Notes on Identification of the Roy Model and the Generalized Roy Model

James Heckman
University of Chicago

Econ 312, Spring 2022

Roy Model

(Y_0, Y_1) potential outcomes

$I^* = Y_1 - Y_0$ choice **index**

Observe Y_1 if $Y_1 \geq Y_0$.

Observe Y_0 if $Y_1 < Y_0$.

Cannot simultaneously observe Y_0 and Y_1.

We can conduct an identification analysis assuming we know

$$I = \frac{I^*}{\sigma_{Y_1 - Y_0}} = \frac{Y_1 - Y_0}{\sigma_{Y_1 - Y_0}}$$

for each person where $D = \mathbf{1}(I > 0)$.

Why do we know this? Conditions established in the literature

[Source: Cosslett (1983), Manski (1988), Matzkin (1992)]

We observe (Y_0, D) and (Y_1, D). We never observe the full triple (Y_0, Y_1, D) for anyone.

- Under conditions specified in the literature, $F(Y_0, I|X, Z)$ and $F(Y_1, I|X, Z)$ are identified where:

$$Y_0 = \mu_0(X) + U_0 \quad E(Y_0 \mid X) = \mu_0(X) \tag{1}$$

$$Y_1 = \mu_1(X) + U_1 \quad E(Y_1 \mid X) = \mu_1(X) \tag{2}$$

$$I^* = \mu_I(X, Z) + U_I \tag{3}$$

$$I = \frac{\mu_I(X, Z)}{\sigma_{U_I}} + \frac{U_I}{\sigma_{U_I}} \tag{4}$$

- Assume $(X, Z) \perp\!\!\!\perp (U_0, U_1, U_I)$.
- Source: Heckman (1990), Heckman and Honoré (1990)
- The key idea in these papers is "sufficient" variation in Z holding X fixed.

Identifying the Index Choice Probability

- From the left-hand side of

$$\Pr(D = 1|X, Z) = \Pr(\mu_I(X, Z) + U_I \geq 0|X, Z),$$

we can identify the distribution of $\frac{U_I}{\sigma_{U_I}}$, as well as $\frac{\mu_I(X,Z)}{\sigma_{U_I}}$.

- Just invert known f_{U_I} to establish $\frac{\mu_I(X,Z)}{\sigma_I}$. **Prove**.
- This is true under normality or for assumed functional forms for the distribution of $\frac{U_I}{\sigma_{U_I}}$.
- Also, we do not have to assume the distribution of U_I is known or that the functional form of $\mu_I(X, Z)$ is linear, e.g. $\mu_I(X, Z) = X\beta_I + Z\gamma_I$.
- See the conditions in the Matzkin (1992) paper and the survey in Matzkin, 2007, *Handbook of Econometrics*.

- Suppose U_I is symmetric around zero:

$$\Pr(D = 1|X, Z) = \int_{-\mu_I(X,Z)}^{\infty} f(U_I) dU_I$$
$$= 1 - F_{U_I}\left(\frac{\mu_I(X, Z)}{\sigma_{U_I}}\right)$$
$$\Rightarrow F_{U_I}^{-1}[1 - \Pr(D = 1|X, Z)] = \frac{\mu_I(X, Z)}{\sigma_{U_I}}$$

- Can recover $\mu_I(X, Z)$ nonparametrically

- Suppose functional form of distribution unknown?
- To approach this, use the following:

$$\Pr(D=1|X,Z) = \Pr(U_I \geq -\mu_I(X,Z)) = \int_{-\mu_I(X,Z)}^{\infty} f(U_I)\,dU_I \qquad (**)$$

- Suppose $\mu_I(X, Z)$ differentiable in Z.
- Z has 2 (or more) elements.

$$\frac{\frac{\partial \Pr(D=1|X,Z)}{\partial Z_1}}{\frac{\partial \Pr(D=1|X,Z)}{\partial Z_2}} = \frac{\left(\frac{\partial \mu_I(X,Z)}{\partial Z_1}\right) f_{U_I}(\mu_I(X,Z))}{\left(\frac{\partial \mu_I(X,Z)}{\partial Z_2}\right) f_{U_I}(\mu_I(X,Z))} = \frac{\frac{\partial \mu_I(X,Z)}{\partial Z_1}}{\frac{\partial \mu_I(X,Z)}{\partial Z_2}}$$

Example

- Suppose $\mu_I(X, Z) = \gamma Z$

$$\frac{\frac{\partial \mu_I(X,Z)}{\partial Z_1}}{\frac{\partial \mu_I(X,Z)}{\partial Z_2}} = \frac{\gamma_1}{\gamma_2}$$

- Normalize $\gamma_1 = 1$; can identify all the other terms.
- To see what is going on, notice that we can define a set of X, Z such that $P(X, Z)$ is constant, which traces out a P isoquant.

- To identify F_{U_I} non-parametrically requires full support of Z and restrictions on $\mu_I(X, Z)$. See Matzkin (1992).
- A key condition is

$$\text{Support}\left(\frac{\mu_I(X,Z)}{\sigma_{U_I}}\right) \supseteq \text{Support}\left(\frac{U_I}{\sigma_{U_I}}\right)$$

  and other regularity conditions.
- Commonly it is assumed that for a fixed X

$$\text{Support}\left(\frac{\mu_I(X,Z)}{\sigma_{U_I}}\right) = (-\infty, \infty).$$

- This is called "identification at infinity." When we vary Z (for each X) we trace out the full support of $\frac{U_I}{\sigma_{U_I}}$.
- **Problem: Prove this using the first line of (**) realizing that you know $\frac{\mu_I}{\delta_I}$.**

Identifying the Joint Distribution of (Y_0, I)

We know the conditional distribution of Y_0:

$$F(Y_0 \mid D = 0, X, Z) = \Pr(Y_0 \le y_0 \mid \mu_I(X, Z) + U_I \le 0, X, Z)$$

Multiply this by $\Pr(D = 0 \mid X, Z)$:

$$F(Y_0 \mid D = 0, X, Z) \Pr(D = 0 \mid X, Z) = \Pr(Y_0 \le y_0, I^* \le 0 \mid X, Z) \quad (*)$$

We can follow the analysis of Heckman (1990), Heckman and Smith (1998), and Carneiro, Hansen, and Heckman (2003).

Left hand side of (*) is known from the data.

Right hand side:

$$\Pr\left(Y_0 \leq y_0, \frac{U_I}{\sigma_{U_I}} < -\frac{\mu_I(X,Z)}{\sigma_{U_I}} \mid X, Z\right)$$

Since we know $\dfrac{\mu_I(X,Z)}{\sigma_{U_I}}$ from the previous analysis, we can vary it for each fixed X.

- If $\mu_I(X,Z)$ gets small ($\mu_I(X,Z) \to -\infty$), recover the marginal distribution Y and in this limit set we can identify the marginal distribution of

$$Y_0 = \mu_0(X) + U_0 \quad \therefore \quad \text{can identify } \mu_0(X) \text{ in limit.}$$

(See Heckman, 1990, and Heckman and Vytlacil, 2007.)

- More generally, we can form:

$$\Pr\left(U_0 \leq y_0 - \mu_0(X), \frac{U_I}{\sigma_{U_I}} \leq \frac{-\mu_I(X,Z)}{\sigma_{U_I}} \mid X, Z\right)$$

- X and Z can be varied and y_0 is a number.
- We can trace out joint distribution of $\left(U_0, \frac{U_I}{\sigma_{U_I}}\right)$ by varying (y_0, Z) for each fixed X (strictly speaking, varying y_0, Z).

$\therefore$ Recover joint distribution of

$$(Y_0, I) = \left(\mu_0(X) + U_0, \frac{\mu_I(X,Z) + U_I}{\sigma_{U_I}} \right).$$

Three key ingredients.

1. The independence of (U_0, U_I) and (X, Z).
2. The assumption that we can set $\dfrac{\mu_I(X,Z)}{\sigma_{U_I}}$ to be very small (so we get the marginal distribution of Y_0 and hence $\mu_0(X)$).
3. The assumption that $\dfrac{\mu_I(X,Z)}{\sigma_{U_I}}$ can be varied independently of $\mu_0(X)$.

Trace out the joint distribution of $\left(U_0, \frac{U_I}{\sigma_{U_I}} \right)$. Result generalizes easily to the vector case. (Carneiro, Hansen, and Heckman, 2003, IER)

Another way to see this is to write:

$$F(Y_0 \mid D = 0, X, Z) \Pr(D = 0 \mid X, Z)$$

This is a function of $\mu_0(X)$ and $\dfrac{\mu_I(X, Z)}{\sigma_{U_I}}$ (Index sufficiency)

Varying the $\mu_0(X)$ and $\dfrac{\mu_I(X,Z)}{\sigma_{U_I}}$ traces out the distribution of $\left(U_0, \dfrac{U_I}{\sigma_{U_I}}\right)$.

This means effectively that we observe the pairs $\left(\frac{I}{\sigma_{U_I}}, Y_1\right)$ and $\left(\frac{I}{\sigma_{U_I}}, Y_0\right)$.

We never observe the triple $\left(\frac{I}{\sigma_{U_I}}, Y_0, Y_1\right)$.

- Use the intuition that we "know" I.
- We observe

$$F(Y_0 \mid I < 0, X, Z)$$

and

$$F(Y_1 \mid I \geq 0, X, Z)$$

and

$$\Pr(I \geq 0 \mid X, Z)$$

and can construct the joint distributions $F(Y_0, I \mid X, Z)$ and $F(Y_1, I \mid X, Z)$.

Roy Normal Case

Armed with normality (or the nonparametric assumptions in Heckman and Honoré, 1990), we can estimate

$$\text{Cov}(I, Y_1) = \frac{\sigma^2_{Y_1} - \sigma_{Y_1,Y_0}}{\sigma^2_{Y_1} + \sigma^2_{Y_0} - 2\sigma_{Y_1,Y_0}}$$

$$\text{Cov}(I, Y_0) = -\frac{\sigma^2_{Y_0} - \sigma_{Y_1,Y_0}}{\sigma^2_{Y_1} + \sigma^2_{Y_0} - 2\sigma_{Y_1,Y_0}}$$

We know Var Y_1, Var Y_0 (e.g. normal selection model or use limit sets)

$\therefore$ Cov(Y_0, Y_1) is identified (actually over-identified).

This line of argument does not generalize if we add a cost component (C) that is unobserved (or partly so).

The intuition is clear. In the Roy model the decision rule is generated solely by (Y_1, Y_0). Knowing agent choices we observe the relative order (and magnitude) of Y_1 and Y_0.

Thus we get a second valuable piece of information from agent choices. This information is ignored in statistical approaches to program evaluation.

But does this analysis generalize?

Generalized Roy Model

Add cost

$$I = Y_1 - Y_0 - C$$

and assume that we do not directly observe C.

$$\text{Observe } Y_1 \mid I > 0,$$
$$\text{Observe } Y_0 \mid I < 0,$$

and

$$I = \frac{Y_1 - Y_0 - C}{\sqrt{\text{Var}(Y_1 - Y_0 - C)}}.$$

We can identify Var Y_1 and can identify Var Y_0.

But we cannot directly identify $\text{Cov}(Y_0, Y_1)$ which measures comparative advantage.

Notice, however, we can determine if

$$E(Y_1 \mid I > 0) > E(Y_1)$$
$$E(Y_0 \mid I < 0) > E(Y_0)$$

(Are people who work in a sector above average for the sector?)

Carneiro, P., K. Hansen, and J. J. Heckman (2003, May). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review 44*(2), 361–422.

Cosslett, S. R. (1983, May). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica 51*(3), 765–82.

Heckman, J. J. (1990, May). Varieties of selection bias. *American Economic Review 80*(2: Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association), 313–318.

Heckman, J. J. and B. E. Honoré (1990, September). The empirical content of the Roy model. *Econometrica 58*(5), 1121–1149.

Heckman, J. J. and J. A. Smith (1998). Evaluating the welfare state. In S. Strom (Ed.), *Econometrics and Economic Theory in* *the Twentieth Century: The Ragnar Frisch Centennial Symposium*, pp. 241–318. New York: Cambridge University Press.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs, and to forecast their effects in new environments. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, Chapter 71, pp. 4875–5143. Amsterdam: Elsevier B. V.

Manski, C. F. (1988, September). Identification of binary response models. *Journal of the American Statistical Association 83*(403), 729–738.

Matzkin, R. L. (1992, March). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*(2), 239–270.

Matzkin, R. L. (2007). Nonparametric identification. In J. J.

Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B. Amsterdam: Elsevier.