14.662 Recitation 7


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
Borjas’ (1987) Roy Notation and Setup

- Potential wages for individual $i$ with schooling level $j \in \{0, 1\}$:
  \[ w_{ij} = E[w_{ij}] + (w_{ij} - E[w_{ij}]) \equiv \mu_j + \varepsilon_{ij} \]

- 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
  \[ w_{i1} - w_{i0} > c \]
  \[ \mu_1 - \mu_0 - c > \varepsilon_{i0} - \varepsilon_{i1} \equiv z \]
  \[ \equiv \nu_i \]

  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 \( E[X] \)
Note first that
\[
\begin{bmatrix}
\varepsilon_{i0} \\
v_i
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_0^2 - \sigma_{01} \\
\sigma_0^2 - \sigma_{01} & \sigma_0^2 + \sigma_1^2 - 2\sigma_{01} \end{bmatrix} \right)
\]

By Fact #1,
\[
E[w_{i0}|z > v_i] = \mu_0 + E[\varepsilon_{i0}|z > v_i] = \mu_0 + E[E[\varepsilon_{i0}|v_i]|z > v_i]
\]

By Fact #2,
\[
E[\varepsilon_{i0}|v_i] = \frac{\sigma_0^2 - \sigma_{01}}{\sigma_v^2} v_i, \quad \text{where} \quad \sigma_v^2 \equiv \sigma_0^2 + \sigma_1^2 - 2\sigma_{01}
\]

So:
\[
E[w_{i0}|z > v_i] = \mu_0 + \frac{\sigma_0^2 - \sigma_{01}}{\sigma_v^2} E[v_i|v_i < z]
\]
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):
  \[
  \rho_{01} - \frac{\sigma_0}{\sigma_1} > 0 \quad \text{and} \quad \frac{\sigma_1}{\sigma_0} - \rho_{01} > 0
  \]
  \[
  \implies \rho_{01} \in \left[ \frac{\sigma_0}{\sigma_1}, \frac{\sigma_1}{\sigma_0} \right]
  \]
  \[
  \implies \text{Distribution of productivity with schooling more unequal}
  \]

- Negative selection (avg. $j = 1$ wage “below avg.” in both sectors):
  \[
  \rho_{01} - \frac{\sigma_0}{\sigma_1} < 0 \quad \text{and} \quad \frac{\sigma_1}{\sigma_0} - \rho_{01} < 0
  \]
  \[
  \implies \rho_{01} \in \left[ \frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1} \right]
  \]
  \[
  \implies \text{Distribution of productivity without schooling more unequal}
  \]

- 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?

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

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

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

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

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

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

- Then IV gives LATE; with Roy selection:

$$E[\mu_1 - \mu_0 + \varepsilon_{i1} - \varepsilon_{i0}|0 < \varepsilon_{i1} - \varepsilon_{i0} - (\mu_1 - \mu_0) \leq 1] \neq \mu_1 - \mu_0$$
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
  \[
  E[w_{ai} - w_{-ai}|A_1i > A_0i] = E[w_{ai} - w_{0i}|A_1i > A_0i, B_{0i} = 0]P(B_{0i} = 0|A_1i > A_0i) \\
  + E[w_{ai} - w_{bi}|A_1i > A_0i, B_{0i} = 1]P(B_{0i} = 1|A_1i > A_0i)
  \]
  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 $E[w_{ai} - w_{-ai} | A_{1i} > A_{0i}]$ 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 $\omega$ 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 $\omega$ 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 $\neq$ 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 \(k\):

\[
\begin{align*}
y &= \sum_{j \neq k} \beta_{jk} d_j + x' \gamma_k + \lambda_{jk} + \varepsilon \\
d_j &= \sum_{j \neq k} \pi_{jk} z_j + x' \psi_{jk} + \eta_{jk} + u, \quad \forall j \neq k
\end{align*}
\]
We use our first stage estimates to compute the probability among our compliers of a next-best field given a completed field. For each preferred field, we report the share of compliers for the two most common next-best field.

Figure 7. Complier weights of alternative fields by completed fields displays the two most common next-best fields for every preferred field. For example, this figure reveals that Science is the typical next-best field for compliers who prefer Technology or Engineering. It is also clear that Humanities, Social Science, and Teaching tend to be close substitutes. By comparing Figure 7 to Figure 1, we can see that the compliers to our instruments are similar to non-compliers in terms of their preferred and next-best field.

By computing the proportion of compliers by preferred and next-best field, we also learn that certain combinations of fields are rare. In particular, few compliers have Law as their next-best field, and virtually no one have Medicine as the next-best field. This means that we do not have support in our data to identify the effect of choosing field $j$ instead of Medicine, and that we have too few compliers to obtain precise estimates of the payoff to choosing field $j$ instead of Law.

7.2 2SLS estimates

Table 5 reports the 2SLS estimates from the model given by equations (14) and (15). By conditioning on next-best field, we are able to estimate the payoffs to different fields while
### KLM (2014) IV Estimates

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

<table>
<thead>
<tr>
<th>Completed field (j):</th>
<th>Humanities</th>
<th>Soc Science</th>
<th>Teaching</th>
<th>Health</th>
<th>Science</th>
<th>Engineering</th>
<th>Technology</th>
<th>Business</th>
<th>Law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humanities</td>
<td>21.38*</td>
<td>-4.72</td>
<td>-22.93*</td>
<td>4.97</td>
<td>-38.51**</td>
<td>6.87</td>
<td>-42.21**</td>
<td>-156.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(10.97)</td>
<td>(9.85)</td>
<td>(12.12)</td>
<td>(11.86)</td>
<td>(14.72)</td>
<td>(48.29)</td>
<td>(10.56)</td>
<td>(437.28)</td>
<td></td>
</tr>
<tr>
<td>Social Science</td>
<td>18.72**</td>
<td>9.84</td>
<td>-10.82</td>
<td>55.46**</td>
<td>-55.36**</td>
<td>-110.38</td>
<td>-28.37**</td>
<td>-76.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(6.73)</td>
<td>(11.55)</td>
<td>(13.00)</td>
<td>(21.45)</td>
<td>(20.60)</td>
<td>(102.97)</td>
<td>(10.66)</td>
<td>(86.42)</td>
<td></td>
</tr>
<tr>
<td>Teaching</td>
<td>22.25**</td>
<td>31.37**</td>
<td>1.82</td>
<td>23.46**</td>
<td>-33.94**</td>
<td>-35.32</td>
<td>-21.08**</td>
<td>22.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.96)</td>
<td>(7.88)</td>
<td>(6.55)</td>
<td>(9.45)</td>
<td>(12.54)</td>
<td>(37.07)</td>
<td>(7.12)</td>
<td>(127.87)</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td>18.75**</td>
<td>30.69**</td>
<td>7.72**</td>
<td>28.87**</td>
<td>-27.87**</td>
<td>-43.38**</td>
<td>-17.39**</td>
<td>-55.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(6.25)</td>
<td>(7.56)</td>
<td>(2.82)</td>
<td>(7.64)</td>
<td>(10.35)</td>
<td>(20.84)</td>
<td>(3.97)</td>
<td>(97.68)</td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>53.71**</td>
<td>69.59**</td>
<td>38.58**</td>
<td>29.63**</td>
<td>-2.21</td>
<td>16.81</td>
<td>-4.92</td>
<td>148.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(18.37)</td>
<td>(22.36)</td>
<td>(14.20)</td>
<td>(11.53)</td>
<td>(14.60)</td>
<td>(18.07)</td>
<td>(10.51)</td>
<td>(276.20)</td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>59.81</td>
<td>-5.53</td>
<td>75.24**</td>
<td>0.16</td>
<td>52.35**</td>
<td>-46.00</td>
<td>-13.03</td>
<td>-57.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(18.09)</td>
<td>(20.56)</td>
<td>(11.53)</td>
<td>(14.60)</td>
<td>(18.07)</td>
<td>(10.51)</td>
<td>(10.51)</td>
<td>(276.20)</td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td>41.87**</td>
<td>58.69**</td>
<td>22.08*</td>
<td>32.45**</td>
<td>68.07**</td>
<td>-5.56</td>
<td>7.03</td>
<td>-53.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(10.84)</td>
<td>(10.09)</td>
<td>(10.09)</td>
<td>(9.63)</td>
<td>(11.95)</td>
<td>(14.47)</td>
<td>(9.49)</td>
<td>(147.53)</td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>48.13**</td>
<td>61.93**</td>
<td>31.02**</td>
<td>30.22**</td>
<td>58.01**</td>
<td>-3.42</td>
<td>28.54*</td>
<td>3.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(11.25)</td>
<td>(12.03)</td>
<td>(8.78)</td>
<td>(10.86)</td>
<td>(12.61)</td>
<td>(15.61)</td>
<td>(9.16)</td>
<td>(83.04)</td>
<td></td>
</tr>
<tr>
<td>Law</td>
<td>46.34**</td>
<td>55.62**</td>
<td>36.60**</td>
<td>21.49*</td>
<td>40.07**</td>
<td>-27.53</td>
<td>-15.55</td>
<td>-1.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.16)</td>
<td>(8.34)</td>
<td>(11.56)</td>
<td>(9.68)</td>
<td>(18.29)</td>
<td>(17.96)</td>
<td>(8.66)</td>
<td>(83.61)</td>
<td></td>
</tr>
<tr>
<td>Medicine</td>
<td>83.34**</td>
<td>79.39**</td>
<td>62.62**</td>
<td>45.57**</td>
<td>81.31**</td>
<td>21.07</td>
<td>40.07**</td>
<td>23.34**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9.76)</td>
<td>(10.65)</td>
<td>(9.02)</td>
<td>(7.01)</td>
<td>(9.71)</td>
<td>(20.67)</td>
<td>(8.79)</td>
<td>(83.61)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-7.00**</td>
<td>-6.25**</td>
<td>-10.31*</td>
<td>-5.62**</td>
<td>-5.27**</td>
<td>-5.07**</td>
<td>-4.07**</td>
<td>-7.00**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.14)</td>
<td>(1.60)</td>
<td>(1.34)</td>
<td>(0.93)</td>
<td>(1.33)</td>
<td>(0.97)</td>
<td>(1.56)</td>
<td>(3.46)</td>
<td></td>
</tr>
<tr>
<td>Application score</td>
<td>-0.62</td>
<td>4.33**</td>
<td>4.01**</td>
<td>1.63**</td>
<td>-0.68</td>
<td>1.06*</td>
<td>-0.09</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.80)</td>
<td>(1.64)</td>
<td>(0.87)</td>
<td>(0.57)</td>
<td>(0.73)</td>
<td>(0.58)</td>
<td>(1.32)</td>
<td>(2.79)</td>
<td></td>
</tr>
<tr>
<td>Average y^k</td>
<td>30.01</td>
<td>23.40</td>
<td>46.15</td>
<td>51.79</td>
<td>27.31</td>
<td>87.85</td>
<td>78.37</td>
<td>75.61</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>8,391</td>
<td>11,030</td>
<td>10,987</td>
<td>3,269</td>
<td>6,422</td>
<td>3,085</td>
<td>1,245</td>
<td>4,403</td>
<td></td>
</tr>
</tbody>
</table>

**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 compilers 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


Figure 10. Comparative advantage

Address this concern, we have re-estimated the model given by equations (14) and (15) with treatment variables defined according to subfields instead of broader fields. The estimates are shown in Appendix Figure A8. They suggest that aggregation to broader fields is not driving the conclusion that compliers tend to prefer fields in which they have comparative advantage.

10.4 Sorting pattern and economic models

The above results suggest that self-selection and comparative advantage are empirically important features of field of study choices. These findings have implications for the type of economic models that can help explain the causes and consequence of individuals choosing different types of post-secondary education.

Much economic analysis and empirical work relies on an efficiency unit framework where there is only one type of human capital which individuals possess in different amounts.

While the presence of comparative advantage is at odds with models based

23 A prominent example is the Ben-Porath model, which assumes efficiency units so different labor skills are perfect substitute. Heckman et al. (1998) extend the standard Ben-Porath model by relaxing

45 Note: This figure graphs the distribution of the differences in relative payoffs to field j versus k between individuals whose preferred field is j and next-best alternative is k and those with the reverse ranking. To con- struct this graph, we use the complier weighted distribution of estimates in Appendix Table A6.

13/14 Courtesy of Lars Kirkebøen, Edwin Leuven, and Magne Mogstand. Used with permission.
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!