Extensions of The Roy Model To Account For Uncertainty

James J. Heckman, Lance Lochner
Petra Todd
The University of Chicago
University of Western Ontario
University of Pennsylvania

August 7, 2005

1 Estimating Distributions of Returns to Schooling

- Restrict the dependence among the (U_0, U_1, U_S) by factor models or other restrictions Urzua (2005).
- A low dimensional set of random variables generates the dependence across the unobservables.
- Such dimension reduction coupled with use of the choice data and measurements that proxy components of the (U_0, U_1, U_S), provides enough information to identify the joint distribution of (Y_1, Y_0) and of (Y_1, Y_0, S).

- Assume separability between unobservables and observables and that Y_1 and Y_0 are scalars:

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1 \\ Y_0 &= \mu_0(X) + U_0. \end{aligned}$$

$$\begin{aligned} S^* &= \mu_S(Z) + U_S \\ S &= \mathbf{1}(S^* \geq 0). \end{aligned}$$

  Allow any X to be in Z.

- To motivate the approach, assume that (U_0, U_1, U_S) is normally distributed with mean zero and covariance matrix Σ_G ("G" for Generalized Roy).

- If the distributions are normal, they can be fully characterized by means and covariances.

- Under normality, standard results in the selection bias literature show that from data on Y_1 given $S = 1$, and X, and data on Y_0 for $S = 0$ and X, and data on choices of schooling given Z, one can identify $\mu_1(X)$, $\mu_0(X)$ and $\mu_S(Z)$, the latter up to scale σ_S (where $\sigma_S^2 = Var(U_S)$).

- In addition, one can identify the joint densities of $(U_0, U_S/\sigma_S)$ and $(U_1, U_S/\sigma_S)$.

- Without further information, one cannot identify the joint density of $(U_0, U_1, U_S/\sigma_S)$.

- To get the gist of the method underlying recent work, we adopt a factor structure model for the U_0, U_1, U_S.

- For simplicity, we assume a one factor model where θ is the factor that generates dependence across the unobservables:

$$\begin{aligned} U_0 &= \alpha_0\theta + \varepsilon_0 \\ U_1 &= \alpha_1\theta + \varepsilon_1 \\ U_S &= \alpha_S\theta + \varepsilon_S. \end{aligned}$$

  Assume $E(U_0) = 0, E(U_1) = 0, E(U_S) = 0$. $E(\theta) = 0$, $E(\varepsilon_0) = 0$, $E(\varepsilon_1) = 0$ and $E(\varepsilon_S) = 0$.

- We assume that θ is a scalar factor (say unmeasured ability)

- $(\varepsilon_0, \varepsilon_1, \varepsilon_S)$ are independent of θ and of each other.

- Under normality or from the general semiparametric identification analysis, we can identify

$$\begin{aligned} COV(U_0, \frac{U_S}{\sigma_S}) &= \frac{\alpha_0 \alpha_S}{\sigma_S} \sigma_\theta^2 \\ COV(U_1, \frac{U_S}{\sigma_S}) &= \frac{\alpha_1 \alpha_S}{\sigma_S} \sigma_\theta^2 \end{aligned}$$

  where $\sigma_S^2 = Var(\varepsilon_S)$.

- From the ratio of the second covariance to the first we obtain $\frac{\alpha_1}{\alpha_0}$.

- Thus we obtain the sign of the dependence between U_0, U_1 because

$$COV(U_0, U_1) = \alpha_0 \alpha_1 \sigma_\theta^2.$$

  From the ratio, we obtain α_1 if we normalize $\alpha_0 = 1$.

- Without further information, we can only identify the variance of U_S up to scale, which can be normalized to 1.
- Knowledge of the sign of $\dfrac{\alpha_1}{\alpha_0}$ is informative on the sign of the correlation between college and high school skills,

Example 1 Access to a single test score

- Assume access to data on Y_0 given $S = 0$, X, Z; to data on Y_1 given $S = 1$, X, Z;

  Data on S given X, Z.

- Suppose that the analyst also has access to a single test score T

$$T = \mu_T(X) + U_T$$

$$U_T = \alpha_T \theta + \varepsilon_T$$

$$T = \mu_T(X) + \alpha_T \theta + \varepsilon_T,$$

  ε_T is independent of ε_0, ε_1, ε_S and (X, Z).

Identify the mean $\mu_T(X)$ from observations on T and X.

- We pick up three additional covariance terms, conditional on X, Z :

$$\begin{aligned} COV(Y_1, T) &= \alpha_1 \alpha_T \sigma_\theta^2, \\ COV(Y_0, T) &= \alpha_0 \alpha_T \sigma_\theta^2, \\ COV(S^*, T) &= \frac{\alpha_S}{\sigma_S} \alpha_T \sigma_\theta^2. \end{aligned}$$

- To simplify the notation we keep the conditioning on X and Z implicit.
- Normalize the loading on the test score to one ($\alpha_T = 1$).
- No longer necessary to normalize $\alpha_0 = 1$ as in the preceding section.

- From the ratio of the covariance of Y_1 with S^* with the covariance of S^* with T, we obtain the left hand side of

$$\frac{COV\,(Y_1, S^*)}{COV\,(S^*, T)} = \frac{\alpha_1 \alpha_S \sigma_\theta^2}{\alpha_S \alpha_T \sigma_\theta^2} = \alpha_1,$$

  $\alpha_T = 1$ (normalization).

- From the preceding argument without the test score, we obtain α_0 since

$$\frac{COV\,(Y_1, S^*)}{COV\,(Y_0, S^*)} = \frac{\alpha_1 \alpha_S \sigma_\theta^2}{\alpha_0 \alpha_S \sigma_\theta^2} = \frac{\alpha_1}{\alpha_0}.$$

- From knowledge of α_1 and α_0 the normalization for α_T, we obtain σ_θ^2 from $COV\,(Y_1, T)$ or $COV\,(Y_0, T)$.

- We obtain α_S (up to scale σ_S) from $COV\,(S^*, T) = \alpha_S \alpha_T \sigma_\theta^2$ since we know α_T $(= 1)$ and σ_θ^2. The model is overidentified.

- Observe that if we write the latent variable determining schooling choices as:

$$S^* = Y_1 - Y_0 - C,$$

$$C = \mu_C(Z) + U_C$$

$$U_C = \alpha_C \theta + \varepsilon_C,$$

- ε_C is independent of θ and the other ε's. $E(U_C) = 0$ and U_C is independent of (X, Z).

$$\begin{aligned} \alpha_S &= \alpha_1 - \alpha_0 - \alpha_C \\ \varepsilon_S &= \varepsilon_1 - \varepsilon_0 - \varepsilon_C \\ Var(\varepsilon_S) &= Var(\varepsilon_1) + Var(\varepsilon_0) + Var(\varepsilon_C). \end{aligned}$$

- Identification of α_0,α_1 and α_S implies identification of α_C.

- Identification of the variance of ε_S implies identification of the variance of ε_C since the variances of ε_1 and ε_0 are known.

- Observe further that the scale σ_{U_S} is identified if there are variables in X but not in Z (Heckman,1976,1979; Heckman and Robb, 1985, 1986; Willis and Rosen, 1979).

From the variance of T given X, we obtain $Var\left(\varepsilon_T\right)$ since we know $Var(T)$ (conditional on X) and we know $\alpha_T^2\sigma_\theta^2$:

$$Var(T) - \alpha_T^2\sigma_\theta^2 = \sigma_{\varepsilon_T}^2.$$

Normality not essential.

Possible to nonparametrically identify the distributions of θ, ε_0, ε_1, ε_S and ε_T.

Example 2 Two (or more) periods of panel data on earnings

Suppose that for each person we have two periods of earnings data in one counterfactual state or the other.

$$\begin{aligned} Y_{1t} &= \mu_{1t}(X) + \alpha_{1t}\theta + \varepsilon_{1t} \qquad t = 1, 2 \\ Y_{0t} &= \mu_{0t}(X) + \alpha_{0t}\theta + \varepsilon_{0t} \qquad t = 1, 2. \end{aligned}$$

Observe one or the other lifecycle stream of earnings for each person, but never both streams for the same person.
Thus in terms of the index

$$S^* = (Y_{12} + Y_{11}) - (Y_{02} + Y_{01}) - C$$

$$S = 1(S^* \geq 0) \text{ and } C \text{ is cost.}$$

Under normality, application of the standard normal selection model allows us to identify

$$\mu_{1t}(X) \text{ for } t = 1, 2,$$
$$\mu_{0t}(X) \text{ for } t = 1, 2,$$
$$\mu_{11}(X) + \mu_{12}(X) - \mu_{01}(X) - \mu_{02}(X) - \mu_C(X),$$
$$\sigma_S \text{ where } U_S = \varepsilon_{11} + \varepsilon_{12} - \varepsilon_{01} - \varepsilon_{02} - \varepsilon_C.$$

- Can recover the scale if there are variables in $(\mu_{11}(X) + \mu_{12}(X) - (\mu_{01}(X) + \mu_{02}(X)))$ not in $\mu_C(Z)$.

  Assume that this condition holds.

- From normality,can recover the joint distributions of (S^*, Y_{11}, Y_{12}) and (S^*, Y_{01}, Y_{02}) but not directly the joint distribution of $(S^*, Y_{11}, Y_{12}, Y_{01}, Y_{02})$.

- Thus, conditioning on X and Z we can recover the joint distribution of (U_S, U_{01}, U_{02}) and (U_S, U_{11}, U_{12}) but apparently not that of $(U_S, U_{01}, U_{02}, U_{11}, U_{12})$.

However, under our factor structure assumptions this joint distribution can be recovered as we next show.

From the available data, we can identify the following covariances:

$$
\begin{aligned}
COV(U_S, U_{12}) &= (\alpha_{12} + \alpha_{11} - \alpha_{02} - \alpha_{01} - \alpha_C)\alpha_{12}\sigma_\theta^2 \\
COV(U_S, U_{11}) &= (\alpha_{12} + \alpha_{11} - \alpha_{02} - \alpha_{01} - \alpha_C)\alpha_{11}\sigma_\theta^2 \\
COV(U_S, U_{01}) &= (\alpha_{12} + \alpha_{11} - \alpha_{02} - \alpha_{01} - \alpha_C)\alpha_{01}\sigma_\theta^2 \\
COV(U_S, U_{02}) &= (\alpha_{12} + \alpha_{11} - \alpha_{02} - \alpha_{01} - \alpha_C)\alpha_{02}\sigma_\theta^2 \\
COV(U_{11}, U_{12}) &= \alpha_{11}\alpha_{12}\sigma_\theta^2 \\
COV(U_{01}, U_{02}) &= \alpha_{01}\alpha_{02}\sigma_\theta^2.
\end{aligned}
$$

- If we normalize $\alpha_{01} = 1$ we can form the ratios

$$\frac{COV(U_S, U_{12})}{COV(U_S, U_{01})} = \alpha_{12} \qquad \frac{COV(U_S, U_{11})}{COV(U_S, U_{01})} = \alpha_{11}$$

$$\frac{COV(U_S, U_{02})}{COV(U_S, U_{01})} = \alpha_{02}.$$

- From these coefficients and the remaining covariances, we identify σ_θ^2 using $COV(U_{11}, U_{12})$ and/or $COV(U_{01}, U_{02})$.

- Thus if the factor loadings are nonzero,

$$\frac{COV(U_{11}, U_{12})}{\alpha_{11}\alpha_{12}} = \sigma_\theta^2 \text{ and } \frac{COV(U_{01}, U_{02})}{\alpha_{01}\alpha_{02}} = \sigma_\theta^2$$

- We can recover σ_θ^2 (since we know $\alpha_{11}\alpha_{12}$ and $\alpha_{01}\alpha_{02}$) from $COV(U_{11}, U_{12})$ and $COV(U_{01}, U_{02})$.

- We can also recover α_C since we know

$$\sigma_\theta^2,\ \alpha_{12}+\alpha_{11}-\alpha_{02}-\alpha_{01}-\alpha_C \ \text{ and } \ \alpha_{11},\alpha_{12},\alpha_{01,}\alpha_{02}.$$

- We can form (conditional on X)

$$\begin{aligned} COV(Y_{11},Y_{01}) &= \alpha_{11}\alpha_{01}\sigma_\theta^2 \\ COV(Y_{12},Y_{01}) &= \alpha_{12}\alpha_{01}\sigma_\theta^2 \\ COV(Y_{11},Y_{02}) &= \alpha_{11}\alpha_{02}\sigma_\theta^2 \\ COV(Y_{12},Y_{02}) &= \alpha_{12}\alpha_{02}\sigma_\theta^2. \end{aligned}$$

- Can identify the joint distribution of $(Y_{01},Y_{02},Y_{11},Y_{12},C)$ since we can identify $\mu_C(Z)$ from the schooling choice equation since we know $\mu_{01}(X),\mu_{02}(X),\mu_{11}(X),\mu_{12}(X)$.

- As in Example 1, this analysis can be generalized.

- The key idea to constructing joint distributions of counterfactuals using the analysis of Cunha, Heckman and Navarro (2005a,b,c,d) is *not* the factor structure for unobservables although it is convenient.

- The motivating idea is the assumption that a low dimensional set of random variables generates the dependence across outcomes.

2 *Ex Ante* and *Ex Post* Returns: Distinguishing Heterogeneity from Uncertainty

2.1 A Generalized Roy Model

Let S_i denote different schooling levels. $S_i = 0$ denotes choice of the high school sector, and $S_i = 1$ denotes choice of the college sector.

$Y_{1,i}$ is the *ex post* present value of earnings in the college sector, discounted over horizon T,

$$Y_{1,i} = \sum_{t=0}^{T} \frac{Y_{1,i,t}}{(1+r)^t},$$

$Y_{0,i}$ is the *ex post* present value,

$$Y_{0,i} = \sum_{t=0}^{T} \frac{Y_{0,i,t}}{(1+r)^t}$$

r is the one-period risk-free interest rate. $Y_{1,i}$ and $Y_{0,i}$ can be constructed from time series of *ex post* potential earnings streams in the two states: $(Y_{0,i,0}, \ldots, Y_{0,i,T})$ for high school and $(Y_{1,i,0}, \ldots, Y_{1,i,T})$ for college.

The variables $Y_{1,i}, Y_{0,i}$, and C_i are *ex post.* Under a complete markets assumption with all risks diversifiable (so that there is risk-neutral pricing)

$$S_i = \begin{cases} 1, \text{ if } E\left(Y_{1,i} - Y_{0,i} - C_i \mid \mathcal{I}_{i,0}\right) \geq 0 \\ 0, \text{ otherwise.} \end{cases} \tag{2.1}$$

Under perfect foresight, the postulated information set would include $Y_{1,i}, Y_{0,i}$, and C_i.

Decision rule is more complicated in the absence of full risk diversifiability

2.2 Identifying Information Sets in the Card Model

Decompose the "returns" coefficient or the gross gains from schooling in an earnings.

- Write discounted lifetime earnings of person i as

$$Y_i = \alpha + \rho_i S_i + U_i, \tag{2.2}$$

  ρ_i is the person-specific *ex post* return, S_i is years of schooling U_i is a mean zero unobservable.

- Decompose ρ_i into two components $\rho_i = \eta_i + \nu_i$, η_i is a component known to the agent ν_i is revealed after the choice is made.

- Schooling choices $S_i = \lambda(\eta_i, Z_i, \tau_i)$, Z_i are other observed determinants.

- τ_i represents additional factors unobserved by the analyst but known to the agent.

- If η_i is known to the agent and acted on, it enters the schooling choice equation.Otherwise it does not.

- ν_i and any measurement errors in $Y_{1,i}$ or $Y_{0,i}$ should not be determinants of schooling choices.

- Suppose that the model for schooling can be written in linear in parameters form, as in the Card:

$$S_i = \lambda_0 + \lambda_1 \eta_i + \lambda_2 \nu_i + \lambda_3 Z_i + \tau_i, \qquad (2.3)$$

  τ_i has mean zero and is assumed to be independent of Z_i.

- Z_i and the τ_i proxy costs

As a simple example, suppose that we observe the cost of funds, r_i, and assume $r_i \perp\!\!\!\perp (\rho_i, \alpha_i)$.

This assumes that the costs of schooling are independent of the "return" ρ_i and the payment to raw ability, α_i.

We can establish identification of $\bar{\rho}$. (If there are observed regressors X determining the mean of $\bar{\rho}$, we identify $\bar{\rho}(X)$, the conditional mean of ρ_i).

- Suppose that agents do not know ρ_i but know $E(\rho_i) = \bar{\rho}$.
- If agents act on this expected return to schooling, decisions are given by

$$S_i = \frac{\bar{\rho} - r_i}{k}$$

ex post earnings after schooling are

$$Y_i = \bar{\alpha} + \bar{\rho} S_i + \{(\alpha_i - \bar{\alpha}) + (\rho_i - \bar{\rho}) S_i\}.$$

- Observe that a regression of S_i on r_i identifies k, $\bar{\rho}$.

In this case,

$$COV\left(Y,S\right)=\bar{\rho}Var\left(S\right)$$

$\left(\rho_i-\bar{\rho}\right)$ is independent of S_i.

Note further that $S_i \perp\!\!\!\perp \left[\left(\alpha_i-\bar{\alpha}\right),\left(\rho_i-\bar{\rho}\right)S_i\right]$.

If, on the other hand, agents know ρ_i, OLS breaks down for identifying $\bar{\rho}$ because ρ_i is correlated with S_i.

We can identify $\bar{\rho}$ and the distribution of ρ_i using the method of instrumental variables.

In this case

$$COV\left(Y_i, S\right) = \bar{\rho} Var\left(S\right) + COV\left(S, \left(\rho - \bar{\rho}\right) S\right).$$

We observe S, can identify $\bar{\rho}$ and can construct $(\rho - \bar{\rho})$ for each S. Can form both terms on the right hand side.

$$S_i = \frac{\rho_i - r_i}{k}$$

We can identify k

$$\text{Can solve } \rho_i = r_i + kS_i$$

Under the assumption that agents do not know ρ but forecast it by $\bar{\rho}$, ρ is independent of S so we can test for independence directly.

In this case the second term on the right hand side is zero and does not contribute to the explanation of $COV\left(\ln y, S\right)$.

Durbin (1954) – Wu (1978) – Hausman (1978) test can be used to compare the OLS and IV estimates, which should be the same under the model that assumes that ρ_i is not known at the time schooling decisions are made and that agents base their choice of schooling on $E(\rho_i) = \bar{\rho}$.

If we add selection bias to the Card model (so $E(\alpha \mid S)$ depends on S), we can identify $\bar{\rho}$ by IV. OLS is no longer consistent even if, in making their schooling decisions, agents forecast ρ_i using $\bar{\rho}$.

Thus the Durbin-Wu-Hausman test is not helpful in assessing what is in the agent's information set.