14.662 Recitation 7

The Roy Model, isoLATEing, and Kirkebøen, Leuven, and Mogstad (2014)

Peter Hull

Spring 2015

Selection: an Applied Microeconomist's Best Friend

- Life (and data) is all about *choices*
  - E.g. schooling (labor), location (urban), insurance (PF), goods (IO)
- How can a dataset of zeros and ones tell us about meaningful latent economic parameters?
- Natural starting point: agents select optimally on potential gains
  - Now obvious, but wasn't always: longstanding belief that job choice "developed by the process of historical accident" (Roy, 1951)
- With enough structure, link from observed to latent is straightforward (e.g. Roy (1951), Heckman (1979), Borjas (1987))
  - Nature of selection characterized by small set of parameters
- Still a lot to do on relaxing structure while staying tractable
  - Recent attempts: Kirkebøen, Leuven, and Mogstad (2014), Hull (2015)

Borjas' (1987) Roy Notation and Setup

- Potential wages for individual i with schooling level $j \in \{0,1\}$:

$$\begin{aligned} w_{ij} &= E[w_{ij}] + (w_{ij} - E[w_{ij}]) \\ &\equiv \mu_j + \varepsilon_{ij} \end{aligned}$$

- Residuals distributed by

$$\begin{bmatrix} \varepsilon_{i0} \\ \varepsilon_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right)$$

- Individual i chooses schooling $j = 1$ iff

$$w_{i1} - w_{i0} > c$$
$$\underbrace{\mu_1 - \mu_0 - c}_{\equiv z} > \underbrace{\varepsilon_{i0} - \varepsilon_{i1}}_{\equiv \nu_i}$$

  Where c denotes relative cost (assume constant for now)

- Question: what is $E[w_{ij} | z > \nu_i]$ for each group?

Some Essential Normal Facts

1. Law of Iterated Expectations (not just normals): for nonrandom $f(\cdot)$,

$$E[Y|f(X)] = E[E[Y|X]|f(X)]$$

2. Linear Conditional Expectations: if X and Y are jointly normal

$$E[Y|X=x] = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x-\mu_X)$$

3. Inverse Mills Ratio: if $X \sim N(\mu, \sigma^2)$, k constant

$$E[X|X>k] = \mu + \sigma \frac{\phi\left(\frac{k-\mu}{\sigma}\right)}{1-\Phi\left(\frac{k-\mu}{\sigma}\right)}$$

$$E[X|X<k] = \mu - \sigma \frac{\phi\left(\frac{k-\mu}{\sigma}\right)}{\Phi\left(\frac{k-\mu}{\sigma}\right)}$$

Key to remembering: $E[X|X<k]$ should be smaller than $E[X]$

Solving Roy

Note first that

$$\begin{bmatrix} \varepsilon_{i0} \\ \nu_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_0^2 - \sigma_{01} \\ \sigma_0^2 - \sigma_{01} & \sigma_0^2 + \sigma_1^2 - 2\sigma_{01} \end{bmatrix} \right)$$

By Fact #1,

$$\begin{aligned} E[w_{i0}|z > \nu_i] &= \mu_0 + E[\varepsilon_{i0}|z > \nu_i] \\ &= \mu_0 + E[E[\varepsilon_{i0}|\nu_i]|z > \nu_i] \end{aligned}$$

By Fact #2,

$$E[\varepsilon_{i0}|\nu_i] = \frac{\sigma_0^2 - \sigma_{01}}{\sigma_\nu^2}\nu_i, \text{ where } \sigma_\nu^2 \equiv \sigma_0^2 + \sigma_1^2 - 2\sigma_{01}$$

So:

$$E[w_{i0}|z > \nu_i] = \mu_0 + \frac{\sigma_0^2 - \sigma_{01}}{\sigma_\nu^2} E[\nu_i|\nu_i < z]$$

Solving Roy (cont.)

By Fact #3,

$$E[w_{i0}|i \text{ chooses } 1] = \mu_0 + \frac{\sigma_0^2 - \sigma_{01}}{\sigma_v^2} E[v_i|v_i < z]$$
$$= \mu_0 + \frac{\sigma_{01} - \sigma_0^2}{\sigma_v} \frac{\phi(z/\sigma_v)}{\Phi(z/\sigma_v)}$$
$$= \mu_0 + \left(\rho_{01} - \frac{\sigma_0}{\sigma_1}\right) \frac{\sigma_0 \sigma_1}{\sigma_v} \frac{\phi(z/\sigma_v)}{\Phi(z/\sigma_v)}$$

The same steps give us

$$E[w_{i1}|i \text{ chooses } 1] = \mu_1 + \left(\frac{\sigma_1}{\sigma_0} - \rho_{01}\right) \frac{\sigma_0 \sigma_1}{\sigma_v} \frac{\phi(z/\sigma_v)}{\Phi(z/\sigma_v)}$$

When are *observed* $j = 1$ workers "above average"?

Positive and Negative Roy Selection

- Positive selection (avg. $j = 1$ wage "above avg." in both groups):

$$\rho_{01} - \frac{\sigma_0}{\sigma_1} > 0 \text{ and } \frac{\sigma_1}{\sigma_0} - \rho_{01} > 0$$

$$\implies \rho_{01} \in \left[\frac{\sigma_0}{\sigma_1}, \frac{\sigma_1}{\sigma_0}\right]$$

$\implies$ Distribution of productivity with schooling more unequal

- Negative selection (avg. $j = 1$ wage "below avg." in both sectors):

$$\rho_{01} - \frac{\sigma_0}{\sigma_1} < 0 \text{ and } \frac{\sigma_1}{\sigma_0} - \rho_{01} < 0$$

$$\implies \rho_{01} \in \left[\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1}\right]$$

$\implies$ Distribution of productivity without schooling more unequal

- Also can have "refugee selection," where $\rho_{01} < \min\left[\frac{\sigma_0}{\sigma_1}, \frac{\sigma_1}{\sigma_0}\right]$ (but can't have the other case, where $\rho_{01} > \max\left[\frac{\sigma_0}{\sigma_1}, \frac{\sigma_1}{\sigma_0}\right] \geq 1$)

Bringing Roy to Data

- Let $D_i = 1$ if i selects $j = 1$. What does OLS of w_i on D_i give?

$$E[w_i|D_i = 0] = \mu_0 + \frac{\sigma_{01} - \sigma_0^2}{\sigma_v} \frac{\phi(z/\sigma_v)}{\Phi(z/\sigma_v)}$$

$$E[w_i|D_i = 1] = \mu_1 + \frac{\sigma_{01} - \sigma_1^2}{\sigma_v} \frac{\phi(-z/\sigma_v)}{\Phi(-z/\sigma_v)}$$

$$E[w_i|D_i = 1] - E[w_i|D_i = 0] = \underbrace{\mu_1 - \mu_0}_{\text{"treatment effect"}} + (\text{selection bias})$$

- Suppose costs are random: $c_i \in \{0,1\}, c_i \perp\!\!\!\perp (\varepsilon_{i1} - \varepsilon_{i0})$:

$$w_i = \mu_0 + (\mu_1 - \mu_0 + \varepsilon_{i1} - \varepsilon_{i0})D_i + \varepsilon_{i0}$$

$$D_i = \mathbf{1}\{\mu_1 - \mu_0 + \varepsilon_{i1} - \varepsilon_{i0} > c_i\}$$

- Then IV gives LATE; with Roy selection:

$$\underbrace{E[\mu_1 - \mu_0 + \varepsilon_{i1} - \varepsilon_{i0}|0 < \varepsilon_{i1} - \varepsilon_{i0} - (\mu_1 - \mu_0) \leq 1]}_{LATE} \neq \underbrace{\mu_1 - \mu_0}_{ATE}$$

Estimation with Multi-Armed Roy

- W/Roy + unrestricted heterogeneity, valid instrument isn't "enough"
- Problem even worse with many sectors; suppose:

$$w_i = \mu_0 + (w_{ia} - w_{i0})A_i + (w_{ib} - w_{i0})B_i + \varepsilon_{i0}$$

- With binary, independent Z_i that reduces cost of sector a, IV identifies

$$\begin{aligned} & E[w_{ai} - w_{\neg ai} | A_{1i} > A_{0i}] \\ = & E[w_{ai} - w_{0i} | A_{1i} > A_{0i}, B_{0i} = 0] P(B_{0i} = 0 | A_{1i} > A_{0i}) \\ & + E[w_{ai} - w_{bi} | A_{1i} > A_{0i}, B_{0i} = 1] P(B_{0i} = 1 | A_{1i} > A_{0i}) \end{aligned}$$

  weighted average across compliers with fallback b and with fallback 0

- Heckman et al. (2006), Heckman and Urzua (2010): unordered treatment and Roy selection demands a parametric model

isoLATEing: a semi-parametric solution

- Want to deconvolute $E[w_{ai} - w_{\neg ai} | A_{1i} > A_{0i}]$ into its two causal parts
- Can identify $\omega \equiv P(B_{0i} = 1 | A_{1i} > A_{0i})$: just the first stage of B_i on Z_i
- If you can split the data into two parts ("strata") where ω differs but $E[w_{ai} - w_{0i} | A_{1i} > A_{0i}, B_{0i} = j]$ doesn't, can solve out ("isoLATE")
- It turns out (see Hull, 2015) two-endogenous variable IV can automate this deconvolution (and give SEs for free!)
- Problem: if $E[w_{ai} - w_{0i} | A_{1i} > A_{0i}, B_{0i} = j]$ also varies across strata (as you'd expect with Roy selection and ω varying), isoLATE is biased
- Possible solution (work in progress!) assume no Roy selection conditional on rich enough covariates X_i, weight cond. IV over X_i
    - Similar to Angrist and Fernandez-Val (2013) solution to LATE != ATE, Angrist and Rokkanen (2016) solution to RD extrapolation

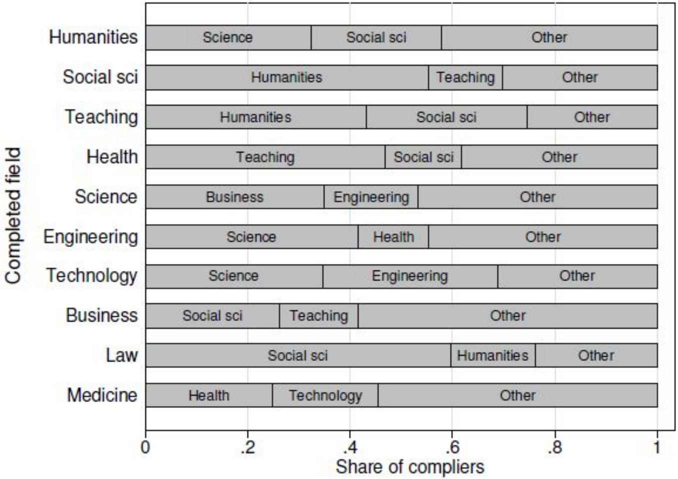
KLM (2014): a data-driven solution

- Kirkebøen, Leuven, and Mogstad (2014) have data on centralized post-secondary admissions and earnings in Norway
  - Interested in estimating the returns to fields and selection patterns
- Note that when $(A_{1i} = A_{0i} = 0) \implies (B_{1i} = B_{0i})$, IV conditional on $(A_{0i} = B_{0i}) = 0$ identifies $E[w_{ai} - w_{0i} | A_{1i} > A_{0i}, B_{0i} = 0]$
  - "Application score" running variable for assignment into ranked fields
  - Sequential dictatorship assignment: truth-telling a dominant strategy
  - Observe completed field/education and earnings
- Assume ranking reveals potential behavior (plausible? Could test); run fuzzy RD for each "next-best" field k:

$$y = \sum_{j \neq k} \beta_{jk} d_j + x' \gamma_k + \lambda_{jk} + \varepsilon$$

$$d_j = \sum_{j \neq k} \pi_{jk} z_j + x' \psi_{jk} + \eta_{jk} + u, \; \forall j \neq k$$

KLM (2014) First Stages



KLM (2014) IV Estimates

Table 4. 2SLS estimates of the payoffs to field of study (USD 1,000)

34

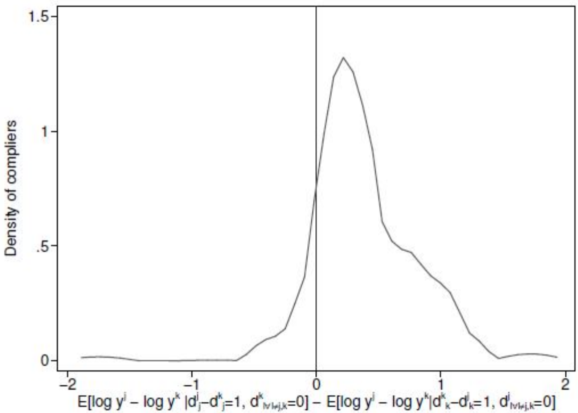
	Next best alternative (k):								
	Humanities	Soc Science	Teaching	Health	Science	Engineering	Technology	Business	Law
Completed field (j):									
Humanities		21.38*	-4.72	-22.93*	4.97	-38.51**	6.87	-42.21**	-156.33
		(10.97)	(9.85)	(12.12)	(11.86)	(14.72)	(48.29)	(10.56)	(437.28)
Social Science	18.72**		9.84	-10.82	55.46**	-55.36**	-110.38	-28.37**	-76.07
	(6.73)		(11.55)	(13.00)	(21.45)	(20.60)	(102.97)	(10.66)	(86.42)
Teaching	22.25**	31.37**		1.82	23.46**	-33.94**	-35.32	-21.08**	22.78
	(4.96)	(7.88)		(6.55)	(9.45)	(12.54)	(37.07)	(7.12)	(127.87)
Health	18.75**	30.69**	7.72**		28.87**	-27.87**	-43.38**	-17.39**	-55.19
	(6.25)	(7.56)	(2.82)		(7.64)	(10.35)	(20.84)	(3.97)	(97.68)
Science	53.71**	69.59**	38.58**	29.63**		-2.21	16.81	-4.92	148.26
	(18.37)	(22.36)	(14.20)	(11.53)		(14.60)	(18.07)	(10.51)	(276.20)
Engineering	59.81	-5.53	75.24**	0.16	52.35**		-46.00	-13.03	-57.66
	(50.59)	(58.17)	(37.50)	(16.36)	(20.98)		(43.89)	(23.70)	(166.60)
Technology	41.87**	58.69**	22.08*	32.45**	68.07**	-5.56		7.03	-53.07
	(10.84)	(10.09)	(12.44)	(10.09)	(9.63)	(11.95)		(9.49)	(147.53)
Business	48.13**	61.93**	31.02**	30.22**	58.01**	-3.42	28.54*		3.53
	(11.25)	(12.03)	(8.78)	(10.86)	(10.48)	(12.61)	(15.61)		(83.04)
Law	46.34**	55.62**	36.60**	21.49*	40.07**	-27.53	-15.55	-1.36	
	(7.16)	(8.34)	(11.56)	(11.46)	(9.68)	(18.29)	(17.96)	(8.66)	
Medicine	83.34**	79.39**	62.62**	45.57**	81.31**	21.07	40.07**	23.34**	14.82
	(9.76)	(10.65)	(9.02)	(7.01)	(9.71)	(20.67)	(11.72)	(8.79)	(83.61)
Female	-7.00**	-6.25**	-10.31*	-5.62**	-5.27**	-5.07**	-4.07**	-7.00**	-10.63
	(1.14)	(1.60)	(1.34)	(0.93)	(1.33)	(0.97)	(1.56)	(3.46)	(6.88)
Application score	-0.62	4.33**	4.01**	1.63**	-0.68	1.06*	-0.09	0.13	13.82
	(0.80)	(1.64)	(0.87)	(0.57)	(0.73)	(0.58)	(1.32)	(2.79)	(14.57)
Average y^k	30.01	23.40	46.15	51.79	27.31	87.85	78.37	75.61	105.83
Observations	8,391	11,030	10,987	3,269	6,422	3,085	1,245	4,403	1,251

Note: From 2SLS estimation of equations (14)-(15), we obtain a matrix of the payoffs to field j as compared to k for those who prefer j and have k as next-best field. Each cell is a 2SLS estimate (with st. errors in parenthesis) of the payoff to a given pair of preferred field and next-best field. The rows represent completed fields and the columns represent next-best fields. The row labeled average y^k reports the weighted average of the levels of potential earnings for compliers in the given next-best field. The final row reports the number of observations for every next-best field. Stars indicate statistical significance, * 0.10, ** 0.05.

Testing for "Comparative Advantage"

- With selection on gains would expect

$$E[Y_j \quad Y_k \; j \quad k] > E[Y_j \quad Y_k \; k \quad j]$$



Roy Takeaways

- Selection on potential gains a powerful, natural assumption
  - Should be comfortable with basic Roy formalization and how to solve
  - Above statistics facts are common labor tools
- Tight link between theory and empirics (all ID roads lead to sorting)
  - Post-credibility revolution, we care more about *what* causal parameters actually represent and how they inform theory
  - Nature of sorting bias can be just as interesting as a treatment effect
- With Roy selection and unknown heterogeneity, a valid instrument is not "enough" (ATE vs. LATE, "fallback" heterogeneity, RD locality)
  - How much structure is needed/plausible?
  - Are "model-free," data-driven assumptions satisfying (e.g. isoLATE, KLM'14)? Or is Heckman right that we need a selection model?
  - Would love to hear your thoughts!

14.662 Labor Economics II
Spring 2015