 0, Y_1 = 0 \mid Z = z) \leq U_{\text{IF}}(z)$$

denote the sharp bounds on $\mathbb{P}(Y_0 = 0, Y_1 = 0 \mid Z = z)$ implied by Assumptions 1, 3 and 4 (Mourifié, Henry and Ruiz provide explicit expressions for $L_{\text{IF}}(z)$ and $U_{\text{IF}}(z)$ in terms of observable probabilities). Substituting these inequalities into the definition of $el(z)$ yields

$$\mathbb{P}(Y = 0 \mid Z = z) - U_{\text{IF}}(z) \;\leq\; el(z) \;\leq\; \mathbb{P}(Y = 0 \mid Z = z) - L_{\text{IF}}(z).$$

Thus, under the Roy model with imperfect foresight, efficiency loss is partially identified and lies in a non-degenerate interval that shrinks to $\{0\}$ when the perfect foresight model holds.

SMIV without Roy selection. If we drop the Roy selection rule altogether and only maintain Assumptions 1 and 4, SMIV alone still imposes stochastic monotonicity restrictions on (Y_0, Y_1) across values of Z. These imply a weaker lower bound

$$L_{\text{SMIV}}(z) := \inf_{\tilde{z} \leq z} \mathbb{P}(Y = 0 \mid Z = \tilde{z}) \;\leq\; \mathbb{P}(Y_0 = 0, Y_1 = 0 \mid Z = z),$$

together with the trivial upper bound $\mathbb{P}(Y_0 = 0, Y_1 = 0 \mid Z = z) \leq \mathbb{P}(Y = 0 \mid Z = z)$. Consequently,

$$0 \leq el(z) \leq \mathbb{P}(Y = 0 \mid Z = z) - L_{\text{SMIV}}(z).$$

Compared to the imperfect foresight Roy case, the SMIV-only model yields a larger identified set for $el(z)$, reflecting the weaker behavioral restrictions.
In the empirical analysis, $el(z)$ is thus a functional of z that is only partially identified. I will estimate the bounds on $el(z)$ under the different assumption sets, and use the intersection-bounds inference methods of **?** to construct confidence bands for these bounds and for scalar summaries such as $\sup_z el(z)$.
My plan is the following:

1. restrict the sample for income groups, gender, minority status.
2. For theorems 1, 2 check monothonicity of outcome and argue that we reject or not Roy Model.→ the idea is to find a test with Y, Z being both discrete variables; discrete support of Z, continuous support of Z.
3. If we reject 1 , build el(z) under the assumption of 1,3,5–¿see what is the cost of rejection. If reject imperfect info, build el(z) under the assumption of 1,5
4. Generate bounds and inference based on Chernozhukov, Rosen, Lee: "Intersection Bounds: inference and methods" from Econometrica. Generate bounds for this last measure.

2 Sample

First I compute 10 different subsamples of my original dataset.

3 Test of Monothonicity

Method

Results

4 Bounds on efficiency measure

Method

Results