Replication Exercise -Set Identification, Mres

Jordi Torres

December 14, 2025

1 Introduction

This paper replicates Mourifié, Henry, and Meango (2020, hereafter MHM). In that paper, the authors propose a framework to test the Roy selection model using partial identification techniques. They also show how to construct a measure of departure from the Roy model and use inference methods from Chernozhukov Lee Rosen (2013) on it. MHM apply their methodology to German and Canadian data, focusing on selection into STEM versus non-STEM higher education fields and on how this selection affects labor market outcomes such as wages and job retention. Their approach allows them both to test whether the Roy model is rejected and, when it is, to quantify the magnitude of its failure across regions.

In my replication, I apply the same methodology to the Belgian context. Rather than focusing on sorting into STEM fields in higher education, I study selection into STEM versus non-STEM tracks at the high school level, and I use outcomes in higher education, in particular the probability of graduating, as the relevant outcome. The main goal is to reproduce a table close to Table 5 in MHM (page 3264), which reports estimates of the departure from Roy across values of the instrument. This exercise serves both as a replication of their methodology and as an application to a different institutional setting.

The remainder of the document is organized as follows. First, I describe the economic setting, following MHM, and explain how the measure of departure from the Roy model is constructed. Second, I briefly describe the dataset and the instruments I use. Third, I explain how I do inference on this measure based on the methods proposed by CLR. Finally, I present and interpret the empirical results I find.

2 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... etc. Finally, let Z denote a set of instruments.
We will make the following assumptions, following MHM:

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, MHM show that we can construct the following bounds on the joint distribution of 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.

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, which can be done by using methods such as Chetverikov (2013), Shi et al (2019).

Although in my code I have tried to test for monothonicity too[2], the main effort of my exercise has been to try to replicate el(z) measure and do inference based on CLR. This is the next step:
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 measure measures the extend towards which there are other factors influencing sorting besides potential outcomes.

Definition 1. ***Efficiency loss.*** *For each z in the support of Z, they 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 $Y_d \in \{0, 1\}$,then $\max\{Y_0, Y_1\} = 1$ except when $(Y_0, Y_1) = (0, 0)$. Therefor:

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

and we can also 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.

[2]For example, I have tried to test for monothonicity of $\mathbf{E}(Y|Z = z)$ with z discrete by simply setting the inequalities and using Andrews and Shi and simulating the critical values by bootstrap. Nevertheless, the test was too restrictive and almost never I rejected that Roy model was consistent with the data.

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 imposes 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,

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

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 my replication exercise I have focused in estimating $el(z)$ under assumptions 1-4 only.
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 $\sup_z el(z)$.

3 Sample

The empirical analysis uses a sample of 2,181 belgian students for whom I observe complete high-school trajectory and up to two years in higher-education. For each individual, I observe whether the student chooses a STEM or non-STEM track in the beginning of high school the type of degree the students enroll in. Given that I don't observe the final graduation, I assume that if students are in year 2 in a STEM degree they will eventually graduate and take this as the outcome of interest[3] In addition, the dataset contains rich background information, including parental income, parental education, IQ measures, gender, and other individual characteristics.
In the analysis, the outcome variable Y is an indicator for graduating from a STEM degree in university, while the treatment variable D indicates whether the student chose a STEM track in high school. The underlying idea is that students who select into STEM tracks early in their educational career are likely to do so based on pre-existing ability and expectations about future outcomes.

To explore heterogeneity in the measure of departure from Roy, I construct ten subsamples of the original dataset by binning individuals according to gender and 5 family income groups. The objective is to examine how the efficiency loss $el(z)$ varies across these dimensions.

I consider two instruments. The first is parental education, measured as a discrete variable with three categories, which is also used as an instrument in MHM. The second is the student's IQ score, measured as a continuous variable during compulsory schooling. I argue that IQ affects both potential outcomes Y_1 and Y_0 and therefore satisfies the stochastic monotonicity assumption. Using both a discrete and a continuous instrument allows me to implement the CLR inference procedure in both settings and to assess the sensitivity of the results to the nature of the instrument.

[3] In the Belgian context this is defensible. There are no entry requirements in universities in terms of grades or background. However, because of this, universities impose strong academic requiremenents in year 1 to filter out students with low ability. Having survived year 1 is thus a good predictor of eventual graduation.

4 Bounds on efficiency measure

Method

I implement the intersection-bounds inference procedure of Chernozhukov, Lee, and Rosen (2013, Section 6.1) using a series estimator, adapted to the present setting. The procedure is applied separately to each subsample.

1. **Series estimation of $\mathbb{P}(Y = 0 \mid Z = z)$.** I estimate the conditional probability $\theta(z) = \mathbb{P}(Y = 0 \mid Z = z)$ by a series regression of $\mathbf{1}\{Y = 0\}$ on a polynomial basis in Z. Let

$$p(Z_i) = (1, Z_i, Z_i^2, \ldots, Z_i^{K-1})',$$

and let P be the $n \times K$ matrix stacking $p(Z_i)'$. The coefficient vector is estimated by OLS,

$$\hat{\beta}_n = (P'P)^{-1} P'Y^{(0)},$$

where $Y^{(0)} = \mathbf{1}\{Y = 0\}$. The fitted bounding function is

$$\widehat{\theta}_n(z) = p(z)'\hat{\beta}_n,$$

which is evaluated on a finite grid of values of Z.

2. **Variance estimation.** I compute a consistent estimator $\hat{\Omega}_n$ of the asymptotic variance of $\sqrt{n}(\hat{\beta}_n - \beta_n)$ using a standard sandwich formula, and obtain its symmetric square root $\hat{\Omega}_n^{1/2}$.

3. **Moment inequalities.** For each grid point z_j, the lower bound

$$\inf_{\tilde{z} \le z_j} \theta(\tilde{z})$$

can be written as a finite collection of inequalities

$$\theta(z_j) - \theta(z_i) \ge 0, \qquad i < j.$$

These inequalities define an index set $V_j = \{1, \ldots, j-1\}$. For each $v \in V_j$, I define

$$\hat{\theta}_n(v) = \widehat{\theta}_n(z_j) - \widehat{\theta}_n(z_v), \qquad p_n(v) = p(z_j) - p(z_v).$$

4. **Standardized moments.** For each $v \in V_j$, I compute

$$\hat{g}(v) = p_n(v)'\hat{\Omega}_n^{1/2}, \qquad s_n(v) = \frac{\|\hat{g}(v)\|}{\sqrt{n}}.$$

5. **Simulation and first-stage critical value.** I set $\tilde{\gamma}_n = 1 - 0.1/\log n$ and simulate R draws $Z_r \sim \mathcal{N}(0, I_K)$. The first-stage critical value $\kappa_{n,V_j}(\tilde{\gamma}_n)$ is computed as the $\tilde{\gamma}_n$-quantile of

$$\left\{ \sup_{v \in V_j} \frac{\hat{g}(v)'Z_r}{\|\hat{g}(v)\|} : r = 1, \ldots, R \right\}.$$

6. **Inequality selection.** Following CLR, I construct the reduced index set $\hat{V}_{n,j}$ by retaining inequalities that are close to binding.

7. **Second-stage critical value and lower bound.** Using $\hat{V}_{n,j}$, I compute a second-stage critical value $\kappa_{n,\hat{V}_{n,j}}(p)$ and form the one-sided lower confidence bound

$$\hat{\theta}_{n,0}(p) = \inf_{v \in \hat{V}_{n,j}} \left[\hat{\theta}_n(v) + \kappa_{n,\hat{V}_{n,j}}(p) s_n(v) \right].$$

Repeating this procedure for each grid point yields a lower confidence band for the bounding function over z.[4]

[4] When Z is discrete, several steps simplify: $\theta(z)$ is estimated by sample means, and the series and variance estimation steps are no longer required and are much simpler.

Results

Figures 1 and 2 report the 90% lower confidence bounds on the efficiency loss function $el(z)$ by gender, income group, and values of the instrument Z. Dark regions correspond to values of (z, income) for which the data are consistent with Roy selection, while lighter regions indicate statistically significant departures, in the sense that the lower confidence bound on $el(z)$ is strictly positive.

For women, departures from Roy selection are limited and concentrated primarily in middle and higher income groups. Positive lower bounds appear mainly at low values of z, and in some cases at higher values of z for intermediate income groups. For women from low-income backgrounds, the Roy model cannot be rejected across most of the support of Z. Overall, the magnitude of the departures is small, with lower confidence bounds rarely exceeding 0.02.

For men, violations of Roy selection are more frequent and occur across a broader range of income groups and ability levels. In particular, positive lower bounds are concentrated in middle income groups and appear both at low and high values of z. By contrast, the lowest and highest income groups exhibit fewer departures, suggesting that Roy-type self-selection is more consistent with the data in these groups.

In appendix, tables 3 and 4 I show the results for the discrete instrument parental education. Here we can see that lower bounds on $el(z)$ are higher, particularly for high income and middle SES students. Efficiency departures are also high for girls with high SES and for boys in middle SES distribution: a possible explanation is that girls may be pushed away because of social rules and boys probably go to STEM when they should not...

Compared to MHM, the qualitative structure of the results is similar, in that departures from Roy selection vary systematically with background characteristics and the instrument. However, the magnitude of the departures is substantially smaller in the Belgian data, indicating more limited deviations from Roy selection than those documented in their application.

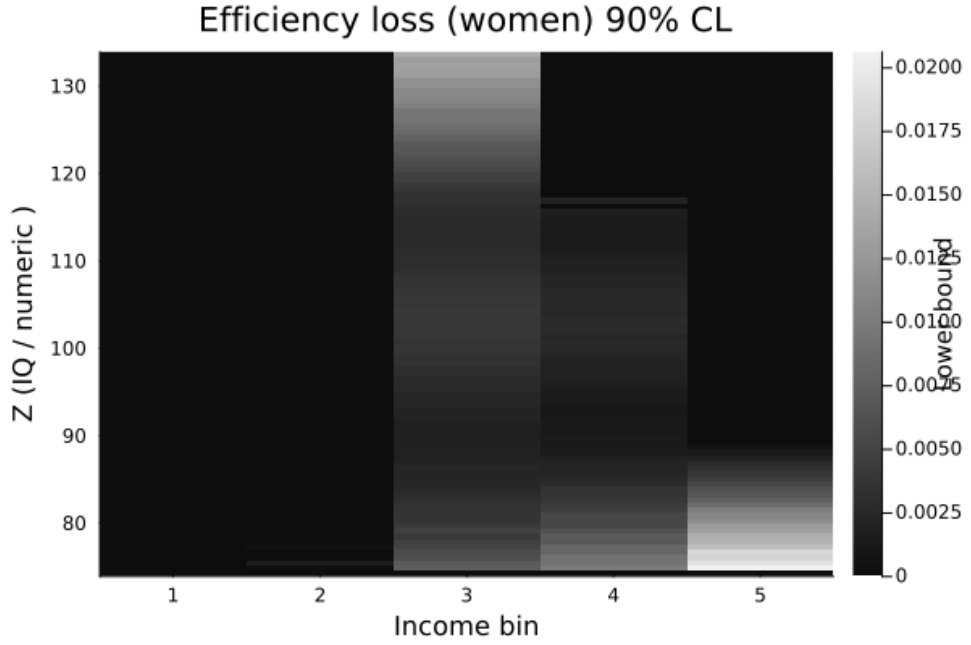

Figure 1: Efficiency loss (women), 90% confidence level

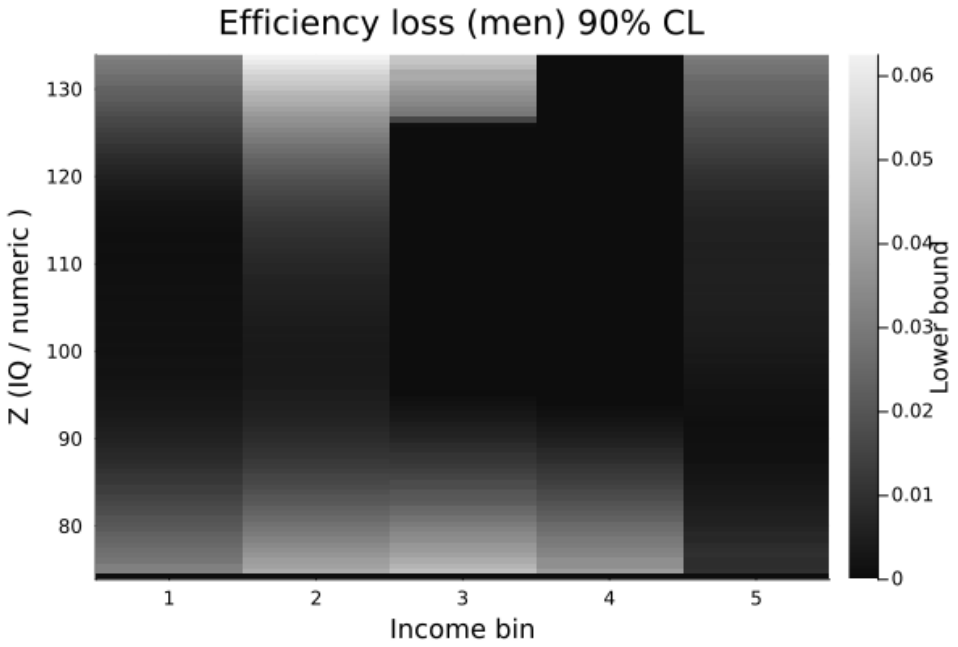

Figure 2: Efficiency loss (men), 90% confidence level

5 Appendix

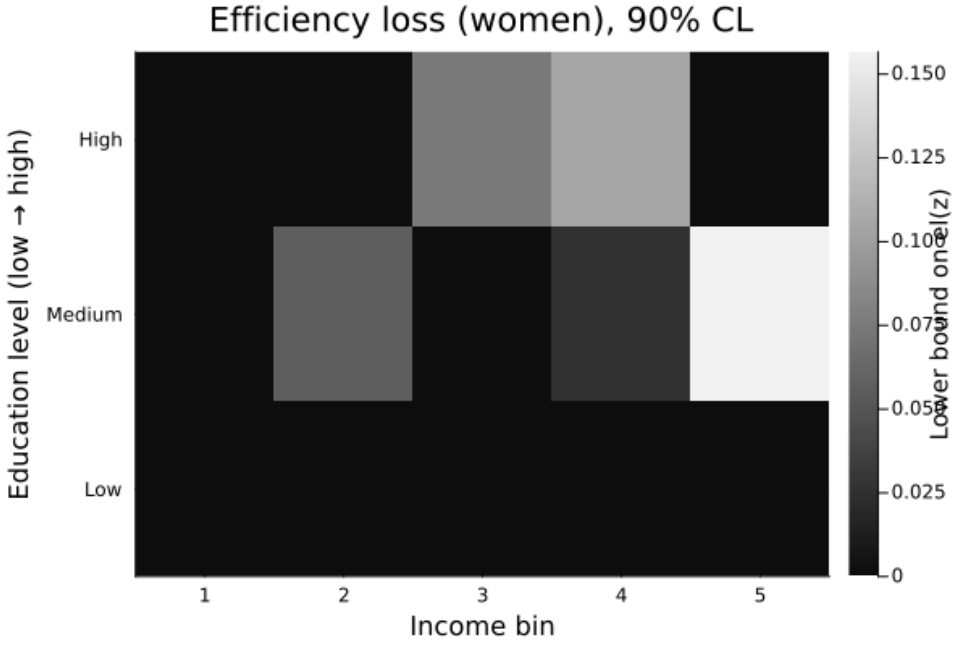

Figure 3: Efficiency loss (women), 90% confidence level

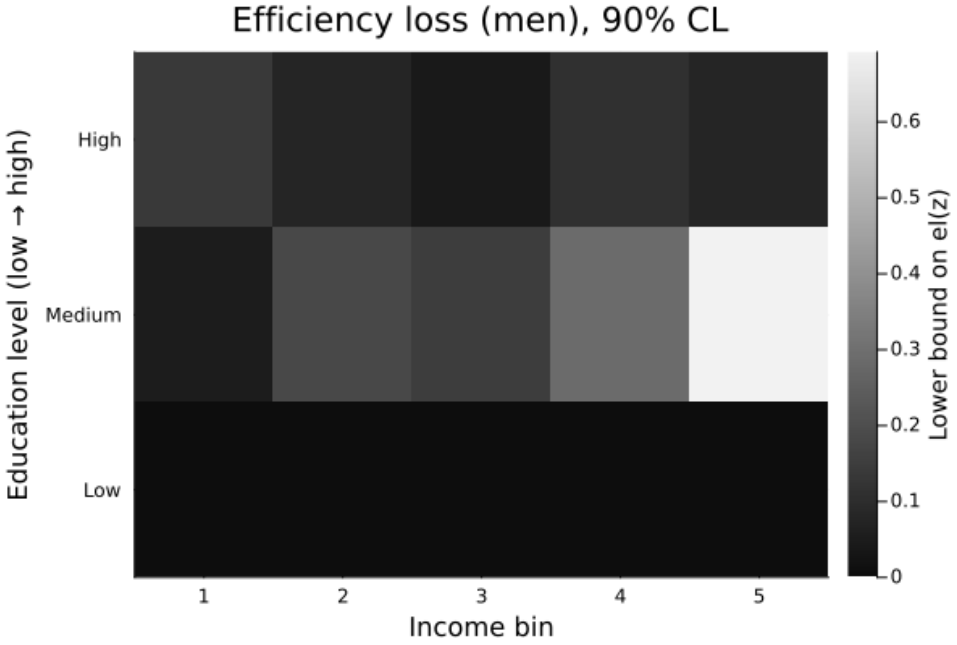

Figure 4: Efficiency loss (men), 90% confidence level