Lecture Note 5: Self-Selection – The Roy Model

David H. Autor
MIT 14.662 Spring 2011

May 3, 2011

1 Introduction

A core topic in labor economics is 'self-selection.' What this term means in theory is that rational actors make optimizing decisions about what markets to participate in – job, location, education, marriage, crime, etc. What it means in practice is that observed economic relationships should generally be viewed as endogenous outcomes of numerous optimizing decisions, rather than as exogenous causal relationships. Understanding self-selection should make you skeptical of treating ecological correlations as causal.

The starting point of formal treatment of this topic in economics is Roy's (1951) "Thoughts on the Distribution of Earnings," which discusses the optimizing choices of 'workers' selecting between fishing and hunting. Roy's key observation is that there are three technological factors that affect this choice:

1. The fundamental distribution of skills and abilities
2. The correlations among these skills in the population
3. The technologies for applying these skills

At the time of Roy's writing, the presumption was that the distribution of income that arises from economic processes is arbitrary. Hence, if we compare the mean earnings of hunters and fishermen, $\bar{y}_h$ and $\bar{y}_f$, then $\bar{y}_h - \bar{y}_f$ is an estimate of the earnings gain or loss that an individual would receive from switching from fishing to hunting. Roy's article explains why this view is incorrect.

The essential departure of Roy's model from previous work is that it is a multiple-index model (in this case, 2 indices): workers have skills in each occupation, but they can only use one skill or the other. Hence, workers self-select the occupation (sector) that gives them the highest expected earnings. Equilibrium in each market equates supply and demand, while a self-selection condition means that the marginal worker is indifferent between the two sectors.

Roy's 1951 paper is amusing to read to a contemporary economist because it so awkwardly straddles the line between older-style narrative economics and modern mathematical economics. Roy is clearly writing with equations and distributions in mind (probably even written out), but he writes mostly about rabbits and fish – only occasionally interjecting that 'therefore' earnings will be log-normal. This is very hard to follow. Borjas' 1987 AER Paper on "Self-Selection and the Earnings of Immigrants" is the first paper that I know that writes down a simple, parametric 2-sector Roy model. The enduring contribution of Borjas' paper is this model (sometimes called a Borjas selection model) rather than the empirical findings. As a labor economist, you should be well versed in this model.

2 Brief review

2.1 The normal selection model

Assume that the full earnings distribution is given by

$$w \sim N\left(\mu_0, \sigma_0^2\right).$$

We only observe wages of those who work. Assume that everyone in the economy has a reservation wage of κ. Thus, the wage is only observed if $w > \kappa$, equivalently $\varepsilon_0 > k - \mu_0$.

What is the expectation of observed wages?

$$\begin{aligned} E\left(w|w>k\right) &= \mu_0 + E\left(\varepsilon_0 | k - \mu_0 > k\right) \\ &= \mu_0 + \sigma_0 E\left(\frac{\varepsilon_0}{\sigma_0}\middle| \frac{\varepsilon_0}{\sigma_0} > \frac{k-\mu_0}{\sigma_0}\right) \\ &= \mu_0 + \sigma_0 \frac{\phi\left(z\right)}{1-\Phi\left(z\right)} \\ &= \mu_0 + \sigma_0 \frac{\phi\left(z\right)}{\Phi\left(-z\right)}, \end{aligned}$$

where $z = \left(k - \mu_0\right)/\sigma_0$, $\phi\left(\cdot\right)$ and $\Phi\left(\cdot\right)$ are the PDF and CDF of the standard normal distribution respectively. Note that we can flip the sign of z due to the symmetry of the normal distribution.

The ratio $\lambda\left(z\right) = \phi\left(-z\right)/\Phi\left(-z\right)$ is typically called the Inverse Mills Ratio, and is a hazard function (the slides contain a graphical depiction), with $\lambda\left(z\right) \geq 0, \lambda'\left(z\right) > 0, \lim\limits_{z\to-\infty} \lambda_t\left(z\right) \to 0, \lim\limits_{z\to\infty} \lambda_t'\left(z\right) \to 1.$

Recall that the conditional distribution of a normal random variable is also a normal random variable. This is a tremendously useful property.

2.2 Correlation coefficient, regression coefficient and conditional normality

The correlation between two random variables y_0 and y_1 is written as:

$$\rho = \frac{\sigma_{01}}{\sigma_0\sigma_1},$$

where σ_{01} is $\mathrm{cov}(\sigma_0, \sigma_1)$.

Similarly, the regression of y_1 on y_0 is:

$$y_1 = \alpha_0 + \beta y_0 + \varepsilon,$$

where $\beta = \frac{\sigma_{01}}{\sigma_0}$. Thus $\rho = \frac{\beta}{\sigma_1}$.

Now add the assumption that

$$\left(\begin{array}{c} y_0 \\ y_1 \end{array} \right) \sim N \left(\begin{array}{c} 0 \\ 0 \end{array}, \begin{array}{cc} \sigma_0 & \cdot \\ \sigma_{01} & \sigma_1 \end{array} \right).$$

What this does for the above regression equation is two things: (1) $\alpha_0 = 0$; and (2) ε is now also normally distributed. This is important to keep in mind about normally distributed variables: the conditional distribution of a normal random variable is also normal; the sum of a series of normal random variables is also normal.

3 Borjas 1987: Self-Selection and the Earnings of Immigrants

Who chooses to immigrate to the United States? One ready-made answer is that workers from low wage countries will immigrate. This may be true on average, but it's probably too simple. The workers immigrating to the United States are probably not a random subset of the Mexican workforce. Rather, we should expect that potential migrants make some rough comparison of their wages in the home country and their expected wages in the U.S. On average, we'd expect those who immigrate to have higher expected earnings in the U.S. than Mexico and vice versa for those who stay.

Before zeroing in on the Borjas immigration example, however, you should note that there is nothing specific to immigration in this model. You can alternatively think of country 0 as the household production sector and country 1 as the market sector. With this substitution, this model can be used to study the relationship between skills, self-selection, and the gender wage gap (as is done by Mulligan and Rubinstein in their 2008 *QJE* paper).

- Consider two countries 0 and 1, denoting the source and host country.

- Log earnings in the source country are given by

$$w_0 = \mu_0 + \varepsilon_0,$$

where $\varepsilon_0 \sim N\left(0, \sigma_0^2\right)$. It's useful to think of ε_o as the demeaned value of worker's 'skill' in the source country.

- If everyone from country 0 were to migrate to the host country, their earnings would be (ignoring any general equilibrium effects!):

$$w_1 = \mu_1 + \varepsilon_1,$$

  where $\varepsilon_1 \sim N\left(0, \sigma_1^2\right)$.

- Assume that the cost of migrating is C, which Borjas puts into 'time equivalent' terms as $\pi = C/w_0$. Borjas further assumes that π is constant, meaning that C is directly proportional to w_0.

- Assume further that each worker knows C, μ_0, μ_1 and his individual epsilons: $\varepsilon_0, \varepsilon_1$.

- You, the econometrician, only observe a worker in one country or the other, and hence you only know ε_0 or ε_1 for any individual.

- What can you infer about what wages for immigrants in the United States would have been had they stayed in their source countries? What would wages in the United States be for non-migrants had they come to the United States? The Roy Model answers these questions.

- The correlation between source and host country earnings is

$$\rho = \frac{\sigma_{01}}{\sigma_0 \sigma_1},$$

  where σ_{01} is $\operatorname{cov}(\sigma_0, \sigma_1)$.

- To implement this model, we need to know ρ, although we do not need to know both $\varepsilon_0, \varepsilon_1$ for any worker.

- A worker will choose to migrate if

$$(\mu_1 - \mu_0 - \pi) + (\varepsilon_1 - \varepsilon_0) > 0, \tag{1}$$

  (Borjas defines the indicator variable I, equal to 1 if this selection condition is satisfied, 0 otherwise).

- Now, define $\nu = \varepsilon_1 - \varepsilon_0$. The probability that a randomly chosen worker from the source country chooses to migrate is equal to

$$\begin{aligned} P &= \Pr\left[\nu > (\mu_0 - \mu_1 + \pi)\right] \\ &= \Pr\left[\frac{\nu}{\sigma_\nu} > \frac{(\mu_0 - \mu_1 + \pi)}{\sigma_\nu}\right] \\ &= 1 - \Phi\left(\frac{(\mu_0 - \mu_1 + \pi)}{\sigma_\nu}\right) \\ &= 1 - \Phi(z) \\ &= \Phi(-z) \end{aligned}$$

where $z = (\mu_0 - \mu_1 + \pi)/\sigma_\nu$, and $\Phi(\cdot)$ is the CDF of the standard normal. Notice that the higher larger is z, the lower is the probability of migration. This is because z is rising in the mean earnings of the home country and the cost of migration. So $\partial P/\partial\mu_0 < 0, \partial P/\partial\mu_1 > 0, \partial P/\partial\pi < 0$.

- These are mean effects. It's useful to assume from here forward that $\mu_1 \approx \mu_0$, so that we can focus on self-selection rather than mean differences.

3.1 Selection conditions

- What is the expectation of earnings in the *source* country for workers who choose to immigrate?

-

$$\begin{aligned} E(w_0|\text{Immigrate}) &= \mu_0 + E\left(\varepsilon_0 \middle| \frac{\nu}{\sigma_\nu} > z\right) \\ &= \mu_0 + \sigma_0 E\left(\frac{\varepsilon_0}{\sigma_0} \middle| \frac{\nu}{\sigma_\nu} > z\right). \end{aligned} \tag{2}$$

- Notice that this equation depends on three things:

1. Mean earnings in home and source country
2. Both error terms $(\varepsilon_0, \varepsilon_1)$ through ν.
3. Implicitly, it also depends on the correlation between the error terms.

- We want to know the expectation of ε_0 given some value ν. Given the normality of $\varepsilon_0, \varepsilon_1$, this is simply equal to the regression coefficient

$$E\left(\varepsilon_0|\nu\right) = \frac{\sigma_{0\nu}}{\sigma_\nu^2}\nu.$$

Applying this to (2), we get

$$\begin{aligned}
E\left(\varepsilon_0|\nu\right) &= \frac{\sigma_{0\nu}}{\sigma_\nu^2}\nu \\
E\left(\frac{\varepsilon_0}{\sigma_0}|\frac{\nu}{\sigma_\nu}\right) &= \frac{\sigma_{0\nu}}{\sigma_\nu^2}\frac{\sigma_\nu}{\sigma_0}\nu \\
&= \frac{\sigma_{0\nu}}{\sigma_0\sigma_\nu}\nu \\
&= \rho_{0\nu}\frac{\nu}{\sigma_\nu}.
\end{aligned}$$

- Hence, we can rewrite (2) as

$$\begin{aligned}
E(w_0|\text{Immigrate}) &= \mu_0 + \sigma_0 E\left(\frac{\varepsilon_0}{\sigma_0}\middle|\frac{\nu}{\sigma_\nu} > z\right) && (3) \\
&= \mu_0 + \rho_{0\nu}\sigma_0 E\left(\frac{\nu}{\sigma_\nu}\middle|\frac{\nu}{\sigma_\nu} > z\right) \\
&= \mu_0 + \rho_{0\nu}\sigma_0\left(\frac{\phi\left(z\right)}{1-\Phi\left(z\right)}\right) \\
&= \mu_0 + \rho_{0\nu}\sigma_0\left(\frac{\phi\left(z\right)}{\Phi\left(-z\right)}\right) && (4)
\end{aligned}$$

where $\phi\left(z\right)/\left(1-\Phi\left(z\right)\right)$ is the Inverse Mills Ratio (IMR), equal to the conditional expectation of a standard normal random variable truncated from the left at point z. We can interchange z and $-z$ in the numerator given the symmetry of $\phi\left(\cdot\right)$. The IMR is a hazard function (the slides contain a graphical depiction), with $\lambda\left(z\right) \geq 0, \lambda'\left(z\right) > 0$, $\lim\limits_{z\to-\infty}\lambda_t\left(z\right) \to 0$, $\lim\limits_{z\to\infty}\lambda_t'\left(z\right) \to 1$. A hazard function answers the question 'what is the probability of an event given that the event has not already occurred?' Here, the IMR answers the question: what is the expectation of epsilon given that epsilon is greater than or equal to z?

- Similarly to above, we can calculate the expected wage in the source country of workers who do migrate as:

$$\begin{aligned}
E(w_1|\text{Immigrate}) &= \mu_1 + E\left(\varepsilon_1\middle|\frac{\nu}{\sigma_\nu} > z\right) && (5) \\
&= \mu_1 + \rho_{1\nu}\sigma_1\left(\frac{\phi\left(z\right)}{\Phi\left(-z\right)}\right)
\end{aligned}$$

3.2 Three cases

It is convenient to express equations (3) and (5) in terms of ρ_{01} rather than $\rho_{0\nu}$. Define the correlation coefficient $\rho_{01} = \frac{\sigma_{01}}{\sigma_0 \sigma_1}$.Rearranging the expression for $\rho_{o\nu}$:

$$\begin{aligned}
\rho_{0\nu} &= \frac{\sigma_{0\upsilon}}{\sigma_0 \sigma_\nu} \\
&= \frac{E\left[\varepsilon_0\left(\varepsilon_1 - \varepsilon_0\right)\right]}{\sigma_0 \sigma_\nu} \\
&= \frac{\sigma_{01} - \sigma_0^2}{\sigma_0 \sigma_\nu} \\
&= \frac{\sigma_0 \sigma_1}{\sigma_\nu}\left[\frac{\sigma_{01}}{\sigma_0 \sigma_1} - \frac{\sigma_0}{\sigma_1}\right] \\
&= \frac{\sigma_0 \sigma_1}{\sigma_\nu}\left[\rho_{01} - \frac{\sigma_0}{\sigma_1}\right].
\end{aligned}$$

Similarly:

$$\begin{aligned}
\rho_{1\nu} &= \frac{\sigma_{1\upsilon}}{\sigma_1 \sigma_\nu} \\
&= \frac{E\left[\varepsilon_1\left(\varepsilon_1 - \varepsilon_0\right)\right]}{\sigma_0 \sigma_\nu} \\
&= \frac{\sigma_1^2 - \sigma_{01}}{\sigma_1 \sigma_\nu} \\
&= \frac{\sigma_0 \sigma_1}{\sigma_\nu}\left[\frac{\sigma_1}{\sigma_0} - \frac{\sigma_{01}}{\sigma_0 \sigma_1}\right] \\
&= \frac{\sigma_0 \sigma_1}{\sigma_\nu}\left[\frac{\sigma_1}{\sigma_0} - \rho_{01}\right].
\end{aligned}$$

Substituting into (3) and (5) yields (with notation $\rho = \rho_{01}$):

$$\begin{aligned}
E(w_0|\text{Immigrate}) &= \mu_0 + \rho_{0\nu}\sigma_0\left(\frac{\phi\left(z\right)}{\Phi\left(-z\right)}\right) \\
&= \mu_0 + \frac{\sigma_0 \sigma_1}{\sigma_\nu}\left(\rho - \frac{\sigma_0}{\sigma_1}\right)\left(\frac{\phi\left(z\right)}{\Phi\left(-z\right)}\right), \qquad (6)
\end{aligned}$$

and

$$\begin{aligned}
E(w_1|\text{Immigrate}) &= \mu_1 + \rho_{1\nu}\sigma_1\left(\frac{\phi\left(z\right)}{\Phi\left(-z\right)}\right) \\
&= \mu_1 + \frac{\sigma_0 \sigma_1}{\sigma_\nu}\left(\frac{\sigma_1}{\sigma_0} - \rho\right)\left(\frac{\phi\left(z\right)}{\Phi\left(-z\right)}\right). \qquad (7)
\end{aligned}$$

Define

$$\begin{aligned} Q_0 &= E(\varepsilon_0 | I = 1) \\ Q_1 &= E(\varepsilon_1 | I = 1) \end{aligned}$$

...we now turn to the three cases.

3.2.1 Positive hierarchical sorting

- This is a case where immigrants are positively selected from the source country distribution and are also above the mean of the host country distribution: $Q_0 > 0, Q_1 > 0$. This will be true iff

$$\frac{\sigma_1}{\sigma_0} > 1 \text{ and } \rho > \frac{\sigma_0}{\sigma_1}.$$

- What do these conditions mean? First, $\frac{\sigma_1}{\sigma_0} > 0$ implies that the host country has a higher 'return to skill' than the source country. Second, $\rho > \frac{\sigma_0}{\sigma_1}$, implies that the correlation between the skills valued in the host and source country is sufficiently high to induce migration of high skilled workers. If you were a skilled worker in the source country, you would not want to migrate to a host country with a very high return to skills if the skills valued in the host country were uncorrelated (or negatively correlated) with skills value in the home country.

- This case embodies the traditional American view of immigration: 'The best and the brightest' leave their home countries for greater opportunity (that is, higher return to skill) in the U.S.

- One way of restating this type of migration is: a source country with low earnings variance 'taxes' the earnings of high skill workers and insures the earnings of low skill workers. High skill workers may want to emigrate, accordingly. But this is not the only possibility.

3.2.2 Negative hierarchical sorting

- This is a case where immigrants are negative selected from the source country distribution and are also below the average of the host country distribution: $Q_0 < 0, Q_1 < 0$. This will be true iff

$$\frac{\sigma_0}{\sigma_1} > 1 \text{ and } \rho > \frac{\sigma_1}{\sigma_0}.$$

- This is simply the converse case. Here, the source country is unattractive to low earnings workers because of high wage dispersion. Again assuming that wages are sufficiently correlated between the source and host country, low skill workers will want to migrate to take advantage of the 'insurance' provided by a narrow wage structure in the host country.

- So, this is the potentially unattractive case (certainly from Borjas' perspective) where a compressed wage structure 'subsidizes' low skill workers, thus attracting low skill workers from abroad.

3.2.3 'Refugee' sorting

- A third case is where $Q_0 < 0, Q_1 > 0$, that is, immigrants are selected from the lower tail of the home country distribution but arrive in the upper tail of the host country distribution. This can only occur if

  $$\rho < \min\left(\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1}\right),$$

  meaning that the correlation between earnings in the two countries is sufficiently low (could be negative).

- This might occur, for example, for a minority group whose opportunities in the host country are depressed by prejudice. Or for the case of migration from a non-market economy where the set of skills rewarded is quite different from the economy in the receiving country (e.g., intellectuals fleeing the crumbling USSR to the West in the 1980s).

3.2.4 A fourth case?

- Note that there is **not** a fourth case where $Q_0 > 0, Q_1 < 0$. Why not? This would suggest irrational migration, where people leave the upper tail of the source country income distribution to join the lower tail of the host country distribution. This is inconsistent with income maximization.

- Mathematically, this case would require that

  $$\rho > \max\left(\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1}\right),$$

  which would imply that $\rho > 1$, which cannot be true for a correlation coefficient.

3.3 Relevance?

- How relevant is the Borjas/Roy selection model to the problem he studies in the 1987 paper, self-selection of immigrants? The evidence is not overwhelming. There are probably more relevant applications of this model.

  1. One is Borjas' 2002 NBER paper on self-selection into government employment in the U.S. Though the returns to skills/education rose rapidly in the U.S. during the 1980s, the wage structure in government jobs was stable, meaning that in relative terms, the public sector wage structure became compressed. You would therefore expect a negative hierarchical sorting case to become potentially relevant: high skill workers leave the government for the private sector and low skill workers remain in government jobs to be sheltered from falling wages. This is what Borjas' paper purports to show, and casual empiricism suggests that this is likely to be right.
  2. A closely related application, but one I've not seen applied, is the changing selection of workers into primary and secondary school teaching. Teaching likely offered relative high 'returns to skills' for educated women prior to the 1970s—for many women, it was one of the few professions that was open to them. This has changed of course. Teaching now probably provides a high floor and a low ceiling for wages for college educated females. This would suggest growing adverse selection into teaching over time. Worth exploring.

- The enduring contribution of Borjas' paper for labor economists is its simple and useful formulation of the Roy model. (Note that a limitation of the model is that it ignores general equilibrium effects whereby large immigrant flows would change the wage structure parameters in the source and host countries.)

- The growing focus of empirical economists on applying instrumental variables to causal estimation is in large part a response to the realization that self-selection (i.e., optimizing behavior) plagues interpretation of ecological relationships. Hence, understanding the importance of self-selection has vastly improved empirical work.

- But instrumental variables are not the only answer to testing cause and effect with observed data.

Self-selection also points to the existence of *equilibrium relationships* that should be observed in ecological data (i.e., those who immigrate should in general do better in the host country than the source country), and these can be tested without an instrument. In fact, there are some natural sciences that proceed almost entirely without experimentation – for example, astrophysics. How do they do it? Models predict non-obvious relationships in data. These implications can be verified or refuted by data, and this evidence strengthens or overturns the hypotheses. Many economists have set aside this methodology.

4 Self-Selection and the Closing of the Gender Gap (Mulligan-Rubinstein)

One of the most significant wage structure developments of the last thirty years has been the closing of the gender gap. Between 1980 and 2000, raw (unadjusted) estimates indicate that roughly 17 of the 40 percentage point gap average earnings gap between female and male workers were erased. Claudia Goldin's 2006 Ely Lecture (published in the AER P&P) presents some of the fascinating facts underlying the transforming labor market status of women.

The rising earnings of women is seen as one of the signs of the health of the US labor market, and is often contrasted to the stagnant earnings of blacks relative to whites. The fact that women gained ground on men during the 1980s at a time when income inequality was rising rapidly is seen as particularly remarkable. As discussed in by Blau and Khan in their 1997 *JOLE* paper "Swimming Upstream," if women had merely stayed at the same *percentile* in the wage distribution during the 1980s, their average earnings would have fallen relative to men simply because, as the earnings distribution widened, a given wage percentile gap would have translated into a larger absolute wage gap. (If the median female earner is below the median of the pooled gender wage distribution, then an increase in wage dispersion—leading to an increase in the wage gap between each percentile—should have caused a decline in female relative wages in a 'single-index' model). These observations imply that women gained during the 1980s *despite* changes in the wage structure.

In contrast, the 2008 QJE paper by Mulligan and Rubinstein argues that women gained (or *appeared* to gain) *because* of changes in the wage structure. M-R use the Roy Model to re-interpret the relationship between the gender gap and wage inequality. Figure 1 of M-R shows quite strikingly that female relative wages rose in remarkable symmetry with the rise in male wage inequality. And this bivariate relationship continued to hold during at least two inflection points of male wage inequality in the last 30 years (1976, 1994).

M-R propose self-selection as an explanation for this correspondence. This hypothesis is not immediately appealing in its conventional form. Typically, we think of variation in self-selection in labor markets as occurring through changes in the participation rate. If female participation increases, and self-selection into the labor market is positive, then the quality of female participants will fall, *raising* the gender gap. Conversely, if women are negatively selected into the labor market, an increase in female participation would reduce the gender gap (by reducing the extent of selection). There were sizable increases in female participation in the 1980s, but this trend stopped or reversed in the 1990s.

The discussion above implicitly assumes that shifts in the participation rate occur against a backdrop of a stable selection mechanism. The hypothesis advanced by M-R is more subtle. They hypothesize that, due to the substantial change in the market reward to skills implied by the changes in male wages during the last several decades, the nature of self-selection of women into the labor market has changed. Specifically, self-selection may have become more positive (if it was originally positive), less negative (if it was originally negative), or could even have flipped signs from negative to positive. In any of these cases, self-selection could reduce the measured gender wage gap without either a substantial change in female participation or a real decline in the latent gender wage gap. This is a nice insight. The evidence favoring it is intriguing, if not overwhelming.

4.1 Model

Let agent i's wages (w_{it}) and reservation wages (r_{it}) in time t be given by:

$$\begin{aligned} w_{it} &= \mu_t^w + \gamma_t + \sigma_t^w \varepsilon_{it}^w, \\ r_{it} &= \mu_t^r + \sigma_t^r \varepsilon_{it}^r, \end{aligned}$$

where ε_{it}^w and ε_{it}^r are person $i's$ year t deviation from the average of persons with his or her gender and observed characteristics. Hence ε_{it}^w and ε_{it}^r can be viewed as person-specific market and non-market "skills" which are priced by the market (σ_t^w) and the non-market "prices" (σ_t^r) in time t. We can normalize ε_{it}^w and ε_{it}^w so their standard deviations are each one with mean zero for each gender at each point in time. (We are assuming that the underlying skills distributions is time-invariant.) In this model, γ_t is the 'true' (unconditional) gender gap in earnings.

Person i works/participates ($L_{it} = 1$) iff $w_{it} > r_{it}$:

$$\varepsilon_{it}^w \frac{\sigma_t^w}{\sigma_t^r} - \varepsilon_{it}^r > -\frac{\mu_t^w + \gamma_t - \mu_t^r}{\sigma_t^r}.$$

Let the person-specific stochastic terms $\sigma_t^w \varepsilon_{it}^w$ and $\sigma_t^r \varepsilon_{it}^r$ follow a bivariate normal distribution:

$$\begin{pmatrix} \sigma_t^w \varepsilon_{it}^w \\ \sigma_t^r \varepsilon_{it}^r \end{pmatrix} \tilde{} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_t^{w2} & \rho_t \sigma_t^w \sigma_t^r \\ \rho_t \sigma_t^w \sigma_t^r & \sigma_t^{r2} \end{pmatrix} \right],$$

where ρ_t is the correlation between ε_{it}^w and ε_{it}^r in time t. Notice that the correlation coefficient is defined as

$$\rho_t = \frac{\sigma_t^{wr}}{\sigma_t^w \sigma_t^r},$$

so $\rho_t \sigma_t^w \sigma_t^r = \sigma_t^{wr}$.

To focus on the effect of prices on participation, ρ_t is assumed to be constant over time—that is, $\rho_t = \rho$. (It is not automatic that this would be true.) The variance/covariance of wages and reservation wages are still allowed to vary over time due to changes in "prices" (σ_t^w, σ_t^r).

Person specific "skills" therefore follow a *standard* bivariate normal distribution:

$$\begin{pmatrix} \varepsilon_{it}^w \\ \varepsilon_{it}^r \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]. \tag{8}$$

Let v_{it} denote person's i deviation from the average net gain from employment of persons with her gender and observed characteristics:

$$\nu_{it} = \varepsilon_{it}^w \frac{\sigma_t^w}{\sigma_t^r} - \varepsilon_{it}^r.$$

What is the *measured* (as opposed to latent) gender gap? It can be written as:

$$\begin{aligned}
G_t &= \gamma_t + E\left(\varepsilon_t^w \sigma_t^w | \, \varepsilon_{it}^w \frac{\sigma_t^w}{\sigma_t^r} - \varepsilon_{it}^r > -\frac{\mu_t^w + \gamma_t - \mu_t^r}{\sigma_t^r} \right), && (9) \\
&= \gamma_t + \sigma_t^w E\left(\varepsilon_t^w | \, \varepsilon_{it}^w \frac{\sigma_t^w}{\sigma_t^r} - \varepsilon_{it}^r > -\frac{\mu_t^w + \gamma_t - \mu_t^r}{\sigma_t^r} \right), && (10) \\
&= \gamma_t + \sigma_t^w \frac{\operatorname{Cov}(\varepsilon_t^w, \nu_t)}{\operatorname{Var}(\varepsilon_t^w)} E\left(\frac{\nu_t}{\sigma_t^\nu} > -\frac{\mu_t^w + \gamma_t - \mu_t^r}{\sigma_t^r \sigma_t^\nu} \right), && (11) \\
&= \gamma_t + \sigma_t^w \sigma_{w\nu}^t \left(\frac{\nu_t}{\sigma_t^\nu} > -\frac{\mu_t^w + \gamma_t - \mu_t^r}{\sigma_t^r \sigma_t^\nu} \right), && (12) \\
&= \gamma_t + \sigma_t^w \rho_{w\nu}^t \lambda(\delta_t) && (13) \\
&= \gamma_t + \sigma_t^w b_t && (14)
\end{aligned}$$

where b_t is the "selection bias" on skills evaluated at market prices σ_t^w, $\delta_t = \frac{\mu_t^w + \gamma_t - \mu_t^r}{\sigma_t^r}$, $\lambda(\delta_t) = \phi(\delta_t)/\Phi(\delta_t)$, and $\rho_{w\nu}^t$ is the correlation between ν and ε^w. Note that $1 \geq \lambda(\delta_t) \geq 0$ and $\lambda'(\delta_t) \leq 0$. If participation reaches 100%, there is no self-selection and so $\lim_{z\rightarrow\infty} \lambda(z) \rightarrow 0$.

Thus,

$$b_t = E\left[\varepsilon_{it}^w | g_i = 1, \nu_{it}/\sigma_t^v > -\delta_t\right] = \rho_{w\nu}^t \lambda(\delta_t),$$

We need to solve for this selection bias

Due to the bivariate normality of the error terms, ν_{it} is normally distributed with a standard deviation that is equal to:

$$\sigma_t^v = \left(1 + \left(\frac{\sigma_t^w}{\sigma_t^r}\right)^2 - 2\rho\frac{\sigma_t^w}{\sigma_t^r}\right)^{1/2}. \tag{15}$$

The correlation between ε^w and ν is equal to:

$$\rho_{w\nu}^t = \frac{cov\left[\left(\varepsilon_{it}^w \frac{\sigma_t^w}{\sigma_t^r} - \varepsilon_{it}^r\right), \varepsilon_{it}^w\right]}{\sigma_t^v \sigma_\varepsilon^w} = \frac{\frac{\sigma_t^w}{\sigma_t^r} var\left(\varepsilon_{it}^w\right) - cov\left(\varepsilon_{it}^r, \varepsilon_{it}^w\right)}{\sigma_t^v}.$$

Given (8) the covariance between ε^r and ε^w is equal to the correlation between these terms, $cov\left(\varepsilon_{it}^r, \varepsilon_{it}^w\right) = \rho$, and $var\left(\varepsilon_{it}^w\right) = 1$ which means that:

$$\rho_{w\nu}^t = \frac{\sigma_t^w/\sigma_t^r - \rho}{\sigma_t^v},$$

which given (15) is equal to:

$$\rho_{w\nu}^t = \frac{\sigma_t^w/\sigma_t^r - \rho}{\left(1 + \left(\frac{\sigma_t^w}{\sigma_t^r}\right)^2 - 2\rho\frac{\sigma_t^w}{\sigma_t^r}\right)^{1/2}}. \tag{16}$$

By substituting (16) back into (**??**) we obtain:

$$b_t = E\left[\varepsilon_{it}^w | g_i = 1, L_{it} = 1\right] = \frac{\sigma_t^w/\sigma_t^r - \rho}{\left(1 + \left(\frac{\sigma_t^w}{\sigma_t^r}\right)^2 - 2\rho\frac{\sigma_t^w}{\sigma_t^r}\right)^{1/2}} \lambda(\delta_t). \tag{17}$$

So, the bias term is equal to:

$$\sigma_t^w b_t = \frac{\sigma_t^w/\sigma_t^r - \rho}{\left(1 + \left(\frac{\sigma_t^w}{\sigma_t^r}\right)^2 - 2\rho\frac{\sigma_t^w}{\sigma_t^r}\right)^{1/2}} \lambda(\delta_t).$$

Putting it all together:

$$G_t = \gamma_t + \sigma_t^w \times \frac{\sigma_t^w/\sigma_t^r - \rho}{\left(1 + \left(\frac{\sigma_t^w}{\sigma_t^r}\right)^2 - 2\rho\frac{\sigma_t^w}{\sigma_t^r}\right)^{1/2}} \times \lambda\left(\delta_t\right)$$

Notice that $\rho_{w\nu}^t$ and hence b_t will be positive iff $\sigma_t^w/\sigma_t^r - \rho > 0$ (equivalently, $\sigma_t^w > \rho\sigma_t^r$). Thus, the sign of self-selection depends on the price of market relative to home-production wages, σ_t^w/σ_t^r, and the correlation between home and market skills, ρ. $\text{Sign}\langle\rho_{w\nu}\rangle = \text{Sign}\langle\sigma_t^w/\sigma_t^r - \rho\rangle$.

4.2 Comparative statics

The key insight of the M-R paper is that if σ_t^w rises, this will tend to make self-selection of females into the labor market more positive (or less negative). And it is possible for the sign of self-selection to flip, so that female self-selection goes from negative to positive (though this is not necessary for the main results).

The following comparative statics are slightly helpful:

$$\left.\frac{\partial b_t}{\partial\left(\sigma_t^w/\sigma_t^r\right)}\right|_{d\lambda(\delta_t)=0} = \frac{1-\rho^2}{\left(1 + \left(\frac{\sigma_t^w}{\sigma_t^r}\right)^2 - 2\rho\frac{\sigma_t^w}{\sigma_t^r}\right)^{3/2}}\lambda\left(\delta_t\right) > 0.$$

This comparative static says that a rise in the relative price of market work makes more positive the correlation between market skill and the gain to market work. This increases positive self-selection or decreases negative self-selection. Either way, it improves the composition of female workers.

The second comparative static,

$$\frac{\partial b_t}{\partial\lambda\left(\delta_t\right)} = \frac{\sigma_t^w/\sigma_t^r - \rho}{\left(1 + \left(\frac{\sigma_t^w}{\sigma_t^r}\right)^2 - 2\rho\frac{\sigma_t^w}{\sigma_t^r}\right)^{1/2}},$$

says that an increase in female participation tends to attenuate whatever selection process is in place. (Recall that $\lambda'(\cdot) < 0$, so a rise in female LFP lowers λ.) So, if self-selection is negative ($\sigma_t^w/\sigma_t^r - \rho < 0$) and more women work, this will raise the wage of female workers. If self-selection is positive ($\sigma_t^w/\sigma_t^r - \rho > 0$) and more women work, this will lower the wage of female workers.

Figure 2 of M-R is quite helpful in seeing how this sign-flipping case could work.

4.3 Aside: Brief intuition of Mulligan-Rubinstein model

- Negative hierarchical sorting

$$\frac{\sigma_0}{\sigma_1} > 1 \text{ and } \rho > \frac{\sigma_1}{\sigma_0}$$

- Positive hierarchical sorting

$$\frac{\sigma_1}{\sigma_0} > 1 \text{ and } \rho > \frac{\sigma_0}{\sigma_1}$$

- What happens if σ_1 rises?

$$\begin{aligned}
\partial\left(\frac{\sigma_1}{\sigma_0}\right)/\partial\sigma_1 &> 0\\
\frac{\partial}{\partial\sigma_1}\left[\rho - \frac{\sigma_0}{\sigma_1}\right] &= \frac{\partial}{\partial\sigma_1}\left[\frac{\sigma_{01}}{\sigma_0\sigma_1} - \frac{\sigma_0}{\sigma_1}\right]\\
&= \left[-\frac{\sigma_{01}}{\sigma_0\sigma_1^2} + \frac{\sigma_0}{\sigma_1^2}\right]\\
&= \left[\frac{\sigma_0}{\sigma_1^2} - \frac{\sigma_{01}}{\sigma_0\sigma_1^2}\right]\\
&= \left[\frac{\sigma_0^2 - \sigma_{01}}{\sigma_0\sigma_1^2}\right] > 0
\end{aligned}$$

- Thus, a rise in σ_1, holding constant σ_0, either increases the extent of positive hierarchical sorting, decreases the extent of negative hierarchical sorting, or flips the nature of selection from negative to positive hierarchical selection. In any of these cases, the gender gap may close due to improved selection of women into the labor market without any change in the latent female/male relative wage.

4.4 Estimation: Control function approach

In theory, one can empirically recover the latent gender gap, γ_t, by controlling for self-selection. This is called the control function approach, and, in the normal selection case (normal distribution that is), it's often called the 'Heckit' estimator (Heckman (1979) two-step). In particular, we can write

$$b_t = \theta\left(\sigma_t^w/\sigma_t^r\right)\lambda\left(P_t\right),$$

where $\theta\left(\sigma_t^w/\sigma_t^r\right)$ proxies for $\rho_{w\nu}^t$, λ is the inverse Mills ratio, and P_t is the fraction of workers participating for the relevant X category (gender, education, etc.). This involves a slight abuse of notation since the argument of the IMR is the standardized selection term, δ_t, rather than the fraction of workers participating, which is $\Phi(\delta_t)$. So, we should write $\lambda\left(\phi\left(\Phi^{-1}(P_t)\right)/P_t\right)$.

Thus, the control function approach essentially involves a wage regression of observed wages on a set of $X's$ and the control function $\lambda(P_t)$. Given the linear relationship between the observed wage, the latent wage mean and the Mills ratio (see equation 9), we 'should' be able to recover γ_t simply by controlling for $\lambda(P_t)$. As is well understood, if there are not excluded instruments to identify $\lambda(P_t)$, this whole approach is simply working off of functional form (since $\lambda(P_t)$ is just a non-linear function of the identical $X's$ included in the linear 2nd stage of the regression). Recognition of this fact has caused this approach (lacking instruments) to go out of vogue. M-R bravely plow ahead, also throwing in some dubious instruments about marriage and kids for good measure (hopefully, these aren't doing much work). But to their credit, M-R are candid about the limitations of this approach and supplement the Heckit findings with several complementary forms of evidence.

Here are the two equations of the Heckit. These are typically estimated in sequence but can be estimated jointly with MLE:

$$\begin{aligned} P_t(Z) &\equiv \Pr(L=1|Z=1) = \Phi(Z\delta_t), \\ w_{it} &= X_{it}\beta_t + g_i\gamma_t + g_i\theta_t\lambda(Z_{it}\delta_t) + \mu_{it}. \end{aligned}$$

Table 1 of M-R implies that selection has indeed gone from negative to positive. That is, controlling for selection, M-R estimate that the latent gender gap was less negative than OLS would suggest during the mid/late 70s, and that the latent gender gap was more negative than OLS would suggest during the mid/late 90s. Their point estimates imply that there has been no closing of the latent gender gap over these 20 years. Should you take this conclusion seriously? It's hardly gospel, especially based on this evidence alone. But it's fascinating.

4.5 Other estimation approaches: ID at infinity

The M-R framework implies that the gender gap should have closed by less among groups of females that already had high LF participation in the 1970s; for these groups, there was not much room for a change in participation. In the limit, if participation is stable at 100%, there is no role for self-selection. This case is called 'identification at infinity' (because as δ_t tends towards infinity, $\lambda(\delta_t)$ tends towards zero, meaning there is no selection problem). Figures 4 and 5 present evidence that there is less gender wage convergence for groups with initially high LFP.

Table 3 also presents evidence suggesting that self-selection into the LF based on IQ has became more positive in the 1980s and 1990s than it was in the 1970s. However, this analysis uses a single cohort of women, so we really cannot be confident that the paper is isolating a time effect as opposed to an age effect.

5 A Roy Model with Spillovers: Chandra and Staiger, 2007

Self-selection models are usually applied in the labor market, but their applicability is much broader. Chandra and Staiger (2007) use a Roy model with productivity spillovers to explain a truly puzzling set of facts in the medical literature. It is well known that there is enormous geographic variation in the use of intensive medical treatments across seemingly comparable places. Surprisingly, the use of higher intensity treatment is not associated with improved satisfaction, outcomes or survival, but is associated with higher costs.

The standard explanation for these facts is ‘flat of the curve’ medicine. Practitioners in certain regions, for unspecified reasons, apply intensive procedures until the marginal return is zero. This argument doesn’t actually make much sense on its face since it would still imply that average outcomes would be better in more intensive regions (unless marginal benefits are zero for everyone). Moreover, there is credible evidence that some intensive treatments produce substantial benefits and are generally under-provided.

Chandra and Staiger propose an economic explanation that is non-obvious and subtle but fits many of the facts. In their model, patients differ in their degree of *suitability* for intensive treatment. More suitable patients always receive the intensive treatment and less suitable patients do not. (This is the Roy model component: intensive versus non-intensive treatments have comparative advantage with different patients). However, the productivity/efficacy of the two treatments is each an increasing function of the rate at which these treatments are applied. Areas that perform intensive treatment in a larger fraction of cases get better at the intensive treatment and *worse* at the non-intensive treatment. (This assumption is plausible in medicine: there is extensive learning by doing.) Accordingly, areas that perform a higher fraction of intensive treatments (for whatever reason) will get better results with intensive treatments for given patient characteristics *but* will get worse results with the non-intensive treatments for given patient characteristics. Therefore, high-intensity areas will *optimally* perform the intensive treatment on a subset of patients who would benefit more from non-intensive treatments *if* given in a non-intensive hospital. (Restated: given a hospital's intensity/specialization, it treats each patient appropriately. But marginal patients, those who are not suited to that hospital's specialty, would fare better if sent to a hospital specializing in the alterative.)

This model gives rise to geographic variation in intensive treatment, greater use of intensive treatment on patients who are only marginally suitable in high-intensity areas, better results for suitable patients in high-intensity areas, and worse results for marginally suitable patients in high-intensity areas. It is therefore ambiguous as to whether or not average outcomes will be better in high-intensity areas.

5.1 Model

Let i index treatments. The survival rate and cost associated with each treatment are:

$$\begin{aligned}\text{Survival}_i &= \beta_i^s Z + \alpha_i^s P_i + \varepsilon_i^s \text{ for } i = 1, 2 \\ \text{Cost}_i &= \beta_i^c Z + \alpha_i^c P_i + \varepsilon_i^c \text{ for } i = 1, 2,\end{aligned}$$

where Z is a vector of patient characteristics and P_i is the proportion of patients treated with treatment i.

The indirect utility of a patient depends on both the survival rate and cost of care:

$$U_i = \text{Survival}_i - \lambda \text{Cost}_i = \beta_i Z + \alpha_i P_i + \varepsilon_i,$$

where $\beta_i = \beta_i^s - \lambda \beta_i^c$, and similarly for α_i and ε_i. (One wonders if the patient or doctor cares about the cost, so it is plausible that $\lambda = 0$). $\beta_i Z$ can be thought of as an index of how suitable a patient is for a given treatment. The term $\alpha_i P_i$ captures the productivity spillover, which is positive if $\alpha_i > 0$. The error term is presumed to be observed by the patient and treating physician but not by the econometrician.

The probability that an individual patient receives intensive treatment ($i = 2$), and bearing in mind that $P_1 = 1 - P_2$:

$$\begin{aligned}\Pr\left[i=2\right] &= \Pr\left[U_2 - U_1 > 0\right] \\ &= \Pr\left[\left(\alpha_1+\alpha_2\right)P_2 - \alpha_1 + \left(\beta_2-\beta_1\right)Z > \varepsilon_1 - \varepsilon_2\right] \\ &= \Pr\left[\alpha P_2 - \alpha_1 + \beta Z > \varepsilon\right],\end{aligned}$$

where $\alpha = \alpha_1 + \alpha_2$, $\beta = \beta_2 - \beta_1$, and $\varepsilon = \varepsilon_1 - \varepsilon_2$.

Among patients who choose the treatment, the expected benefit is:

$$E\left[U_2 - U_1 | U_2 - U_1 > 0\right] = \beta Z + \alpha P_2 - \alpha_1 + E\left[\varepsilon | U_2 - U_1 > 0\right].$$

Thus, patients who choose the intensive treatment benefit if they are relatively more appropriate (larger βZ) or live in more intensive regions (higher αP_2).

Equilibrium in this model is complicated. There may be multiple equilibria because the benefits to intensive treatment increases with the number of patients taking the treatment. Equilibrium is achieved when the proportion of patients choosing the intensive treatment generates benefits that are consistent with this proportion choosing the treatment. The following equilibrium condition must hold:

$$P_2 = \int_z \Pr\left(\alpha P_2 - \alpha_1 + \beta Z > \varepsilon\right) f\left(Z\right) dZ \equiv G\left(P_2\right)$$

This equilibrium is a fixed point of the equation. Figure 1 shows examples with single and multiple equilibria.

The major takeaway of the model is found in Figure 2. Figure 2A shows that for a given patient, the model implies that the patient receives optimal treatment *given* the area in which he is treated. However, the marginal patient treated intensively an intensive area would have fared better if he had been treated non-intensively in a non-intensive area. (However, this patient would have fared still worse if he had been treated non-intensively in the intensive area.)

Q: Is the market equilibrium likely to be Pareto efficient? Is it likely to be welfare maximizing?

5.2 Evidence

The measure of intensive care is cardiac catherization, which is a surgical procedure. Medical management is the alternative procedure. All patients should be prescribed beta blockers–though not all are, and this is a measure of the quality of medical management. The geographical unit is the Hospital Referral Region (HRR). Given the endogeneity of the treatment that a patient receives, some models use as instruments the Differential Distance (DD) from a patient's home zip code to the nearest cath *relative to* nearest non-cath hospital.

Estimating equation:

$$\text{Outcome}_{ijk} = \beta_{0k} + \beta_{1k}\text{Intensive}_i + X_i\Pi_k + \upsilon_{ijk},$$

where i is a person, j is a HRR, and k is some measure of suitability for treatment (so the coefficients will differ by suitability). This model requires instruments for Intensive treatment.

However, for estimates where outcomes are modeled as a function of area Cath rates, estimates are done by OLS. Why? Because the model suggests these relationships should obtain in equilibrium. Of course, if large differences in patient pops drives P_2 across regions, that's a problem.

Adjusted area cath rates are:

$$\Pr\left(\text{Cath}_{ij}\right) = G\left(\theta_0 + \theta_j + X_i\Phi\right).$$

The θ_j's are the risk adjusted rates. Often, HRR's are split into two groups, above and below median.

Here are some key results:

- Table 1. 2SLS estimates indicate that the benefits of cath for those who receive it (treatment on treated) are much larger (and much, much cheaper) for those with greater suitability, measured either by empirical likelihood from cross-section analysis or by age group (over age 80, cath is not generally advised). Thus, patients who are most suitable benefit the most from Cath. (This is a necessary baseline fact to establish.)

- Table 2 presents evidence of instrument validity. Columns 1 and 2 show that DD is a very strong predictor of the likelihood of cath, both overall and within suitability groups. Columns 3 and 4 allow a Wald estimate of the effect of DD on survival. The implied Wald estimate is similar to the models in Table 1 that include many additional controls. This suggests that the instrument is largely orthogonal to the control variables, which is generally good news. Columns 5 through 8 suggest that there is little observable difference in underlying suitability or expected survival rates among patients according to DD. Again, good news for 2SLS.

- Table 3 tests whether marginal patients are less suitable for cath in high cath areas by estimating the following equation *for patients receiving cath*:

  $$\text{Appropriateness}_{ij} = \mu_0 + \mu_1 \ln\left(\text{Cath Rate}\right)_j + \varepsilon_i.$$

  If marginal patients who receive cath are less appropriate in high cath areas, we would expect that $\mu_1 < 0$. More specifically, μ_1 estimates the gap between the marginal and average suitability of cath patients:

  $$\frac{\partial\left(A/C\right)}{\partial \ln C} = \frac{\partial C\left(A/C\right)}{\partial C} = \frac{\partial A}{\partial C} - \frac{CA}{C^2} = \frac{\partial A}{\partial C} - \frac{A}{C},$$

  where A/C is the average suitability of patients who are catheterized. So, μ_1 has a subtle interpretation. One can similarly show that $\partial \ln\left(A/C\right)/\partial \ln C = \left(\partial A/\partial C - A/C\right)/\left(A/C\right)$, meaning that the log-log regression gives the gap between the marginal and average outcomes stated in percentage terms.

- Table 4. Risk adjusted use of beta-blockers is negatively correlated with intensivity of Cath, which is indicator of low-quality medical management in these high-Cath areas.

- Table 5 seems like a poor idea. As far as I can tell, this a regression of Y on $\bar{Y}$, which is a regression you do not want to run.

- Table 6 finds that the patients most appropriate for cath do comparatively better in high cath regions.

- Table 7 offers what are arguably the most compelling results in the paper. Line A shows that costs are higher but average outcomes are not better in high cath regions. Yet, for, patients suitable for cath, survival rates considerably higher *and* for patients not suitable for cath, survival rates are significantly lower. This is a fairly amazing finding.

5.3 Summary of Chandra-Staiger

Perhaps even more than M-R, Chandra-Staiger demonstrate how a little bit of economic theory can go a very long way in explaining a set of puzzling empirical relationships. One unusual feature of the paper is that it uses a mixture of OLS and 2SLS estimates *as appropriate* to test the model's empirical implications. Implications tested at the individual level use 2SLS because the question of interest is the effect of intensity on individual outcomes conditional on suitability. Here, random assignment is key. However, estimates that study area relationships between cath propensity and characteristics of patients receiving cath (e.g., Table 3, studying the relationship between area cath rates and the average and marginal characteristics of patients receiving cath, and Table 7, studying the relationship between patient suitability for cath and area outcomes as a function of cath rates) are fit using OLS because the model implies that these relationships should be detectable in equilibrium (i.e., without random assignment). [Table 7 could also be estimated by 2SLS, however.]

6 Basu (2002) "Sexual Harassment in the Workplace."

The Basu paper on your syllabus provides an insightful (and controversial) objection to the assumed efficiency of the operation of implicit markets. This is not exactly a Roy model (it's a compensating-diffs model, which is close but not identical to a Roy model). Still, it's interesting...

I lay out the basic model and allow you to ponder the implications on your own. (Note that the MIT Working Paper 02-11 available via SSRN is more rigorous than the Journal of Economics Perspectives Version from 2003: http://papers.ssrn.com/abstract_id=303184 .) Here's the model:

- Firms produce output Y using only labor, where n is the number of laborers:

$$Y = f(n),\ f'(n) > 0, f''(n) < 0.$$

- Assume that sexual harassment is legal, and that employers get perverse gratification, $\theta > 0$ (measured in units of output), from each worker they are are allowed to harass.
- Write the wages of harassed and non-harassed workers as w_N, w_H.
- In this case, the firm's payoff from production is:

$$\pi(n_N, n_H) = f(n_N + n_H) - n_H w_H - n_N w_N + n_H \theta.$$

- If firms employ both types of labor in equilibrium, in must be the case that

$$w_H = w_N + \theta,$$

and

$$f'(n_N + n_H) = w_N.$$

- The population of workers is measured on the interval $[0, P]$.
- Assume that all workers have the following labor supply function

$$s(w) \text{ with } s'(w) > 0.$$

- Workers differ however in their disutility of being sexually harassed. Write the monetized cost of harassment to worker i as $c(i)$.
- Index the worker with the highest disutility as 0 and the lowest disutility as P, with workers having decreasing disutility as we move from 0 to P.
- Assume the following is true:

$$c(0) > \theta > c(P).$$

- Hence, in equilibrium, some workers will choose to be harassed and others not. Another way of saying this is:

$$\begin{aligned} s(w_N) &> s(w_N + \theta - c(0)) \text{ and} \\ s(w_N) &< s(w_N + \theta - c(P)). \end{aligned}$$

- Define w_N^B as the wage level prevailing in a world where sexual harassment as banned.
- Define w_N^A as the wage level prevailing *for non-harassed workers* in a world where sexual harassment is legal.
- Let $\iota(\cdot)$ be the inverse of the $c(\cdot)$ function.
- When harassment is banned, labor supply is simply:

$$L_s^B = P \cdot s\left(w_N^B\right)$$

- When harassment is allowed, labor supply is:

$$L_s^A = \int_{\iota(\theta)}^{P} s\left(w_N^A + \theta - c(i)\right) di + \iota(\theta)\, s\left(w_N^A\right).$$

- Observe that for women in the interval $\iota(\theta)$, labor supply under the harassment regime is higher than it would be under the non-harassment regime.
- Basu shows (and it's easy to see why) that $w_N^B > w_N^A$, that is, non-harassed workers are better off (higher paid) if harassment is illegal, even if these workers are not harassed *in either case.* It's also case that some workers who accept harassment in the legalized regime are also better off in the world without harassment.