Self-selection: The Roy model

Heidi L. Williams

MIT 14.662

Spring 2015

1. Preliminaries: Overview of 14.662, Part II

2. A model of self-selection: The Roy model

3. Application: Immigration
   - Chiswick (1978) and Borjas (1985): Assimilation
   - Abramitzky, Boustan, and Eriksson (2014): Assimilation
   - Borjas (1987): A model of self-selection
   - Abramitzky, Boustan, and Eriksson (2012): Testing the Roy model

4. Looking ahead

Preliminaries: Overview of 14.662, Part II

Much of 661/662 focuses on theories of earnings distributions

- Many so far: Human capital, Rosen-style superstar models...
- Neal and Rosen (2000) *Handbook* chapter a useful overview
  - Facts: empirical regularities in earnings distributions
    - ★ Right skewed
    - ★ Mean earnings greatly differ across groups (*e.g.* education, gender)
  - Theories: provide a synthesized review

Preliminaries: Overview of 14.662, Part II

Roadmap for the rest of the semester

- Four models with implications for earnings distributions:
  1. Roy model
  2. Compensating differentials model
  3. Discrimination models
  4. Models of rent-sharing
- Three related topics which speak to other empirically important determinants of the distribution of labor earnings:
  1. Management practices
  2. Intergenerational mobility
  3. Early life determinants of long-run outcomes

Preliminaries: Overview of 14.662, Part II

Logistics

- Remaining lectures: Continue comments on assigned papers
- Two additional problem sets (due 4/22 and 5/6)
- Research paper proposal (due 4/28)

1. Preliminaries: Overview of 14.662, Part II

2. A model of self-selection: The Roy model

3. Application: Immigration
   - Chiswick (1978) and Borjas (1985): Assimilation
   - Abramitzky, Boustan, and Eriksson (2014): Assimilation
   - Borjas (1987): A model of self-selection
   - Abramitzky, Boustan, and Eriksson (2012): Testing the Roy model

4. Looking ahead

The Roy model: Roy (1951)

How does occupation self-selection impact the income distribution?

- Motivation: Contemporaries assumed distribution of incomes was arbitrary: "developed by the process of historical accident"
- Core of Roy's model is to ask how the distribution of earnings is affected if individuals purposively select their occupation

Definitely worth reading, but not an easy read (verbal math)

- Instead will walk through (formally identical) Borjas (1987) model
- 'Standard' formalization: important for you to be comfortable with
  - Notes walk through more mechanics than I will cover in class
  - Section this week will also walk through this model

The Roy model: Roy (1951)

Two occupations: (rabbit) hunting and fishing

Goal was to understand self-selection:

- Will the individuals best suited for fishing choose to fish?
- Will the individuals best suited for hunting choose to hunt?

Core idea: individuals will not randomly sort across occupations

- Implies that the wage gap will reflect not only a "real" difference in potential earnings, but will also be a function of occupational sorting

Applications of the Roy model

Roy-style selection applicable to essentially every sub-field of economics

We will focus on three applications:

1. Immigration: Borjas (1987), Abramitsky et al. (2012, 2014)
2. Health care: Chandra and Staiger (2007)
3. Redistribution: Abramitzky (2009)

Other applications on the syllabus:

- Borjas (2002): sorting of workers into the public sector
- Dahl (2002): geographic variation in returns to education
- Kirkeboen, Leuven, and Mogstad (2014): fields of study
- Rothschild and Scheuer (2013): optimal tax
- Willis and Rosen (1979): sorting into college attendance

1. Preliminaries: Overview of 14.662, Part II

2. A model of self-selection: The Roy model

3. Application: Immigration
   - Chiswick (1978) and Borjas (1985): Assimilation
   - Abramitzky, Boustan, and Eriksson (2014): Assimilation
   - Borjas (1987): A model of self-selection
   - Abramitzky, Boustan, and Eriksson (2012): Testing the Roy model

4. Looking ahead

Borjas (1987) application of the Roy model

Motivation: Understanding native-immigrant earnings differences, with a focus on the self-selection induced by the migration decision

- Model written from perspective of an immigrant thinking of migrating from her home (non-US) country to the US
- Idea: Individuals compare potential income in the US with income in home country, make migration decision based on income differential (net of migration costs)
- Induces self-selection $\Rightarrow$ empirically testable predictions
  1. If US has higher returns to skill (higher income inequality), migrants disproportionately drawn from top of home country's skill distribution
  2. Vice versa if US has lower returns to skill (lower income inequality)

Context for this paper

Borjas (1999) *Handbook* chapter on economics of immigration:

1. Why do some people move? Our focus
2. What happens when they do? 14.661: Card (1990), Borjas (2003)

We will focus in particular on skill composition of immigrants

- Important for interpreting native-immigrant earnings differences
- Of course, economic impact of immigration (question #2) depends on the skill distributions of natives and immigrants

Pre-Roy model of migration decisions

Chiswick (1978): *"Economic theory suggests that migration in response to economic incentives is generally more profitable for the more able and more highly motivated"*

- Footnote outlining a simple model generating that prediction
- Key assumption: ability has same effect on earnings in home, US
- Roy model relaxes this assumption: selection critically depends on correlation between value of ability in home, US $\Rightarrow$ in Roy model, self-selection will not always imply immigrants are positively selected

1. Preliminaries: Overview of 14.662, Part II
2. A model of self-selection: The Roy model
3. Application: Immigration
   - Chiswick (1978) and Borjas (1985): Assimilation
   - Abramitzky, Boustan, and Eriksson (2014): Assimilation
   - Borjas (1987): A model of self-selection
   - Abramitzky, Boustan, and Eriksson (2012): Testing the Roy model
4. Looking ahead

Interpreting native-immigrant earnings differences: Chiswick-Borjas

Chiswick (1978): How does "time in the US" affect immigrant earnings?

- Estimated standard cross-sectional Mincer-style human capital earnings functions that included variables for "years since migration"
- Possible because, for the first time since 1930, the (recently released) 1970 US Census asked a question about year of arrival
- Chiswick's conclusions thus based on cross-sectional comparison of different cohorts in 1970

Chiswick (1978) analysis

In the 1970 Census data, Chiswick estimated regressions like the following:

$$\ln(earnings_i) = \mathbf{X_i'}\theta + \delta I_i + \alpha_1 I_i \text{Years}_i + \alpha_2 I_i \text{Years}_i^2 + \epsilon_i$$

where:

- $\mathbf{X_i'}$: covariates such as education and potential experience
- I_i: indicator for foreign-born
- Years_i: years since migration

Chiswick (1978) estimates

TABLE 2

REGRESSION ANALYSIS OF EARNINGS FOR NATIVE- AND FOREIGN-BORN ADULT WHITE MEN, 1970

	NATIVE BORN	NATIVE AND FOREIGN BORN			FOREIGN BORN
	(1)	(2)	(3)	(4)	(5)
EDUC	.07154	.07058	.07004	.07164	.05740
	(53.78)	(55.68)	(55.18)	(54.11)	(12.93)
T	.03167	.03050	.03071	.03097	.02028
	(22.99)	(22.86)	(22.99)	(23.10)	(3.47)
T2	−.00052	−.00049	−.00050	−.00051	−.00031
	(−20.77)	(−20.45)	(−20.78)	(−20.93)	(−3.18)
LN WW	1.10335	1.10326	1.10169	1.10111	1.07151
	(81.75)	(84.78)	(84.70)	(84.67)	(21.97)
RURALEQ1	−.17222	−.16970	−.17080	−.16915	−.05821
	(−20.28)	(−20.25)	(−20.39)	(−20.18)	(−1.13)
SOUTHEQ1	−.12090	−.12620	−.12530	−.12389	−.21587
	(−14.17)	(−15.01)	(−14.91)	(−14.74)	(−4.38)
NOTMSP	−.30647	−.31078	−.30947	−.30874	−.34498
	(−27.76)	(−28.97)	(−28.86)	(−28.79)	(−7.66)
FOR	*	.02951	−.16359	.00990	*
		(1.75)	(−4.32)	(0.18)	
(FOR) (YSM)	*	*	.01461	.01555	.01500
			(3.98)	(4.23)	(3.87)
(FOR) (YSM2)	*	*	−.00016	−.00018	−.00019
			(−2.47)	(−2.79)	(−2.82)
(FOR) (EDUC)	*	*	*	−.01619	*
				(−4.23)	
CONSTANT	−1.03646	−1.01537	−1.00016	−1.02156	−.78891
Observations (*N*)	34,321	36,245	36,245	36,245	1,924
R	.55423	.55455	.55533	.55564	.58194
R^2	.30717	.30753	.30839	.30873	.33866
Standard error	.70900	.71008	.70966	.70949	.71676

SOURCE.—U.S. Bureau of the Census 1972.
NOTE.—*t*-ratios in parentheses; dependent variable: natural logarithm of earnings in hundreds of dollars.
* Variable not entered.

Chiswick conclusion #1

The experience-earnings profile of immigrants is steeper than the experience-earnings profile of natives with the same measured skills.

- Estimated coefficients evaluated at 10 years of experience ($T = 10$) and 5 years of residency ($YSM = 5$)
- Concluded that the return to experience for immigrants (2.718%) is steeper than for natives (2.07%)

Chiswick conclusion #2

The experience-earnings profile of immigrants crosses the experience-earnings profile of natives about 10-15 years after immigration.

- Estimated coefficients, holding constant schooling and total labor market experience
- For $YSM = 10$, predicted % difference in earnings between natives and foreign born is $\approx -3.349\%$; for $YSM = 15$, this is $\approx +1.956\%$
- Hence, he concluded that the immigrant experience-earnings profile crossed that of natives between 10 and 15 years after immigration
- Chiswick interpreted this fact as evidence of self-selection in migration in favor of "high ability, highly motivated workers, and workers with low discount rates for human capital investments."

Problem: Age-time-cohort effects

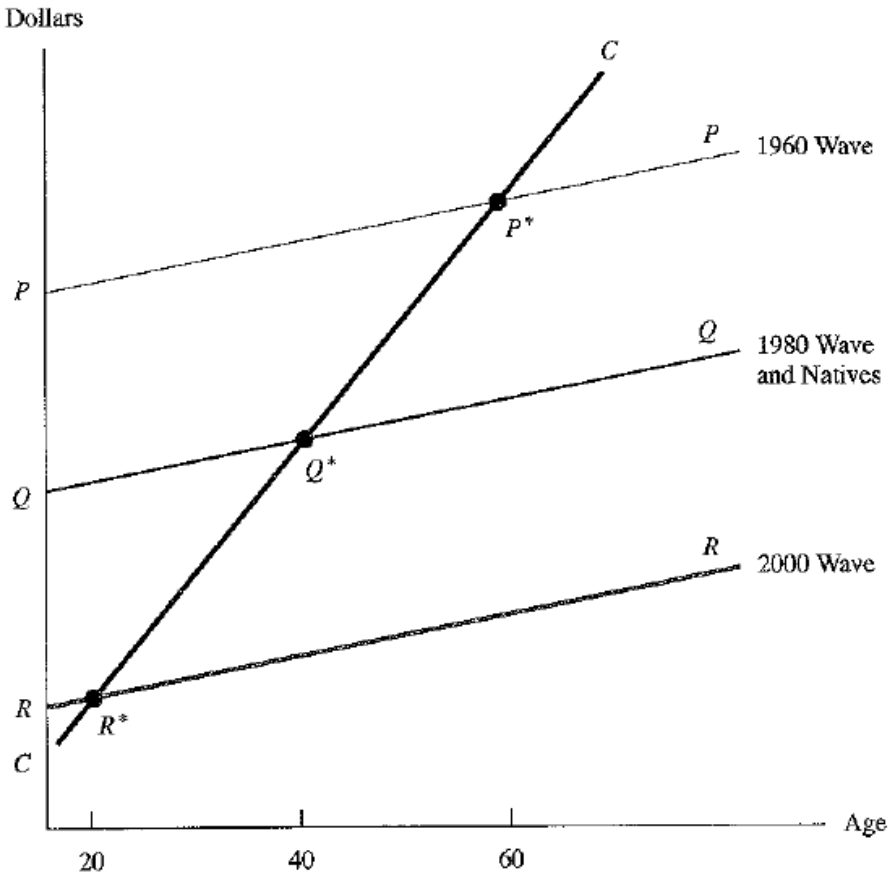
However, because the 1970 Census is a single cross-section, the "years since migration" variable may confound two effects:

1. A true "assimilation" effect
2. Fixed quality differences across immigrant cohorts: quality of immigrant cohorts – in terms of their earnings – could change over time as a function of, *e.g.*, changes in immigration policies.

Figure 8-5 in Borjas's *Labor Economics* text illustrates why the Chiswick type cross-sectional analysis can erroneously estimate patterns in the age-earnings profile that may be driven by fixed differences across cohorts.

Problem: Age-time-cohort effects (continued)

FIGURE 8-5 Cohort Effects and the Immigrant Age-Earnings Profile
The typical person migrating in 1960 is skilled and has age-earnings profile *PP;* the 2000 immigrant is unskilled and has age-earnings profile *RR;* the 1980 immigrant has the same skills as the typical native and has age-earnings profile *QQ*. Suppose all immigrants arrive at age 20. The 2000 census cross section reports the wages of immigrants who have just arrived (point R^*); the wage of immigrants who arrived in 1980 when they are 40 years old (point Q^*); and the wage of immigrants who arrived in 1960 when they are 60 years old (point P^*). The cross-sectional age-earnings profile erroneously suggests that immigrant earnings grow faster than those of natives.



Source: Figure 8-5 in Borjas's *Labor Economics* text (Fifth Edition, p. 333).

Problem: Age-time-cohort effects (continued)

A true "assimilation" effect and fixed cohort differences are indistinguishable in the 1970 Census because:

$$\text{(year of migration)} + \text{(years in US)} = 1970$$

Stated differently, the Chiswick-style cross-section approach encountered a version of the (now) well-known problem that it is impossible to separately identify age and cohort effects in a single cross-section.

Borjas (1985)

Borjas (1985) realized progress can be made by using repeated cross-section or longitudinal data (and a "test version of Stata")

- Took advantage of (recently released) 1980 US Census
- Contribution was to combine 1970 and 1980 US Census data to examine how well Chiswick's cross-sectional predictions about earnings growth predicted the actual earnings growth experienced by specific immigrant cohorts during the period 1970-1980

Borjas (1985): Method

In order to identify both the assimilation effect and cohort effects while controlling for year effects, a restriction must be imposed

- Borjas assumed time-specific shocks have the same effect on log earnings of natives and immigrants
- In a pooled sample of native-born and foreign-born individuals, effectively uses natives to estimate the Census year indicators
- Implicit assumption: factors fixed within Census year have same effect on log earnings of natives and immigrants
  - For factors like inflation, that assumption seems reasonable
  - However, other year-specific factors – such as business cycles – may have differential effects on natives and immigrants

Borjas (1985): Conclusions

- Chiswick: immigrants adapt quite rapidly into US labor market
- Borjas reached a different conclusion:
  - Documents relatively slower rates of earnings growth for immigrants
  - Implies a decline in the quality of immigrant cohorts

Take-away #1: Age-time-cohort effects

Methodological point: The impossibility of identifying age, time, and cohort effects in a linear model comes up in a variety of contexts

- Useful framework to keep in mind while reading papers, attending seminars, working on your own research
- Example: Dave Molitor's MIT dissertation on physician practices

Take-away #2: Substantive conclusions

How did the substantive conclusions of this Chiswick-Borjas exchange relate to Borjas's later Roy model application?

- Chiswick (1978): interpreted the fact that experience-earnings profile of immigrants crosses that of natives as evidence of self-selection in migration in favor of "high ability, highly motivated workers"
- Borjas (1985): clarified that this could instead reflect cohort effects
- Raises the question of how cohort effects relate to self-selection
- This question provides the starting point for Borjas's application of the Roy model

1. Preliminaries: Overview of 14.662, Part II
2. A model of self-selection: The Roy model
3. Application: Immigration
   - Chiswick (1978) and Borjas (1985): Assimilation
   - Abramitzky, Boustan, and Eriksson (2014): Assimilation
   - Borjas (1987): A model of self-selection
   - Abramitzky, Boustan, and Eriksson (2012): Testing the Roy model
4. Looking ahead

Abramitzky, Boustan, and Eriksson (2014)

- Re-examine this question using data on European immigrants to the US labor market during the Age of Mass Migration (1850-1913)
- Motivation for analyzing this period: Contemporaries were concerned about the ability of migrants to assimilate into the US economy
  - Congressional commission in 1907 concluded immigrants - particularly from southern/eastern Europe - would be unable to assimilate
  - Report fueled subsequent legislation to restrict immigrant entry via a literacy test (1917) and quotas (1924)
  - Subsequent analyses suggested - contrary to the commission's report - immigrants caught up with natives after 10 to 20 years in the US
  - However, all of these studies are subject to:
    1. Borjas (1985) critique on cohort effects
    2. Bias due to selective return migration

Novel data

- Ambitious new data effort: Construct a novel panel data set that follows native-born workers and immigrants from 16 sending countries through the US censuses of 1900, 1910, and 1920
- Match individuals by first/last name, age, country/state of birth
- Because these censuses do not contain data on wages or income, they assign individuals the median income in their reported occupation

Empirical specifications

1. Cross-section model
   - Compare occupation (proxy for labor market earnings) of native-born and immigrant workers as a function of time spent in the US, indicators for year and country of origin, and age controls
   - Note: arrival cohort indicators not included
2. Repeated cross-section model
   - Add arrival cohort indicators
   - Comparison with cross-section model allows them to infer how much of the earnings difference between natives and immigrants is attributable to differences in the quality of arrival cohorts
3. Panel model
   - Follows individuals across census years
   - Comparison with repeated cross-section allows them to infer whether and to what extent return migrants were positively or negatively selected from the immigrant population

Figure 2

Panel estimates suggest that the average immigrant did not face a substantial occupation-based earnings penalty upon first arrival, and experienced occupational advancement at the same rate as natives

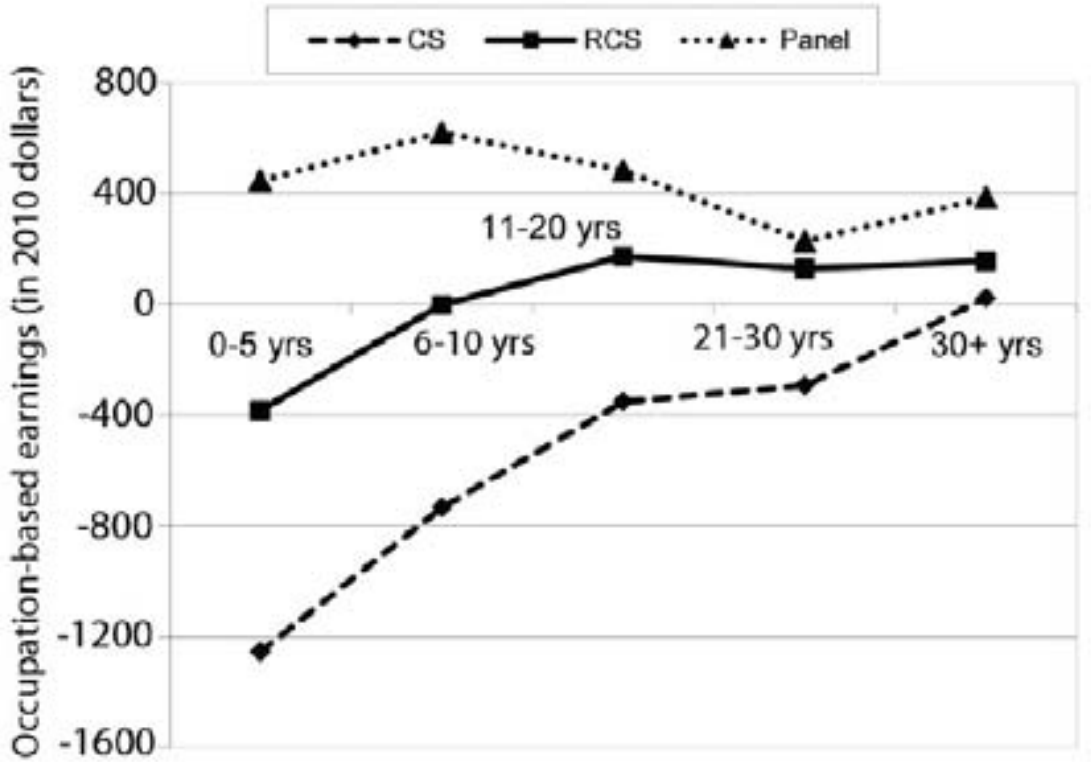


FIG. 2.—Convergence in occupation score between immigrants and native-born workers by time spent in the United States, cross-sectional and panel data, 1900–1920. The graph plots coefficients for years spent in the United States indicators in equation (1). Note that for the panel line, we subtract the native-born dummy from the years in the United States indicators (because the omitted category in that regression is natives in the panel sample). See table 4 for coefficients and standard errors.

1. Preliminaries: Overview of 14.662, Part II
2. A model of self-selection: The Roy model
3. Application: Immigration
   - Chiswick (1978) and Borjas (1985): Assimilation
   - Abramitzky, Boustan, and Eriksson (2014): Assimilation
   - Borjas (1987): A model of self-selection
   - Abramitzky, Boustan, and Eriksson (2012): Testing the Roy model
4. Looking ahead

Basic set-up of the model

- Two countries: country 0 (home) and country 1 (US)
- Decompose earnings into observables (μ), unobservables (ϵ_i):

$$\ln w_{ij} = \mu_j + \epsilon_{ij}$$

$$\begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{0,1} \\ \sigma_{0,1} & \sigma_1^2 \end{pmatrix} \right)$$

- From here, drop i subscripts
- Correlation coefficient of ϵ_0, ϵ_1: $\rho_{0,1} = \frac{cov(\epsilon_0, \epsilon_1)}{\sigma_0 \sigma_1} = \frac{\sigma_{0,1}}{\sigma_0 \sigma_1}$

Basic set-up of the model (continued)

- Migration cost C
- "Time-equivalent" migration costs $\pi = \frac{C}{w_o}$

Individual decision to migrate determined by sign of I:

$$\begin{aligned} I &= \ln\left(\frac{w_1}{w_o + C}\right) \\ &= \ln(w_1) - \ln(w_0(1+\pi)) \\ &= \mu_1 + \epsilon_1 - \mu_0 - \epsilon_0 - \ln(1+\pi) \\ &\approx (\mu_1 - \mu_0 - \pi) + (\epsilon_1 - \epsilon_0) \end{aligned}$$

Defining $v \equiv \epsilon_1 - \epsilon_0$, migration rate P is:

$$\begin{aligned} P &= \Pr[\epsilon_1 - \epsilon_0 > -(\mu_1 - \mu_0 - \pi)] \\ &= \Pr[v > (\mu_0 - \mu_1 + \pi)] \end{aligned}$$

Basic set-up of the model (continued)

- Define $z = \frac{\mu_0 - \mu_1 + \pi}{\sigma_v}$
- ϕ, Φ: PDF and CDF of standard normal distribution
- $v = \epsilon_1 - \epsilon_0 \Rightarrow s = \frac{v}{\sigma_v}$ follows a standard normal

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

Migration rate increasing in mean US wages ($\frac{\partial P}{\partial \mu_1} > 0$), decreasing in mean home wages ($\frac{\partial P}{\partial \mu_0} < 0$), and decreasing in costs of migrating ($\frac{\partial P}{\partial \pi} < 0$)

Useful facts (in case anyone is rusty)

[Property 1.] If a vector of random variables $X \sim N(\mu, \Sigma)$, then $AX + b \sim N(A\mu + b, A\Sigma A')$.

[Property 2.] If $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_y^2 \end{pmatrix}\right)$, then $(Y|X = x) \sim N\left(\mu_y + \rho_{X,Y}(\frac{\sigma_y}{\sigma_x})(x - \mu_x), \sigma_y^2(1 - \rho_{X,Y}^2)\right)$.

[Property 3.] For any non-stochastic function $f(\cdot)$ and $X = f(W)$, $E(Y|X) = E(E(Y|W)|X)$.

[Property 4.] Let $\phi(z)$ and and $\Phi(z)$ denote the PDF and CDF of the standard normal distribution, respectively. If $\frac{v}{\sigma_v} \sim N(0, 1)$, then $E\left(\frac{v}{\sigma_v} | \frac{v}{\sigma_v} > z\right) = \frac{\phi(z)}{1-\Phi(z)}$; we refer to this expression as the Inverse Mills Ratio. Because $\phi(z) = \phi(-z)$ and $1 - \Phi(z) = \Phi(-z)$, we can also write the Inverse Mills Ratio as $\lambda(z) = \frac{\phi(-z)}{\Phi(-z)}$.

Analyzing self-selection

To analyze self-selection, Borjas derives expressions comparing $E(\ln w_0 | I > 0)$ and $E(\ln w_1 | I > 0)$: that is, for individuals who immigrate compare average earnings in country 0 and average earnings in country 1

Let's start with $E(\ln w_o | I > 0)$, which can be re-written as follows:

$$\begin{aligned} E(\ln w_0 | I > 0) &= E\left(\mu_0 + \epsilon_0 | \frac{v}{\sigma_v} > z\right) \\ &= \mu_0 + E\left(\epsilon_0 | \frac{v}{\sigma_v} > z\right) \\ &= \mu_0 + \sigma_0 E\left(\frac{\epsilon_0}{\sigma_0} | \frac{v}{\sigma_v} > z\right) \end{aligned}$$

Analyzing self-selection (continued)

Let's derive a simplified version of the $E\left(\frac{\epsilon_0}{\sigma_0}|\frac{v}{\sigma_v} > z\right)$ term:

1. Because ϵ_0 and ϵ_1 are jointly normally distributed, applying Property 1 you can show that ϵ_0 and $v \equiv \epsilon_1 - \epsilon_0$ are jointly normally distributed: $\begin{pmatrix} \epsilon_0 \\ \epsilon_1 - \epsilon_0 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{0,1} - \sigma_0^2 \\ \sigma_{0,1} - \sigma_0^2 & \sigma_0^2 + \sigma_1^2 - 2\sigma_{0,1} \end{pmatrix}\right)$.
2. Given that ϵ_0 and $v \equiv \epsilon_1 - \epsilon_0$ are jointly normally distributed, applying Property 2 you can show that $E(\epsilon_0|v) = \rho_{0,v}(\frac{\sigma_0}{\sigma_v})v$, where $\rho_{0,v} = \frac{\sigma_{0,v}}{\sigma_0 \sigma_v}$. Simplifying implies $E(\epsilon_0|v) = \frac{\sigma_{0,v}}{\sigma_v^2}v$.
3. Applying Property 3, you can show that $E(\frac{\epsilon_0}{\sigma_0}|\frac{v}{\sigma_v} > z) = E(E(\frac{\epsilon_0}{\sigma_0}|\frac{v}{\sigma_v})|\frac{v}{\sigma_v} > z)$.

Analyzing self-selection (continued)

Putting this together, let's simplify $E(\frac{\epsilon_0}{\sigma_0}|\frac{v}{\sigma_v})$. Let $s = \frac{v}{\sigma_v} \sim N(0,1)$. Applying Property 2, $E(\epsilon_0|s) = \frac{\sigma_{0,s}}{\sigma_s^2}s$. Substituting $\rho_{0,v} = \frac{\sigma_{0,v}}{\sigma_0\sigma_v}$ gives:

$$\begin{aligned}
E(\frac{\epsilon_0}{\sigma_0}|\frac{v}{\sigma_v}) &= \frac{1}{\sigma_0}E(\epsilon_0|s) \\
&= \frac{1}{\sigma_0}\frac{\sigma_{0,s}}{\sigma_s^2}s \\
&= \frac{1}{\sigma_0}\frac{\frac{1}{\sigma_v}cov(v,\epsilon_0)}{1}\frac{v}{\sigma_v} \\
&= \frac{\sigma_{0,v}}{\sigma_0\sigma_v}\frac{v}{\sigma_v} \\
&= \rho_{0,v}\frac{v}{\sigma_v}
\end{aligned}$$

Analyzing self-selection (continued)

The Inverse Mills Ratio $\frac{\phi(z)}{1-\Phi(z)}$ is the conditional expectation for a standard normal truncated on the left by z. Using this notation:

$$\begin{aligned} E(\ln w_0 | I > 0) &= \mu_0 + \sigma_0 \rho_{0,v} E\left(\frac{v}{\sigma_v} | \frac{v}{\sigma_v} > z\right) \\ &= \mu_0 + \sigma_0 \rho_{0,v} \left(\frac{\phi(z)}{1-\Phi(z)}\right) \end{aligned}$$

- Similar expression for $E(\ln w_1 | I > 0) : \mu_1 + \sigma_1 \rho_{1,v} \left(\frac{\phi(z)}{1-\Phi(z)}\right)$

Analyzing self-selection (continued)

- Useful to re-write equations for $E(\ln w_0 | I > 0)$, $E(\ln w_1 | I > 0)$
- Substituting $\sigma_{0,v} = cov(\epsilon_0, v) = E[\epsilon_0 \cdot (\epsilon_1 - \epsilon_0)] = \sigma_{0,1} - \sigma_0^2$:

$$\begin{aligned} E(\ln w_0 | I > 0) &= \mu_0 + \sigma_0 \rho_{0,v} \left(\frac{\phi(z)}{1 - \Phi(z)} \right) \\ &= \mu_0 + \frac{\sigma_0 \sigma_1}{\sigma_v} \left(\rho_{0,1} - \frac{\sigma_0}{\sigma_1} \right) \left(\frac{\phi(z)}{1 - \Phi(z)} \right) \end{aligned}$$

- Substituting $\sigma_{1,v} = \sigma_1^2 - \sigma_{0,1}$:

$$\begin{aligned} E(\ln w_1 | I > 0) &= \mu_1 + \sigma_1 \rho_{1,v} \left(\frac{\phi(z)}{1 - \Phi(z)} \right) \\ &= \mu_1 + \frac{\sigma_0 \sigma_1}{\sigma_v} \left(\frac{\sigma_1}{\sigma_0} - \rho_{0,1} \right) \left(\frac{\phi(z)}{1 - \Phi(z)} \right) \end{aligned}$$

Analyzing self-selection (continued)

In order to understand the position of migrants in the distribution of workers in each country, we want to know the signs of Q_0 and Q_1:

$$Q_0 \equiv E(\epsilon_0 | I > 0) = \frac{\sigma_0 \sigma_1}{\sigma_v} \left(\rho_{0,1} - \frac{\sigma_0}{\sigma_1} \right) \left(\frac{\phi(z)}{1 - \Phi(z)} \right)$$

$$Q_1 \equiv E(\epsilon_1 | I > 0) = \frac{\sigma_0 \sigma_1}{\sigma_v} \left(\frac{\sigma_1}{\sigma_0} - \rho_{0,1} \right) \left(\frac{\phi(z)}{1 - \Phi(z)} \right)$$

Four cases of immigrant selection

1. **Positive selection:** $Q_0 > 0$ **and** $Q_1 > 0$**.** Arises $\Leftrightarrow \rho_{0,1} > \frac{\sigma_0}{\sigma_1}$. Migrants drawn from upper tail, fall in upper tail. Borjas's example: high-skilled workers migrating from Western Europe.
2. **Negative selection:** $Q_0 < 0$ **and** $Q_1 < 0$**.** Arises $\Leftrightarrow \rho_{0,1} > \frac{\sigma_1}{\sigma_0}$. Migrants drawn from lower tail, fall in lower tail. Borjas's example: US safety net may draw low-skilled immigrants.
3. **Refugee selection:** $Q_0 < 0$ **and** $Q_1 > 0$**.** Arises $\Leftrightarrow \rho_{0,1} < \min(\frac{\sigma_0}{\sigma_1}, \frac{\sigma_1}{\sigma_0})$. Migrants drawn from lower tail, fall in upper tail. Borjas's example: Communist takeover.
4. **No fourth case:** $Q_0 > 0$ **and** $Q_1 < 0$**.** Mathematically, this case is ruled out because it would require $\rho_{0,1} > 1$.

Note: Joint normality assumption

As an econometrician, what you observe is individuals' migration decisions (whether they moved to US or stayed), data on US wages of migrants $E(\ln w_1 | I > 0)$, and data on home wages of non-migrants $E(\ln w_0 | I \leq 0)$.

- Given this data, we would like to know the joint distribution of $\ln w_0$ and $\ln w_1$ so that we can make statements about where migrants fall in the home and US country income distributions.
- Heckman and Honore (1990) show that the joint normality assumption in the original Roy model allows you to identify the joint distribution of $\ln w_0$ and $\ln w_1$ in a single cross section of data, but that without this assumption the model is no longer identified.
- French and Taber *Handbook* chapter gives some intuition

1. Preliminaries: Overview of 14.662, Part II
2. A model of self-selection: The Roy model
3. Application: Immigration
   - Chiswick (1978) and Borjas (1985): Assimilation
   - Abramitzky, Boustan, and Eriksson (2014): Assimilation
   - Borjas (1987): A model of self-selection
   - Abramitzky, Boustan, and Eriksson (2012): Testing the Roy model
4. Looking ahead

Testing the Roy model

Mixed evidence on the Roy model

- Chiquiar and Hanson (2005): evidence against negative selection of Mexican migrants (as would be predicted by the Roy model)
- Focus here: Abramitsky, Boustan, Eriksson (2012)
  - Age of mass migration (1850-1913): open borders
  - Focus on Norwegian migrants: Roy model predicts negative selection

1900 income distributions: US and Norway

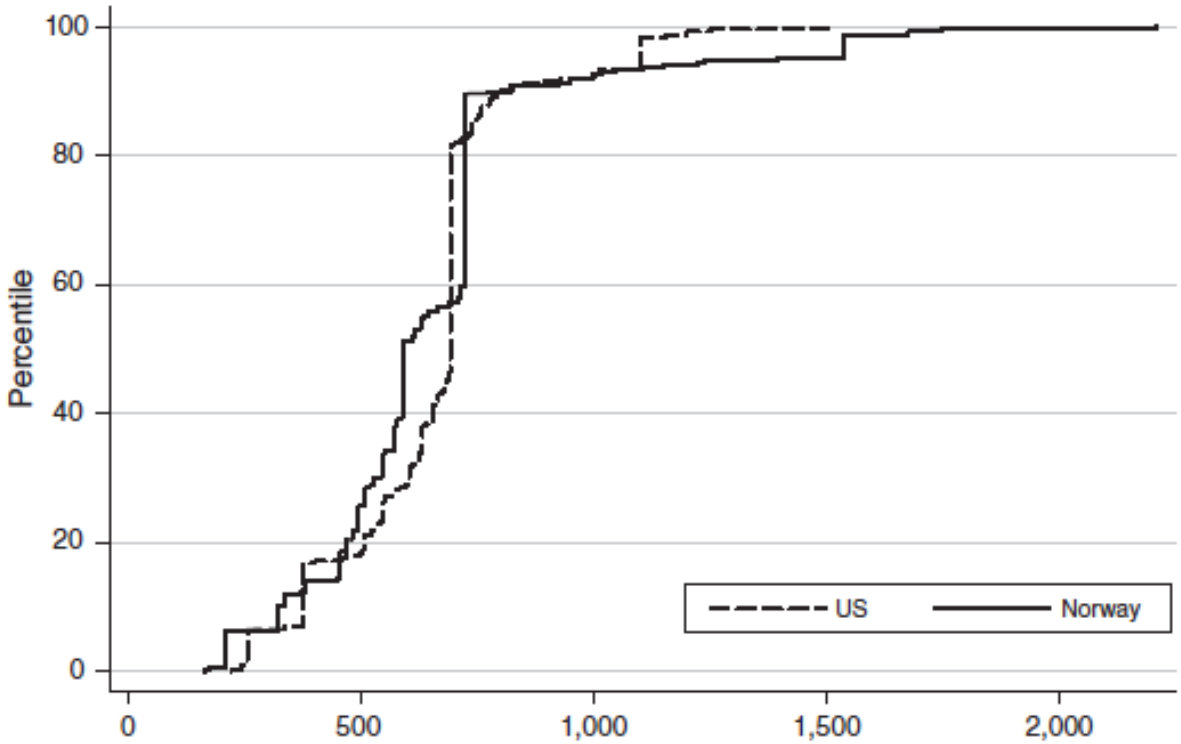


FIGURE 1. CUMULATIVE INCOME DISTRIBUTION FUNCTIONS IN THE UNITED STATES AND NORWAY IN 1900

Notes: US and Norwegian distributions contain all men aged 38 to 50 in the respective censuses of 1900. The x-axis is scaled in 1900 US dollars. Individuals are assigned the mean earnings for their occupation and are arrayed from lowest- to highest-paid occupations. The Norwegian distribution is rescaled to have the same mean as the US distribution (the actual Norwegian and US means are US$(1900)350 and US$(1900)643, respectively).

Data and analysis

Another heroic data effort:

- Two fully digitized Norwegian censuses (1865 and 1900)
- Newly-digitized dataset of all Norwegian-born men in the US in 1900 using now-publicly-available census records
- Match migrants and stayers based on names and ages
- Earnings-related outcome: Occupation

Evidence of negative selection in urban sample (mixed for rural sample)

- Two pieces of direct evidence:
  1. Compare occupational distributions of leavers/stayers
  2. Compare fathers of migrants/non-migrants
- Indirect evidence: compare OLS/family FE returns to migration

Comparing occupational distributions of leavers/stayers

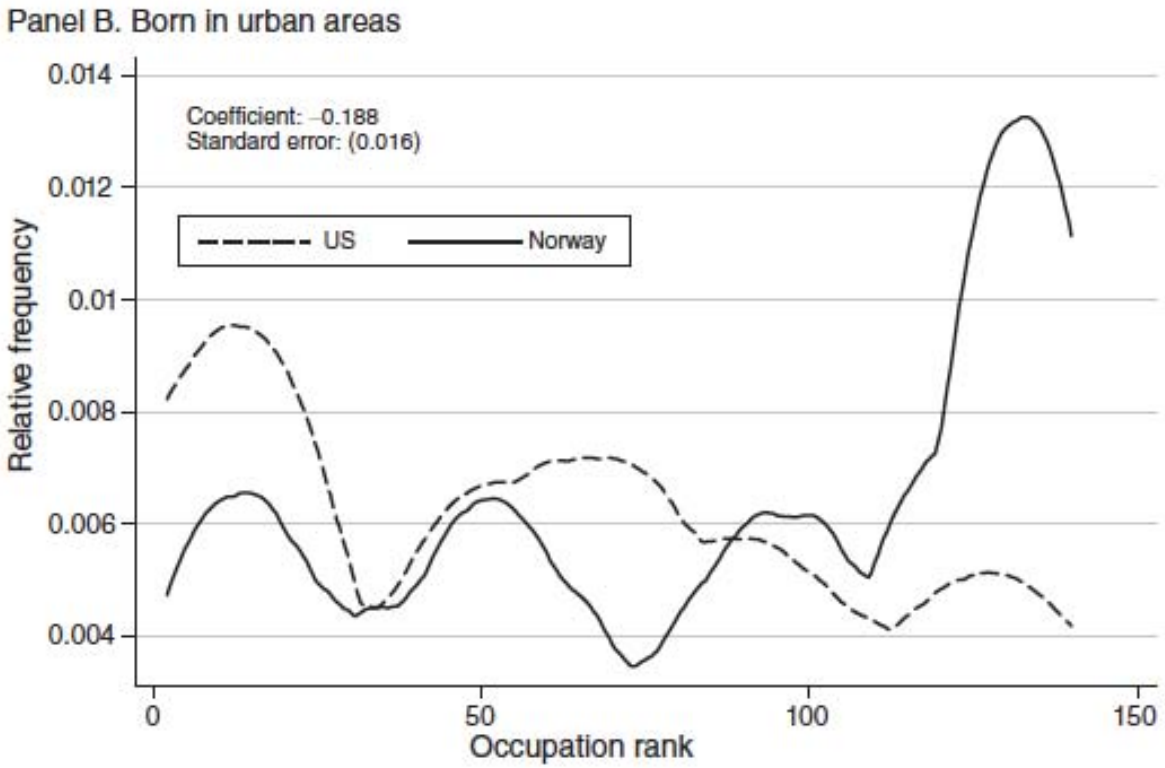


FIGURE 3. COMPARING THE OCCUPATIONAL DISTRIBUTIONS OF NORWEGIAN-BORN MEN IN THE UNITED STATES AND NORWAY IN 1900

Notes: Each figure presents the relative frequency of 144 earning categories (representing 189 distinct occupations) for Norwegian-born men in the United States and in Norway. All men are assigned the mean US earnings in their occupation. Men are divided by rural or urban place of birth. Farmers, the largest occupational category, is excluded from the figure for reasons of scale. We report coefficients and standard errors from OLS regressions of ln(earnings) on a dummy for living in the United States controlling for a quadratic in age.

Comparing fathers of migrants/non-migrants

TABLE 4—ECONOMIC OUTCOMES OF HOUSEHOLD HEADS WITH MIGRANT AND NONMIGRANT SONS, 1865

	All households		Households with matched sons	
Dependent variables	Mean	Coefficient on migrant HH	Mean	Coefficient on migrant HH
Panel A. Urban				
Occ. > median	0.593	−0.001	0.583	−0.030
		(0.022)		(0.042)
Occupational income	428.27	−27.501	440.47	−26.555
		(10.216)		(20.798)
Assets	0.260	−0.030	0.252	−0.058
		(0.018)		(0.035)
N		4,038		1,074
Panel B. Rural				
Occ. > median	0.608	0.008	0.577	−0.054
		(0.014)		(0.032)
Occupational income	321.21	6.092	315.30	−9.077
		(3.847)		(9.072)
Assets	0.665	−0.032	0.613	−0.035
		(0.012)		(0.028)
Match tax records	0.130	−0.037	0.134	−0.040
		(0.009)		(0.021)
Property tax bill	2.759	−0.372	2.821	0.044
$N = 1{,}410; 300$		(0.307)		(0.887)
N		12,177		2,499

Notes: Results for Match 1. The left-hand panel includes all households while the right-hand panel considers only households in which all sons can be matched to 1900. In each panel, sample means are reported in Column 1 and the coefficient and standard error from a regression of each dependent variable on an indicator for being the head of a migrant household is reported in Column 2. Regressions also control for a quadratic in head's age and a series of province dummies. We assign income levels to household heads using mean Norwegian earnings by occupation in 1900. Above-median occupations are those earning more than US$(1900)311 (urban) and US$(1900)393 (rural) per year. Assets is an indicator variable equal to 1 for men who own a business, own land, or are master craftsmen in an artisanal workshop. Tax records refers to the 1886 Land Registers. For fathers who match to the tax records, we report the value of the property tax bill in 1900 US dollars. The data is provided in speciedalers and marks. In 1875, one speciedaler was equal to four Norwegian kroner. The number of households that can be matched to the property tax records is reported below that variable.

Comparing OLS/family FE returns to migration

1. OLS: compare earnings of migrants with earnings of stayers
2. FE: compare earnings of migrants with earnings of stayer brothers

If the OLS estimate measures the return to migration plus a selection term, and if migrants are negatively selected, then the OLS estimate will be smaller than the family fixed effect estimate.

Of course, family FE estimate is not free of selection concerns

- Appendix presents IV analysis using gender composition of a man's siblings and birth order as instruments for migration

Comparing OLS/family FE returns to migration

TABLE 3—OLS AND WITHIN-HOUSEHOLD ESTIMATES OF THE RETURN TO MIGRATION. HOUSEHOLDS WITH TWO OR MORE MEMBERS IN THE MATCHED SAMPLE

Dependent variable = ln(earnings); Coefficient on = 1 if migrant	Full sample, 1865	Rural, 1865	Urban, 1865
Panel A. Unweighted			
OLS	0.545	0.607	0.384
	(0.027)	(0.034)	(0.044)
Within household	0.511	0.508	0.508
	(0.035)	(0.045)	(0.057)
Chi-squared	1.49	7.47	8.31
p-value	0.2218	0.0063	0.0039
N	2,655	1,823	832
Number of migrant-stayer pairs	326	167	159
Panel B. Weighted			
OLS	0.586	0.609	0.443
	(0.029)	(0.033)	(0.067)
Within household	0.542	0.529	0.561
	(0.039)	(0.042)	(0.049)
Chi-squared	2.13	4.60	5.65
p-value	0.1441	0.0320	0.0175
N	2,241	1,666	306
Number of migrant-stayer pairs	269	140	129

Notes: Each cell contains coefficient estimates and standard errors from regressions of ln(earnings) on a dummy variable equal to one for individuals living in the United States in 1900. Regressions also include controls for age and age squared. In each panel, the first row conducts an OLS regression for the restricted sample of households that have at least two matched members in the dataset and the second row adds household fixed effects. Panel B contains results from regressions weighted to reflect the urban status (full sample only), asset holdings, and occupational distribution of fathers in the full population. We conduct chi-squared tests of the null hypothesis that the OLS and within-household coefficients are equal.

Take-aways

1. First, this is a very recent paper providing new, interesting evidence testing the predictions of the Roy model. This is a classic question, but that doesn't mean that there isn't room for good papers on it!
2. Second, this paper highlights the value of looking for the 'right' empirical setting and of constructing the 'right' data
   - Testing for selection: Open borders
   - Data: empirical estimates are basically just summary statistics, but that's because the authors did an enormous amount of work to construct data that enables transparent empirical tests
3. Finally, this is a great example of how economic history can overlap nicely with core questions in labor economics
   - Useful to keep in mind for your own research, in addition to more 'traditional' focus of economic history, which is shedding light on the long-run impacts of economic phenomena

1. Preliminaries: Overview of 14.662, Part II
2. A model of self-selection: The Roy model
3. Application: Immigration
   - Chiswick (1978) and Borjas (1985): Assimilation
   - Abramitzky, Boustan, and Eriksson (2014): Assimilation
   - Borjas (1987): A model of self-selection
   - Abramitzky, Boustan, and Eriksson (2012): Testing the Roy model
4. Looking ahead

Looking ahead

Two additional applications of the Roy model:

- Health care: Chandra and Staiger (2007)
- Redistribution: Abramitsky (2009)

Please comment on Chandra and Staiger (2007)

14.662 Labor Economics II

Spring 2015