NBER WORKING PAPER SERIES

SELF-SELECTION OF EMIGRANTS: THEORY AND EVIDENCE ON STOCHASTIC DOMINANCE IN OBSERVABLE AND UNOBSERVABLE CHARACTERISTICS

George J. Borjas
Ilpo Kauppinen
Panu Poutvaara

Working Paper 21649
http://www.nber.org/papers/w21649

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2015

We thank participants at Norface Migration Network Conference, Journées Louis-André Gérard-Varet, EEA and CEMIR Junior Economist Workshop in 2013, the Alpine Population Conference, UCFS Workshop, CESifo ESP area conference and VfS annual conference in 2015 and seminars at UC Irvine, ETH Zurich, Labour Institute for Economic Research, VATT, University of Linz, and University of Salzburg for valuable comments. Financial support from Leibniz Association (SAW-2012-ifo-3) is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Self-Selection of Emigrants: Theory and Evidence on Stochastic Dominance in Observable and Unobservable Characteristics

George J. Borjas, Ilpo Kauppinen, and Panu Poutvaara

NBER Working Paper No. 21649
October 2015
JEL No. F22,J61

ABSTRACT

We show that the Roy model has more precise predictions about the self-selection of migrants than previously realized. The same conditions that have been shown to result in positive or negative selection in terms of expected earnings also imply a stochastic dominance relationship between the earnings distributions of migrants and non-migrants. We use the Danish full population administrative data to test the predictions. We find strong evidence of positive self-selection of emigrants in terms of pre-emigration earnings: the income distribution for the migrants almost stochastically dominates the distribution for the non-migrants. This result is not driven by immigration policies in destination countries. Decomposing the self-selection in total earnings into self-selection in observable characteristics and self-selection in unobservable characteristics reveals that unobserved abilities play the dominant role.

George J. Borjas
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138
and NBER
gborjas@harvard.edu

Ilpo Kauppinen
VATT Institute for Economic Research
Arkadiankatu 7, 00101
Helsinki, Finland
ilpo.kauppinen@vatt.fi

Panu Poutvaara
Ifo Institute
Poschingerstraße 5
81679 Munich
Germany
and University of Munich
poutvaara@ifo.de

Self-Selection of Emigrants: Theory and Evidence on Stochastic Dominance in Observable and Unobservable Characteristics*

George J. Borjas, Ilpo Kauppinen, and Panu Poutvaara

1. Introduction

A central finding in the economic literature on international migration is that emigrants are not randomly selected from the population of the source countries. The nature of the non-random selection affects the level and the distribution of welfare through two major channels. First, the skill distribution of migrants affects the wage structure in both sending and receiving countries (Borjas 2003). A second effect takes place through the public sector. Immigration creates a fiscal surplus in the receiving country if and only if the net present value of the tax payments of immigrants exceeds the net present value of the costs they impose. Both the immigration of net recipients and the emigration of net payers pose a challenge to the public treasury (Wildasin 1991; Sinn 1997).

Beginning with Borjas (1987), there has been a great deal of interest in deriving and empirically testing models that predict how migrants differ from non-migrants. Many of these studies rely on an application of the Roy model of occupational self-selection. As long as skills are sufficiently transferable across countries, the sorting of persons across countries is mainly determined by international differences in the rate of return to skills. A country like the United States would then attract high-skilled workers from more egalitarian countries (i.e., countries offering relatively low rates of return to skills) and low-skilled workers from countries with greater income inequality (i.e., countries offering higher rates of return to skills). The evidence indeed suggests a negative cross-section

* We thank participants at Norface Migration Network Conference, Journées Louis-André Gérard-Varet, EEA and CEMIR Junior Economist Workshop in 2013, the Alpine Population Conference, UCFS Workshop, CESifo ESP area conference and VfS annual conference in 2015 and seminars at UC Irvine, ETH Zurich, Labour Institute for Economic Research, VATT, University of Linz, and University of Salzburg for valuable comments. Financial support from Leibniz Association (SAW-2012-ifo-3) is gratefully acknowledged.

correlation between the earnings of immigrants in the United States and income inequality in the source countries.[1]

Although the existing literature on immigrant selection focuses mainly on the U.S. context or on migration flows from poor to rich countries, there are also sizable migration flows between rich countries. According to the United Nations (2013), 21.9 million persons from EU15 countries now live outside their birthplace, with 42 percent of these migrants living in other EU15 countries and an additional 13 percent living in the United States.[2]

This paper examines the self-selection of emigrants from Denmark, one of the richest and most redistributive European welfare states. In 2013, over a quarter million Danes lived outside Denmark (corresponding to about 5 percent of the Danish-born population), with 50 percent of the migrants living in other EU15 countries and 13 percent in the United States (United Nations, Department of Economic and Social Affairs 2013). Because the returns to skills in Denmark are relatively low, the canonical Roy model predicts that the emigrants should be positively selected in the sense that the expected earnings of the migrants exceed the expected earnings of the stayers.[3] However, there have been few systematic studies of the self-selection of migrants from a relatively egalitarian country to see whether this is indeed the case.[4]

---

[1] Related cross-country studies include Cobb-Clark (1993) and Bratsberg (1995). Grogger and Hanson (2011) examine the selection of migrants across a broad range of countries using an alternative theoretical framework where individuals maximize linear utility and migration is driven by absolute earnings differences between high and low-skilled workers.

[2] The EU15 countries refer to the member states of the European Union prior to the expansion on May 1, 2004.

[3] For comparisons of gross wage premia from tertiary education across countries see Boarini and Straus (2010). A recent paper studying returns to cognitive skills is Hanushek et al. (2015). The study finds significant cross-country differences. Moreover, the returns are relatively low in Denmark as well as in other Nordic countries, and high in the United States, Germany and the United Kingdom, which also are among the most popular destinations of Danish migrants.

[4] Studies of the selection of migrants across developed countries include Lundborg (1991), Pirttilä (2004), Kleven et al. (2014), and Junge et al. (2014). Many studies also examine selection issues in a historical context; see Wegge (1999, 2002), Abramitzky and Braggion (2006), Abramitzky, Boustan, and Eriksson (2012), Ferrie (1996), and Margo (1990).

Our theoretical analysis shows that the canonical framework does not only have predictions about the difference between the expected earnings of migrants and non-migrants, which is the basis for the standard definition of positive or negative selection in the literature, but also about the stochastic ordering of the two earnings distributions. We show that the same conditions that predict that migrants are positively self-selected in the sense of a difference in expected incomes also predict that the income distribution of the migrants will first-order stochastically dominate the income distribution of the non-migrants. The theory also distinguishes between selection in observable and selection in unobservable characteristics.

Our empirical analysis uses the Danish full population administrative data to analyze how migrants and non-migrants differ in their pre-emigration earnings and other observable characteristics. To shed light on the role of unobservable characteristics in the selection process, we investigate how migrants and non-migrants differ in terms of unobservable earnings ability, as measured by residuals from Mincerian earnings regressions. Our empirical results are in line with the predictions of the model: Danish emigrants are indeed positively self-selected both in terms of earnings and in terms of residuals from the wage regressions. Following our reframing of the canonical Roy framework in terms of the concept of stochastic dominance, our study specifically tests for whether the earnings distribution of the emigrants stochastically dominates that of the stayers (as would be predicted by the model). The evidence confirms this strong theoretical prediction over most of the support of the earnings distribution.

Our study is related to the flurry of recent papers that examine the selection of migrants from Mexico to the United States. The pioneering analysis of Chiquiar and Hanson (2005) merged information from the U.S. census on the characteristics of the Mexican migrants with information from the Mexican census on the characteristics of the Mexican non-migrants. Because the merged data did not report the earnings of migrants *prior* to the move, pre-migration earnings were predicted based on observable characteristics of the migrants. This "counterfactual" empirical exercise suggested that Mexican emigrants were located in the medium-high range of the Mexican wage distribution. The finding of

intermediate selection in the Mexican context does not seem consistent with the basic implications of the Roy model because the rate of return to skills is far larger in Mexico than in the United States. More recent studies by Fernández-Huertas Moraga (2011) and Kaestner and Malamud (2014) use survey data that report the *actual* pre-migration earnings and find evidence of negative selection. They also conclude that part of the negative selection can be traced to the unobservable characteristics that determine a migrant's earnings.

The important role played by unobservable characteristics implies that constructing a counterfactual earnings distribution for the migrants based on observable characteristics can greatly bias the nature of the selection revealed by the data. Our findings suggest that the use of such a counterfactual distribution will tend to *understate* the true selection in earnings, so that the selection implied by the counterfactual distribution is far weaker than the true selection—regardless of whether there is positive or negative selection. The numerical bias that results from using the counterfactual estimation is sizable in the Danish context: more than half of the difference between the expected earnings of migrants and non-migrants arises because of differences in unobserved characteristics.

The paper is organized as follows. Section 2 sketches the economic theory underlying the analysis and derives theoretical predictions concerning the self-selection of emigrants, using the notion of stochastic dominance as a unifying concept. Section 3 introduces and describes the unique population data that we use and reports some summary statistics. Sections 4 and 5 present the main empirical findings. In section 4, we examine the selection in terms of observed pre-migration earnings. We present a statistical method for testing the theoretical implication that the earnings distribution of the emigrants should stochastically dominate the corresponding distribution of the non-migrants. Section 5 extends the empirical work by examining the selection that occurs in the unobserved component of earnings. Section 6 evaluates the bias that results from predicting the pre-migration earnings of emigrants from the earnings distribution of non-migrants. Section 7 examines whether the selection of persons moving to other EU15 countries differs from the selection of migrants moving to countries where immigration restrictions come into play.

We find that immigration restrictions have little effect on the selection of emigrants. Finally, Section 8 summarizes the study and draws some lessons for future research.

2. Theoretical framework

Previous literature on the self-selection of migrants has focused on the conditional expectations of earnings distributions among migrants and stayers. In this section, we derive a novel result: the Roy model implies that under certain conditions, the earnings distribution of migrants first-order stochastically dominates, or is stochastically dominated by, the earnings distribution of stayers. In a bivariate normal framework, it turns out that the conditions required for stochastic dominance are identical to the conditions that determine the nature of self-selection in terms of expected earnings.

We also decompose self-selection into two components, one that is determined by differences in returns to observable skills between source and host country, and one that is determined by differences in returns to unobservable skills. The distinction between observable and unobservable skills, of course, depends on the empirical framework and on the data that is being used; observable skills include the variables explaining earnings that are included in the data, while the component of earnings that is left unexplained by the data is the unobservable skill component. Even though the content of the two components differs among data sets, it is likely that a major part of migrant self-selection is determined by the unobservable component simply because "observables" tend to explain a relatively small fraction of the variance in earnings.

We take as our starting point the migration decision faced by potential migrants in a two-country framework, in line with Borjas (1987) and subsequent literature. Residents of the source country (country 0) consider migrating to the destination country (country 1), and the migration decision is assumed to be irreversible. To simplify the presentation, we focus on a single observed skill characteristic s and suppress the subscript that indexes a particular individual. For concreteness, the variable s can be thought of as giving the worker's years of educational attainment, but it includes all the characteristics affecting

individual's income that are observed in a given set of data. Residents of the source country face the earnings distribution:

(1) $$\log w_0 = \alpha_0 + r_0 s + \varepsilon_0,$$

where w_0 gives the wage in the source country; r_0 gives the rate of return to observable skills; and the random variable ε_0 measures individual-specific productivity shocks resulting from unobserved characteristics and is normally distributed with mean zero and variance σ_0^2. The distribution of observable skills in the source country's population is given by $s = \mu_s + \varepsilon_s$, where the random variable ε_s is also assumed to be normally distributed with mean zero and variance σ_s^2.

If the entire population of the source country were to migrate, this population would face the earnings distribution:

(2) $$\log w_1 = \alpha_1 + r_1 s + \varepsilon_1,$$

where the random variable ε_1 is normally distributed with mean zero and variance σ_1^2.

For analytical convenience, we assume that $Cov(\varepsilon_0, \varepsilon_s) = Cov(\varepsilon_1, \varepsilon_s) = 0$, so that the individual-specific unobserved productivity shocks (i.e., the "residuals" from the regression line) are independent from observable characteristics.[5] The correlation coefficient between ε_0 and ε_1 equals ρ_{01}. It is also worth noting that the random variable ε_s is individual-specific and has the same value for the same individual in both countries, whereas ε_0 and ε_1 are both individual- and country-specific.

[5] A more realistic assumption would be that the correlation between observed and unobserved skills is positive. However, allowing for positive correlation does not change the qualitative predictions of the model.

Equations (1) and (2) completely describe the earnings opportunities available to persons born in the source country. Assume that the migration decision is determined by a comparison of earnings opportunities across countries net of migration costs C. Define the index function:

$$(3) \qquad I = \log\left(\frac{w_1}{w_0+C}\right) \approx [(\alpha_1 - \alpha_0) + (r_1 - r_0)\mu_s - \pi] + [(r_1\varepsilon_s + \varepsilon_1) - (r_0\varepsilon_s + \varepsilon_0)]$$
$$= \Delta\mu + (v_1 - v_0),$$

where π gives a "time-equivalent" measure of migration costs ($\pi = C/w_0$). The cross-country difference in earnings net of the time-equivalent migration cost for an individual with average observed and unobserved characteristics is given by $\Delta\mu = [(\alpha_1 - \alpha_0) + (r_1 - r_0)\,\mu_s - \pi]$. The difference in earnings attributable to individual deviation from average characteristics is given by $(v_1 - v_0)$, where $v_i = (r_i\,\varepsilon_s + \varepsilon_i)$ for $i \in \{0,1\}$. A person emigrates if the index $I > 0$, and remains in the origin country otherwise.

Migration costs vary among persons—but the sign of the correlation between costs (whether in dollars or in time-equivalent terms) and skills (both observed and unobserved) is ambiguous and difficult to determine. The heterogeneity in migration costs can be incorporated to the model by assuming that the distribution of the random variable π in the source country's population is given by $\pi = \mu_\pi + \varepsilon_\pi$, where μ_π is the mean level of migration costs in the population, and ε_π is a normally distributed random variable with mean zero and variance σ_π^2. However, Borjas (1987) and Chiquiar and Hanson (2005) show that time-equivalent migration costs do not play a role in the algorithm that determines the selection of emigrants if either those costs are constant (so that $\sigma_\pi^2 = 0$), or if the costs are uncorrelated with skills. For analytical convenience, we assume that time-equivalent

migration costs are constant, so that $\pi = \mu_\pi$.[6] The outmigration rate from the source country is then given by:

$$(4) \qquad Pr(I > 0) = Pr[v^* > -\Delta\mu^*] = 1 - \Phi(-\Delta\mu^*),$$

where $v^* = (v_1 - v_0)/\sigma_v$ is a standard normal random variable; $\Delta\mu^* = \Delta\mu/\sigma_v$; $\sigma_v^2 = Var(v_1 - v_0)$; and Φ is the standard normal distribution function.[7]

In addition to identifying the determinants of the outmigration rate in equation (4), the Roy model lets us examine *which* persons find it most worthwhile to leave the source country.[8] In the following, we examine the self-selection of emigrants along two dimensions: selection in terms of observable skills *s* and selection in terms of unobservable skills ε_0, which together combine into selection in terms of total productivity or earnings, as measured by log w_0.

Let $\mathbf{F_M}(z)$ and $\mathbf{F_N}(z)$ represent the (cumulative) probability distributions of skills or earnings for migrants and non-migrants in the source country, respectively, where z denotes a particular measure of skills (e.g., observable or unobservable characteristics or income). By definition, the probability distribution of migrants $\mathbf{F_M}(z)$ first-order stochastically dominates that of stayers $\mathbf{F_N}(z)$ if:

$$(5) \qquad \boldsymbol{F_M}(z) \leq \boldsymbol{F_N}(z) \forall\, z,$$

---

[6] If π were negatively correlated with skills, the negative correlation would tend to induce the more skilled to migrate, creating a positively selected migrant flow. This would strengthen positive self-selection, and weaken negative self-selection.

[7] It is straightforward to study equation (4) to confirm that the migration rate rises, when mean income in the source country falls, mean income in the host country rises, returns to observed skills in the source country fall, returns to observed skills in the host country rise, time-equivalent migration costs fall and when mean observed skills rise if $r_1 > r_0$ or fall if $r_1 < r_0$.

[8] Throughout the analysis, we assume that $\Delta\mu^*$ is constant. The migration flow is effectively assumed to be sufficiently small that there are no feedback effects on the labor markets of either the source or destination countries.

and there is at least one value of z for which a strict inequality holds.[9] From now on, whenever we refer to stochastic dominance, we mean first-order stochastic dominance.

Equation (5) implies that a larger fraction of the migrants have skills *above* any threshold z^*. Put differently, for any level of skills z^*, the population described by the probability distribution $\mathbf{F_M}$ is more skilled because a larger fraction of the group exceeds that threshold. The migrants, in short, are positively selected. Negative selection, of course, would occur if the reverse was true and $\mathbf{F_N}(z) \leq \mathbf{F_M}(z)\ \forall\ z$, with a strict inequality holding for at least one value of z.

If the skill distribution of migrants stochastically dominates that of non-migrants, the stochastic dominance then also implies the typical definition of positive selection that is based on conditional expectations:

$$E(z|I > 0) > E(z|I \leq 0), \tag{6}$$

so that migrants, on average, are more skilled than stayers. Conversely, if the probability distribution of stayers stochastically dominates that of migrants, and there was negative selection, it would also follow that $E(z|I > 0) < E(z|I \leq 0)$. The converse, however, is not true for a general distribution: A claim of positive selection in expectations, as defined by equation (6), does not imply that the skill distribution of migrants stochastically dominates that of non-migrants.

To derive the stochastic ordering of the skill distributions of migrants and non-migrants, let $f(x, v)$ be a bivariate normal density function, with means (μ_x, μ_v), variances (σ_x^2, σ_v^2) and correlation coefficient ρ. Further, let the random variable v be truncated from below at

[9] An alternative and perhaps more intuitive definition of stochastic dominance is in terms of quantiles. Let $Q_M(P)$ and $Q_N(P)$ be the quantile functions of order P of the skill distributions of migrants and non-migrants. $\mathbf{F_M}(z)$ stochastically dominates $\mathbf{F_N}(z)$ if and only if $Q_M(P) \geq Q_N(P)$ for all $0 \leq P \leq 1$ and there is at least one value of P for which a strict inequality holds.

point *a* and from above at point *b*. Arnold et al. (1993, p. 473) show that the (marginal) moment generating function of the standardized random variable $(x - \mu_x)/\sigma_x$, given the truncation of *v*, is given by:

(7) $$m(t) = \left[\frac{\Phi(\beta-\rho t)-\Phi(\alpha-\rho t)}{\Phi(\beta)-\Phi(\alpha)}\right] e^{t^2/2},$$

where $\alpha = (a - \mu_v)/\sigma_v$; and $\beta = (b - \mu_v)/\sigma_v$.

In terms of the migration decision, the truncation in the random variable $v = v_1 - v_0$ in the sample of migrants is from below and implies that $\alpha = -\Delta\mu^* = k$, and $\beta = \infty$, where *k* is the truncation point. In the sample of stayers, the truncation in *v* is from above, and the truncation points are $\alpha = -\infty$ and $\beta = k$. By substituting these definitions into equation (7), it can be shown that the moment generating functions for the random variable giving the conditional distributions of skill characteristic *x* for migrants and stayers reduce to:

(8)
$$m_F(t) = \left[\frac{1 - \Phi(k - \rho t)}{1 - \Phi(k)}\right] e^{t^2/2}$$

and

(9)
$$m_G(t) = \left[\frac{\Phi(k - \rho t)}{\Phi(k)}\right] e^{t^2/2}.$$

Consider any two distribution functions $\mathbf{F}(z)$ and $\mathbf{G}(z)$. Thistle (1993, p. 307) shows that **F** will stochastically dominate **G** if and only if:

(10) $$m_F(-t) < m_G(-t), \forall\, t > 0,$$

where m_F is the moment generating function associated with distribution **F**; m_G is the moment generating function associated with **G**.

The ranking of the moment generating functions in equation (10) implies we can determine the stochastic ranking of the two distributions by simply solving for the relevant correlation coefficient ρ, and comparing equations (8) and (9). Such a comparison implies that:

$$(11) \qquad \begin{aligned} \boldsymbol{F_M}(z) &< \boldsymbol{F_N}(z), \quad if \quad \rho > 0 \\ \boldsymbol{F_M}(z) &> \boldsymbol{F_N}(z), \quad if \quad \rho < 0. \end{aligned}$$

In other words, migrants are positively selected if ρ > 0, and are negatively selected otherwise. Consider initially the stochastic ranking in observable characteristics. The random variable $x = \varepsilon_s$, and the relevant correlation coefficient ρ is defined by:

$$(12) \qquad \rho = Corr(\varepsilon_s, v_1 - v_0) = \frac{r_0 \sigma_s}{\sigma_v}\left(\frac{r_1}{r_0} - 1\right).$$

Equation (12) shows that the stochastic ordering of the distributions of observable skills of migrants and non-migrants depends only on international differences in the rate of return to observable skills. The skill distribution of migrants will stochastically dominate that of stayers when the rate of return to skills is higher abroad. Conversely, the skill distribution for non-migrants will stochastically dominate the distribution for migrants if the rate of return to observable skills is larger at home.

Consider next the stochastic ordering in the conditional distributions of unobservable skills ε_0. The relevant correlation for determining this type of selection is given by:

$$(13) \qquad \rho = Corr\,(\varepsilon_0, v_1 - v_0) = \frac{\sigma_0}{\sigma_v}\left(\rho_{01}\frac{\sigma_1}{\sigma_0} - 1\right).$$

It follows that the distribution of unobservable skills for migrants stochastically dominates that for non-migrants when $\rho_{01}\frac{\sigma_1}{\sigma_0} > 1$. Note that the necessary condition for positive selection has two components. First, the unobserved characteristics must be "transferable" across countries, so that ρ_{01} is sufficiently high. Second, the residual variance in earnings is larger in the destination country than in the source country. The residual variances σ_0^2 and σ_1^2, of course, measure the "price" of unobserved characteristics: the greater the rewards to unobserved skills, the larger the residual inequality in wages.[10] As long as unobserved characteristics are sufficiently transferable across countries, emigrants are positively selected when the rate of return to unobservable skills is higher in the destination.

Finally, consider the stochastic ranking in "total" productivity. The earnings distribution in the source country given by equation (1) can be rewritten as:

$$\log w_0 = (\alpha_0 + r_0 \mu_s) + (r_0 \varepsilon_s + \varepsilon_0) = (\alpha_0 + r_0 \mu_s) + v_0, \tag{14}$$

where the normally distributed random variable v_0 has mean zero and variance σ_{v0}^2. The relevant correlation for determining the stochastic ranking of the earnings distributions of migrants and non-migrants is:

$$\rho = Corr(v_0, v_1 - v_0) = \frac{1}{\sigma_v}\left[\gamma\left(\frac{r_1}{r_0} - 1\right) + (1-\gamma)\left(\rho_{01}\frac{\sigma_1}{\sigma_0} - 1\right)\right], \tag{15}$$

where $\gamma = r_0^2\sigma_s^2/\sigma_{v0}^2$ and $1 - \gamma = \sigma_0^2/\sigma_{v0}^2$.

The sign of the correlation in equation (15), which determines the nature of the selection in pre-migration earnings, depends on the sign of a weighted average of the selection that occurs in observable and unobservable characteristics. Interestingly, the weight is the

[10] This interpretation of the variances follows from the definition of the log wage distribution in the host country in terms of what the population of the source country would earn if the entire population migrated there. This definition effectively holds constant the distribution of skills.

fraction of the variance in earnings that can be attributed to differences in observable and unobservable characteristics, respectively.

If there is positive (negative) selection in both "primitive" types of skills, there will then be positive (negative) selection in pre-migration earnings. If, however, there are different types of selection in the two types of skills, the selection in each type is weighted by its importance in creating the variance of the earnings distribution. It is well known that observable characteristics (such as educational attainment) explain a relatively small fraction of the variance in earnings (perhaps less than a third). As a result, equation (15) implies that it is the selection in *unobservables* that is most likely to determine the nature of the selection in the pre-migration earnings of emigrants. This implication plays an important role in explaining why the evidence reported in Fernández-Huertas Moraga (2011) and Kaestner and Malamud (2014) conflicts with that of Chiquiar and Hanson (2005).

As mentioned earlier, the stochastic dominance results necessarily imply selection in terms of conditional expectations. In the case of bivariate normal distributions, it follows that the expectation of the earnings distribution of migrants $E(\log w_0 \mid v^* > -\Delta\mu^*)$ is given by:

$$(16) \quad E(logw_0|v^* > -\Delta\mu^*) = \alpha_0 + r_0\mu_s + \frac{r_0\sigma_s^2}{\sigma_v}\left(\frac{r_1}{r_0} - 1\right)\lambda(-\Delta\mu^*) + \frac{\sigma_0^2}{\sigma_v}\left(\rho_{01}\frac{\sigma_1}{\sigma_0} - 1\right)\lambda(-\Delta\mu^*),$$

where $\lambda(-\Delta\mu^*) = \phi(-\Delta\mu^*)/[1 - \Phi(-\Delta\mu^*)] > 0$, and ϕ is the density of the standard normal distribution. As can be seen by examining equation (16), the conditions that determine the self-selection in terms of expectations are the same as the conditions that determine the stochastic ordering of the skill distributions of migrants and non-migrants. In the normal distribution framework that underlies the canonical Roy model, stochastic dominance implies selection in expectations, and vice versa.

In empirical applications, however, the prediction of stochastic dominance is likely to be much less robust than the predictions concerning expectations because testing for

stochastic dominance will require a more rigorous test than simply comparing the average incomes or skills of migrants and non-migrants. If one just compares the averages to find out how migrants are self-selected, the findings can be compatible with the predictions of the Roy-model even if a large number of individuals in the data behave against the stochastic dominance predictions of the model. As a result, establishing an empirical pattern of stochastic dominance provides very strong evidence that differences in skill prices are indeed important in migration decisions.

3. Data

Our analysis uses administrative data for the entire Danish population from 1995 to 2010. The data is maintained and provided by Statistics Denmark and it derives from the administrative registers of governmental agencies that are merged using a unique social security number.[11]

For each year between 1995 and 2004, we identified all Danish citizens aged 25-54 who lived in Denmark during the entire calendar year.[12] We restrict the analysis to persons who worked full time.[13] Migration decisions of part-time workers or of workers outside the labor force may be driven by different factors, and the observed income of these workers may not be indicative of their true earnings potential. The income variable for each year is

[11] All residents in Denmark are legally required to have a social security number. This number is necessary to many activities in daily life, including opening a bank account, receiving wages and salaries or social assistance, obtaining health care, and enrolling in school.

[12] A person's age is measured as of January 1st the year after the reference year.

[13] The administrative data allows the calculation of a variable that measures the amount of "work experience gained" during the calendar year. The maximum possible value for this variable is 1,000. We restrict our sample to workers who have a value of 900 or above, so that our sample roughly consists of persons who worked full time at least 90 percent of the year. In order to measure the work experience gained during a given year, we subtract the value from the previous year from the current value of the variable. Persons who had a missing value for work experience in either of the two years were dropped from the sample. Missing values in this variable typically indicate that the person spent time abroad.

constructed by adding the worker's annual gross labor income and positive values of freelance income.[14]

We merged this information with data from the migration register for the years 1995 through 2010. The migration register reports the date of emigration and the country of destination. Even though it is possible for Danish citizens to emigrate without registering, we expect that the numbers of persons who do so is small as it is a legal requirement for Danish citizens to report emigration. Danish tax laws provide further incentives for migrants to register when they emigrate.

After identifying the population of interest, we determined for each person whether he or she emigrated from Denmark during the following calendar year. If we found that a particular person emigrated, we searched for the person in the migration register for subsequent years to determine if the migrant returned to Denmark at some point in the future, and recorded the date of possible return migration. The migration register includes near-complete information on return migration, as registration in Denmark is required for the return migrant to be eligible for income transfers and to be covered by national health insurance.

To focus on migration decisions that are permanent in nature, we restrict the analysis to migration spells that are at least five years long.[15] We define a migrant as an individual who is found in one of the 1995-2004 cross-sections, who emigrates from Denmark during the following year to destinations outside Greenland or the Faroe Islands, and who stays abroad for at least five years.[16] Individuals who emigrated for less than five years were

---

[14] The information on earnings is taken from the tax records for each calendar year. This variable is considered to be of high quality by Statistics Denmark. Some persons also report negative values for freelance income. These negative values are likely to be due to losses arising from investments and do not reflect the productive characteristics of the individual.

[15] Having stayed abroad for five years predicts longer migration spells. For example 72% of men and 71% of women who left Denmark in 1996 and were still abroad after five years were also abroad after ten years.

[16] Greenland and the Faroe Islands are autonomous regions but still part of Denmark. We have excluded these destinations as many of these migrants could have originated in Greenland or the Faroe Islands, and

removed from the data, and the rest of the population is then classified as non-migrants.[17] The analysis of both migrants and non-migrants is further restricted to only include Danish citizens who do not have an "immigration background."[18]

Table 1 reports summary statistics from the Danish administrative data. The panel data set contains over 6.4 million male and 5.1 million female non-migrants. The construction of the data implies that non-migrants appear in the data multiple times (potentially once in each cross-section between 1995 and 2004). We were able to identify 7323 male and 3436 female migrants. By construction, these migrants are persons who we first observe residing in Denmark and who left the country at some point between 1996 and 2005. As Table 1 shows, the Danish emigrants are younger than the non-migrants, regardless of gender. Despite the age difference, the emigrants earned higher annual incomes in the year prior to the migration than the non-migrants.

We construct a simple measure of "standardized earnings" that adjusts for differences in age, gender, and year effects. Standardized earnings are defined by the ratio of a worker's annual gross earnings to the mean gross earnings of workers of the same age and gender during the calendar year.[19] Table 1 shows that emigrants earn more than non-migrants in terms of standardized earnings. In particular, male emigrants earn about 30 percent more than non-migrants, and female emigrants earn about 20 percent more.

Table 2 reports the number of emigrants moving to different destinations. The largest destinations for both men and women are two other Nordic countries, Sweden and Norway,

---

many would actually be returning home rather than emigrating from Denmark. The exact duration requirements were 1,825 days or longer for long-term migrants.

[17] We also examined the selection of short-term migrants and the qualitative results are similar to those reported below, although the intensity of selection is weaker.

[18] Statistics Denmark defines a person to have "no immigrant background" if at least one of the parents was born in Denmark and the person is/was a Danish citizen. We searched the population registers from 1980 to 2010 for the parents of the persons in our sample, and if a parent was found he or she was required to be a Dane with no immigrant background as well.

[19] Both migrants and non-migrants, as well as shorter-term migrants, are included in these calculations.

as well as the United States, the United Kingdom and Germany.[20] These five countries account for 57 percent of all emigration.

Finally, it is also interesting to summarize the link between education and emigration. Table 3 reports the education distributions for non-migrants and migrants. It is evident that the migrants tend to be more educated than the non-migrants, among both men and women. For example, 50 percent of Danish male non-migrants have a vocational education, as compared to only 30 percent of migrants to non-Nordic destinations. Similarly, the fraction of male migrants to non-Nordic destinations with a Master's degree is 24 percent, whereas only 7 percent of male non-migrants have a Master's degree.

In order to add time dimension, the evolution of the emigration rate is presented in figure 1a for men and in figure 1b for women separately for the whole population and for those with higher education and without higher education. As we are looking at long-term migration, the emigration rates are small, but there is an upward trend. The rate is higher for men and for those with higher education. We also computed the difference between the average of the log standardized earnings, or a degree of selection, for migrants and non-migrants for each year from 1995 to 2004 for men and women separately. The results are reported in figures 2a and 2b. There is a downward trend in the difference for both men and women. The finding makes sense: when the migrants are positively self-selected and the emigration rate gets bigger the average standardized earnings of migrants should get smaller. However, the variation across years is small, so that pooling the data is justified.

To summarize, the descriptive findings suggest a strong degree of positive selection—at least as measured by education and differences in the conditional means of earnings.

[20] If we relax the constraints on labor market status and age to enter the sample, the United Kingdom emerges as the largest destination because of the large number of Danish students who pursue their education there.

4. Selection in pre-migration earnings

This section presents empirical evidence on the self-selection of emigrants from Denmark in terms of standardized pre-emigration earnings. The main empirical finding is that long-term emigrants from Denmark were, in general, much more productive prior to their migration than individuals who chose to stay.

Of course, the summary statistics reported in Table 1 already suggest positive selection among emigrants because their standardized earnings exceeded those of non-migrants. However, differences in conditional averages could be masking substantial differences between the underlying probability distributions. Our theoretical framework predicts that the distribution of earnings for migrants should stochastically dominate that of non-migrants. As a result, our empirical analysis will mainly consist of comparing cumulative distributions of standardized earnings between migrants and non-migrants. An advantage of simply graphing and examining the cumulative distributions is that the analysis does not require any type of kernel density estimation, and that we do not need to impose any statistical assumptions or parametric structure on the data. We will also present kernel density estimates of the earnings density functions as an alternative way of presenting the key insights. Finally, we will derive and report statistical tests to determine if the data support the theoretical prediction of stochastic dominance.

Figure 3a illustrates the cumulative earnings distributions for male migrants to Nordic countries, male migrants to destinations outside Nordic countries, and for male non-migrants. The values of the standardized earnings are truncated at -2 and 2 to make the graphs more tractable.[21] The figure confirms that migrants were positively selected during

[21] The truncation does not alter the results considerably as the shares of observations below the lower and above the upper truncation points are small. Further, the following analysis of differences between cumulative distribution functions does not use truncation. 0.07% of non-migrants, 0.19% migrants to other Nordic countries and 0.11% of migrants to other destinations lie below the lower truncation point. Correspondingly, 0.03% of non-migrants and 0.21% of migrants to destinations outside Nordic countries lie above the upper truncation point. There are no migrants to other Nordic countries above the upper truncation point.

the study period. The cumulative distribution function of standardized earnings of migrants to destinations outside the Nordic countries is clearly located to the right of the corresponding cumulative distribution for non-migrants, as would be the case if the cumulative distribution of migrants stochastically dominates that of non-migrants. The figure also shows that the distribution function for migrants to other Nordic countries is located to the right of that for non-migrants. However, the selection of the migrants to Nordic countries seems weaker. This weaker selection may arise because the rate of return to skills in Nordic countries is relatively low when compared to that in other potential destinations.[22] Figure 3b presents corresponding evidence for women.[23] The main findings are qualitatively similar, but the positive selection seems weaker.

Figure 4a presents the corresponding kernel estimates of the density functions of the logarithm of standardized earnings for men, while Figure 4b presents the respective graphs for women.[24] The density functions again reveal the positive selection of migrants moving outside the Nordic countries, both for men and women.

As is evident from the figures, Kolmogorov-Smirnov tests comparing the earnings distributions for different groups reject the hypothesis that the underlying earnings distributions are the same at a highly significant level. In addition to showing that the cumulative distributions are different, it is also important to determine if the evidence statistically supports the theoretical prediction that the cumulative distribution function of migrants stochastically dominates that of non-migrants. Statistical tests for first-order stochastic dominance are highly sensitive to small changes in the underlying distributions,

---

[22] Moreover, some Danes may live in southern Sweden but work in Denmark. As this type of migration is not related to returns to skills in the destination country this should decrease the estimated selection to Nordic countries.

[23] For women, 0.06% of non-migrants lie below the lower truncation point and 0.00% of non-migrants lie above the higher truncation point. There are no migrants lying below the lower or above the higher truncation point.

[24] Following Leibbrandt et al. (2005) and Fernandes-Huertas Moraga (2011), we use Silverman's reference bandwidth multiplied by 0.75 to prevent over-smoothing. The same bandwidth is used also in all the kernel density estimates reported in subsequent calculations.

making it difficult to rank distributions in many empirical applications.[25] As noted by Davidson and Duclos (2013), it may be impossible to infer stochastic dominance over the full support of empirical distributions if the distributions are continuous in the tails, simply because there is not enough information in the tails for meaningful testing of any statistical hypothesis. It would then make sense to focus on testing stochastic dominance over a restricted range of the distribution. We apply an approach that characterizes the range over which the value of the cumulative distribution function for non-migrants is statistically significantly larger than that of non-migrants.

In particular, we calculate the difference between the cumulative distribution functions with confidence intervals. To calculate the confidence intervals we use tools that were introduced in Araar (2006) and Araar et al. (2009).[26] More formally, we test the following null hypothesis for each $w \in U$, where U is the joint support of the two distributions:

(17) $$H_0: \Delta(\mathbf{F}(w)) = \mathbf{F_N}(w) - \mathbf{F_M}(w) < 0,$$

against the alternative hypothesis

(18) $$H_1: \Delta(\mathbf{F}(w)) = \mathbf{F_N}(w) - \mathbf{F_M}(w) \geq 0$$

and characterize any relevant range of w where we are able to reject the null.

Let $\hat{\sigma}(w)$ be the standard deviation of the estimator $\hat{\Delta}(\mathbf{F}(w))$, and let $z(\theta)$ be the $(1 - \theta)^{th}$ quantile of the standard normal distribution.[27] Davidson and Duclos (2000) show that the

[25] This can lead to difficulties in empirical work, and less restrictive concepts such as *restricted first order stochastic dominance* (Atkinson, 1987) and *almost stochastic dominance* (Leshno and Kevy, 2002) have been proposed.

[26] The calculations are implemented using the DASP Stata module presented in Araar and Duclos (2013).

[27] The asymptotic variance of $\hat{\Delta}(w)$ is derived in Araar et al. (2009).

estimator $\hat{\Delta}(\mathbf{F}(\mathrm{w}))$ is consistent and asymptotically normally distributed. We can then define the lower bound for a one-sided confidence interval for $\Delta\big(\boldsymbol{F}(w)\big)$ as:[28]

(19) $$\widehat{LB}_{\Delta(\mathbf{F}(\mathrm{w}))} = \hat{\Delta}(\mathbf{F}(\mathrm{w})) - \hat{\sigma}(w)z(\theta).$$

We estimate the standard errors using a Taylor linearization and allow for clustering at the individual level. We then implement the procedure by calculating the lower bounds of the confidence intervals for the estimate $\hat{\Delta}(\mathbf{F}(\mathrm{w}))$ defined in equation (19).

Table 4 reports the shares of migrants and non-migrants whose earnings are outside the range over which the migrant distribution stochastically dominates at a 95 percent confidence level. Consider first the distributions of non-migrant men and men migrating to destinations outside the Nordic countries. Although it is not clearly visible from figure 3a, the cumulative distribution functions cross near the lower tails of the distributions. Figure 5a depicts $\hat{\Delta}(\mathbf{F}(\mathrm{w}))$ and lower and upper bounds for a 95% confidence interval.[29] The lower bound of the confidence interval is positive on most of the range covering the supports of the distributions. Only 2.0 percent of the migrants and 3.4 percent of the non-migrants lie below the lower bound of the range where the lower bound of the confidence interval is positive, whereas the shares of migrants and non-migrants above the upper bound of the range are 0.1 and 0.0 percent. Put differently, the earnings of almost 98 percent of male migrants to destinations outside Nordic countries are on the range where the cumulative distribution function for non-migrants is statistically significantly above the function for migrants.

Figure 5b depicts $\hat{\Delta}(\mathbf{F}(\mathrm{w}))$ and the bounds for a 95% confidence interval for non-migrant women and women migrating to destinations outside Nordic countries. Only 2.8 percent of the migrants and 4.1 percent of the non-migrants have earnings below the range where the

[28] Chow (1989) proved the theorem for the case of independent samples. Davidson and Duclos (2000) show that the results also extend to the case of paired incomes from the same population.

[29] The upper bounds are calculated similarly to the lower bounds.

lower bound of the confidence interval is positive, and an even smaller 0.2 percent of the migrants and 0.0 percent of the non-migrants have earnings above this range. We interpret these findings as support for the stochastic dominance prediction for both men and women migrating outside Nordic countries.

Figures 6a and 6b and the bottom panel of Table 4 present a corresponding analysis by comparing the cumulative distributions of persons who migrate to other Nordic countries with that of non-migrants. Almost 12 percent of male migrants and 16 percent of male non-migrants have earnings that lie below the range where $\widehat{LB}_{\Delta(\mathbf{F}(w))}$ is positive, and another 1.5 percent of the migrants and 0.7 percent of the non-migrants have earnings above the range. Put differently, about 87 percent of the male migrants to Nordic countries have incomes on the range where $\widehat{LB}_{\Delta(\mathbf{F}(w))}$ is positive. For women, it can be seen in Table 4 that almost 95 percent of the migrants going to Nordic countries have earnings on the range where $\widehat{LB}_{\Delta(\mathbf{F}(w))}$ is positive. To sum up, the findings offer support to the stochastic dominance prediction for male and female migrants regardless of their destination, although the evidence is weaker for men who migrated to Nordic countries.

Additional support for our theory comes from Mexico. Our theory predicts that the earnings distribution of migrants from Mexico to the United States should be stochastically dominated by the earnings distribution of non-migrants. Fernández-Huertas Moraga (2011) presents these distributions for men. Although he does not present confidence intervals as we do, the figures suggest a pattern that mirrors what we find for Denmark, reversing the curves for migrants and non-migrants. In Mexico, the wage distribution of non-migrants stochastically dominates that of migrants, apart from an overlap for a few percent at the bottom and converging at the top.

5. Selection in unobserved characteristics

In the previous section, we documented the selection that characterizes the migrants using the total pre-migration earnings (after adjusting for age and year). We now examine a

specific component of earnings, namely the component due to unobserved characteristics. In particular, we now adjust for differences in educational attainment between migrants and non-migrants (as well as other observable variables) by running earnings regressions, and determine whether the distribution of the residuals differs between the two groups.[30]

By construction, the residuals from a Mincerian wage regression reflect the part of earnings that is uncorrelated with the observed measures of skill. Obviously, the decomposition is somewhat arbitrary because it depends on the characteristics that are observed and can be included as regressors in the wage equation. Nevertheless, the study of emigrant selection in terms of wage residuals is important for a number of reasons.

First, selection in terms of unobservable characteristics sheds light on the importance of the quality of job matches relative to the skill component that is internationally transferable. The theory predicts that the nature of the selection in unobservable characteristics depends on the magnitude of the correlation coefficient measuring how the source and destination countries value these types of skills. As long as this correlation is strongly positive (so that unobserved characteristics are easily transferable across countries), Danish emigrants would be positively selected in unobservables. After all, the payoff to these types of skills is likely to be greater in the destination countries. However, it could be argued that the correlation between the wage residuals in Denmark and abroad may be "small". For example, the residuals from the wage regression may be largely reflecting the quality of the existing job match in the Danish labor market, rather than measuring the worker's innate productivity. To the extent that the quality of the job match plays an important role in generating the residual, the correlation in this residual across countries would be expected to be weak (in fact, a pure random matching model would suggest that it would be zero). As a result, there would be negative selection in unobserved characteristics simply because Danish workers with good job matches (and hence high values of the residual) would not move.

[30] In the earnings regressions we use non-standardized annual earnings as the dependent variable. We include age and year fixed effects and run the regressions separately for men and women.

Second, the theory also suggests that the nature of the selection in pre-migration earnings depends on a weighted average of the selection that occurs in observable and unobservable characteristics, with the weights being the fraction of earnings variance attributable to each type of skill. Because observable characteristics play only a limited role in explaining the variance of earnings in the population, it is crucial to precisely delineate the nature of selection in unobservable characteristics.

Table 5 reports the Mincerian wage regressions that we use to calculate the residuals. The sample includes the whole population of prime aged full time workers pooled over the entire 1995-2004 period. In addition to vectors of fixed effects giving the worker's age and educational attainment, we also include the worker's marital status and number of children. The regressions are estimated separately for men and women.

Figure 7a presents the cumulative distributions of wage residuals for male migrants to Nordic countries, male migrants to destinations outside Nordic countries, and male non-migrants. The values of the residuals are truncated at -2 and 2, a range that covers practically all of the population.[31] The cumulative distribution function of residuals for emigrants who moved outside the Nordic countries is located to the right of the cumulative distribution for migrants to Nordic countries, which in turn is located to the right of the cumulative distribution of the non-migrants. The visual evidence, therefore, provides a strong indication that migrants are positively selected in terms of unobserved characteristics. Figure 7b presents the analogous evidence for women.[32] The figure shows that female migrants are also positively selected in terms of wage residuals. As was the case when comparing the measure of pre-migration earnings in the previous section, the

[31] For men, 0.05% of non-migrants, 0.19% of migrants to other Nordic countries and 0.11% of migrants to other destinations lie below the lower truncation point. Correspondingly, 0.03% of non-migrants and 0.23% of migrants to destinations outside Nordic countries lie above the upper truncation point. There are no migrants to other Nordic countries above the upper truncation point.

[32] For women, 0.05% of non-migrants lie below the lower truncation point and 0.00% of non-migrants lie above the higher truncation point. There are no migrants lying below the lower or above the higher truncation point.

selection in unobserved characteristics is less pronounced for women than for men. One explanation for this could be that men are typically primary earners in couples.

We also performed Kolmogorov-Smirnov tests on the distributions of residuals for non-migrants and migrants to other Nordic countries and for migrants to other destinations (separately for men and women). All the tests clearly rejected the null hypothesis, confirming that the distributions of residuals indeed differ among the groups.[33]

The evidence on the positive selection of migrants in unobserved characteristics obviously implies that the selection in pre-migration earnings documented in the previous section cannot be attributed solely to the fact that migrants are more educated. Instead, we find that there is positive selection *within* education groups. This result also has implications on the interpretation of earnings regression residuals in general. The residuals from wage regressions are sometimes interpreted as reflecting the value of the job match between the worker and the employer. If a high value for the residual only reflects a good match, we would then expect to find that workers with large residuals would be less likely to change jobs and less prone to migrate. Our findings clearly reject this interpretation. Comparing results on the self-selection to other Nordic countries and the rest of the world suggests that search for a better job match to those who have a bad job match in Denmark is more pronounced among migrants to other Nordic countries.[34]

As in the previous section, we also calculated the difference between the cumulative distribution functions with confidence intervals to determine whether empirical evidence supports the stochastic dominance prediction. The test results are summarized in Table 6. Figure 8a depicts $\hat{\Delta}(\mathbf{F}(\mathrm{w}))$ and the lower and upper bounds for a 95% confidence interval

[33] The *p*-value for the test between women migrating to other Nordic countries and to other destinations was 0.015 and all the other *p*-values were 0.000, so that all tests clearly reject the hypothesis that the observations are drawn from the same distribution.

[34] For this group, returns to unobserved productivity are not as important a criterion for self-selection as among migrants to the rest of the world, simply because differences in returns to skills between Denmark and other Nordic countries are minor. As a result, the mechanism of searching for a better match quality is more pronounced.

for the comparison between non-migrant men and men migrating to destinations outside Nordic countries. The lower bound of the 95% confidence interval is positive on the range of residuals covering most of the support of the two distributions. 9.9 percent of the migrants and 15.2 percent of the non-migrants have wage residuals below the lower bound of this range, whereas the shares of migrants and non-migrants above the upper bound of the range are 0.1 and 0.0 percent.[35]

Figure 9a depicts $\hat{\Delta}(\mathbf{F}(w))$ and the bounds for a 95% confidence interval for non-migrant men and men migrating to other Nordic countries. A 13 percent share of migrants and 15 percent of non-migrants have values of the wage residual that are below the lower bound of the range where the lower bound of the 95% confidence interval is positive, and shares of 2 percent and 1 percent of migrants and non-migrants have values of the residual that would place them above this range. Put differently, the residuals of over 84 percent of male migrants to destinations outside Nordic countries are on the range where the cumulative distribution function for non-migrants is statistically significantly above the function for migrants.[36] Interestingly, there is a sizable area in the left tail of the distribution of residuals where the upper bound of the confidence interval is negative.[37]

We conclude by summarizing the evidence as follows: there is strong positive selection in unobservable characteristics in the sample of migrants that moved outside the Nordic countries and weaker evidence of positive selection in the sample of migrants who moved to other Nordic countries.

---

[35] For women, 20 percent of migrants and 25 percent of non-migrants have earnings residuals below the lower bound of the range where the lower bound of the confidence interval is positive, and shares of migrants and non-migrants above the range are less than one percent.

[36] For women, 20 percent of migrants and 25 percent of non-migrants have earnings residuals below the lower bound of the range where the lower bound of the confidence interval is positive, and shares of migrants and non-migrants above the range are 3 percent and 2 percent.

[37] A 2 percent share of migrants and 2 percent share of non-migrants have residuals in this area, and the interpretation would be that male migrants to other Nordic countries are negatively selected in terms of residuals in the left tail of the distribution.

6. Bias in counterfactual predictions

The fact that emigrants are self-selected in their unobserved characteristics implies that using the observable characteristics of migrants to predict their counterfactual earnings had they chosen not to migrate will lead to biased results. Due to data constraints, this is precisely the empirical exercise conducted by Chiquiar and Hanson (2005), who adopt the methodology introduced by DiNardo, Fortin, and Lemieux (1996) and build a counterfactual wage density of what the Mexican immigrants would have earned in Mexico had they stayed. The actual wage density of Mexican "stayers" is then compared to the counterfactual density for migrants. By construction, this approach ignores the role of unobservable characteristics in the estimation of the counterfactual wage distribution.

A clear advantage of the Danish administrative data is that the earnings of emigrants can be observed before they emigrate, so there is no need to build a counterfactual density. One just needs to compare the earnings distribution of non-migrants to the actual distribution of future migrants, as we have done in the preceding analysis. The administrative data, however, allows us to precisely measure the extent of the bias resulting from carrying out the counterfactual exercise in Chiquiar and Hanson (2005). In particular, we can contrast the predicted counterfactual wage distribution of migrants had they not moved to the actual wage distribution of migrants prior to their move. We carry out this exercise by precisely replicating the various steps in the Chiquiar-Hanson calculations. It is worth emphasizing that this type of bias will arise not only in studies that examine the selection of migrants, but in *any* study that relies on observables to predict a counterfactual wage distribution.

Let w represent the logarithm of standardized annual earnings as defined earlier (i.e. earnings adjusted for age, gender, and year effects). Let $f(w|x)$ be the density function of wages in Denmark, conditional on a set of observable characteristics x. Also, let I be an indicator variable equal to one if the individual migrates the following year and equal to zero otherwise. Define further $h(x|I = 0)$ as the conditional density of observed

characteristics among workers in Denmark who choose not to migrate, and $h(x|I=1)$ be the corresponding conditional density among migrants. The observed wage density for the non-migrants is

(20) $$g(w|I=0) = \int f(w|x, I=0)h(x|I=0)\,dx.$$

Similarly, the observed density for the migrants is

(21) $$g(w|I=1) = \int f(w|x, I=1)h(x|I=1)\,dx.$$

Up to this point, the analysis reported in this paper consists of directly estimating and comparing the distribution functions associated with the densities in equations (20) and (21). Suppose that the pre-migration earnings density for non-migrants were not available. We would instead attempt to estimate it from the observable characteristics of the migrants. The implied counterfactual distribution is:

(23) $$\hat{g}(w|I=1) = \int f(w|x, I=0)h(x|I=1)\,dx.$$

Equation (23) corresponds to the density of income for non-migrants, but it is instead integrated over the density of observable characteristics for migrants. Note that the counterfactual density in (23) can be rewritten as:

(24)

$$\hat{g}(w|I=1) = \int f(w|x, I=0)h(x|I=0)\frac{h(x|I=1)}{h(x|I=0)}dx$$

$$= \int \theta f(w|x, I=0)h(x|I=0)\,dx,$$

where $\theta = \frac{h(x|I=1)}{h(x|I=0)}$. To compute θ, we use Bayes' law to write:

(25) $$h(x) = \frac{h(x|I=0)Pr(I=0)}{\Pr(I=0|x)} \text{ and } h(x) = \frac{h(x|I=1)Pr(I=1)}{\Pr(I=1|x)},$$

where $h(x)$ is the unconditional density of observed characteristics.

We can then combine these two equations to solve for θ:

(26) $$\theta = \frac{Pr(I=1|x)}{1-Pr(I=1|x)} \frac{Pr(I=0)}{Pr(I=1)}.$$

The proportion $Pr(I = 0)/Pr(I = 1)$ is a constant related to the proportion of migrants in the data. It can be set to one in kernel density estimation without loss of generality. The weight we use in the estimation is then given by:

(27) $$\theta^e = \frac{Pr(I=1|x)}{1-Pr(I=1|x)}.$$

As in Chiquiar and Hanson (2005), the individual weights θ^e are calculated by estimating a logit model where the dependent variable indicates if a person emigrated. The regressors include a vector of age fixed effects, a vector of schooling fixed effects, variables indicating whether the worker is married and the number of children (and an interaction between these two variables), and a vector of year fixed effects.[38] Table 7 reports the logit regressions estimated separately by gender. The coefficients are then used to compute the weights for each non-migrant person in the sample.[39] Figures 10a and 10b present the

[38] We also tried specifications with age, age squared and interactions of explanatory variables, but we do not report these analyses as the resulting counterfactual distributions did not practically differ from the distributions resulting from this simpler specification.

[39] As earlier, we use Silverman's reference bandwidth multiplied by 0.75.

resulting counterfactual density functions of the logarithm of standardized earnings as well as the actual distributions for migrants and non-migrants.[40]

The difference between the actual density for non-migrants and the counterfactual density for migrants reflects the part of self-selection that is due to observable characteristics. Similarly, the difference between the counterfactual and actual densities for migrants reflects the part of selection that is due to unobserved characteristics (i.e., all those variables that could not be included in the logit model).

One simple way of quantifying these distributional differences is to compute the averages of the various distributions. These calculations are reported in Table 8. Consider initially the results in the male sample. The difference between the mean of the actual distributions for migrants and non-migrants is 0.245 log points, but the difference between the counterfactual distribution and the distribution for non-migrants is 0.073. This implies that only about 30 percent of the positive selection in pre-migration earnings can be attributed to the observable characteristics included in the logit model, while about 70 percent is attributable to unobservable determinants of productivity.

The calculations in the female sample yield a difference of 0.157 log points between the means of the actual distributions for migrants and non-migrants and a difference of 0.074 points between the counterfactual distribution and the distribution for non-migrants. As a result, observable and unobservable characteristics each account for about half of the positive self-selection in the pre-migration earnings of women.[41] The key lesson is clear: selection in unobservable characteristics plays a crucial role in determining the skill composition of emigrants.

---

[40] To conduct the counterfactual analysis we pool the sample of all migrants (regardless of whether they moved to Nordic countries or not).

[41] The component of self-selection that is due to unobservable characteristics plays a somewhat smaller role for women. One reason could be that women are more often tied migrants, and the migration decision may be mainly based on the skills of the spouse. The variance in income is also smaller for women, which also makes the selection both in terms of observable and unobservable characteristics weaker.

The distinct role of observables and unobservables in determining the selection in the pre-migration earnings of migrants is evident if we return to the Roy model and equation (16), which presents the conditional expectation $E(\log w_0 \mid v^* > -\Delta\mu^*)$.

Equation (16) yields an interesting and potentially important insight. The nature of the selection in pre-migration earnings, of course, is given by the sum of the selection in observables and the selection in unobservables. Note, however, that each of these selection terms has a weighting coefficient that represents the variance in earnings attributable to observable characteristics ($r_0^2\sigma_S^2$) or to unobservable characteristics (σ_0^2). As noted earlier, observable characteristics explain a relatively small fraction of the variance in earnings. Put differently, equation (16) implies that it is the selection in *unobservables* that is most likely to determine the nature of the selection that characterizes the emigrant sample.

To the extent that both types of selections (i.e., in observables and unobservables) work in the same direction, the counterfactual exercise described in this section will inevitably underestimate the true extent of positive selection in pre-migration earnings. Conversely, the counterfactual exercise will also attenuate the extent of "true" negative selection if there is negative selection in both components of skills. In fact, Fernández-Huertas Moraga (2011) presents a corresponding analysis using survey data from Mexico and finds that counterfactual estimates greatly underestimate the extent of negative selection in the pre-migration earnings of Mexicans who move to the United States. Put differently, the counterfactual exercise may lead to qualitatively right conclusions about the nature of the selection, but it may also generate a sizable bias, greatly underestimating the true extent of either positive or negative selection.

7. Selection and immigration restrictions

As applied in the immigration literature, the Roy model focuses solely on the economic factors that motivate labor flows across international borders. The modeling typically ignores the fact that these flows occur within a policy framework where some receiving countries enact detailed restrictions specifying which potential migrants are admissible and which are not.

We can use the administrative data from Denmark, combined with the unique political circumstances that guarantee free migration within Europe, to partially address the question of whether immigration policy affects selection all that much in the end. Specifically, we can subdivide the group of migrants who moved outside Nordic countries into two groups: those who moved to a country in the EU15 or to Switzerland, and those who moved to a country outside the EU15 and Switzerland. Movement of labor was unrestricted between Denmark and other EU15 countries and Switzerland in the period under study, but was obviously restricted by immigration regulations to destinations outside the EU15, such as the United States.

It turns out that these different immigration policies pursued by the EU15 and Switzerland and the rest of the world barely matter in determining the selection of Danish emigrants. Figure 11a depicts the cumulative distribution functions of the logarithm of standardized annual income for men and figure 11b for women. It is evident that the distribution functions of standardized earnings are very similar for the two groups of migrants.[42] We also conducted the analysis using the wage residuals (not shown), and the distributions of residuals are also similar between the two groups.

There is an important sense in which these policy restrictions cannot matter much. Suppose, for example, that a receiving country enacts a policy that limits entry only to high-skill immigrants. If the high-skill immigrants from a sending country do not find it optimal to move, the policy cannot force those high-skill workers to migrate. All the policy can do is

[42] For women, a Kolmogorov-Smirnov test is not able to reject the null-hypothesis that the observations for the two groups of migrants come from the same underlying distribution.

essentially cut the migration flow from that particular sending country down to zero. The low-skill workers would like to move but are not admitted, and the high skill workers are admissible but they do not want to move.

In sum, the positive self-selection that is so evident in the Danish emigrant data cannot be explained by immigration restrictions. Even though labor flows to the EU15 and Switzerland were unrestricted, there is no evidence of weaker positive selection to these countries than to the rest of the world. Our results also have tentative implications for the question of whether migration patterns reflect differences in rate of returns or migration costs. Although it is plausible that migration costs are higher when moving to other continents, our results suggest that such differences do not play a significant role in the sorting of emigrants between Europe and other continents, most notably North America.

8. Conclusion

This paper shows that the Roy model has more dramatic predictions on the self-selection of emigrants than previously examined. The same conditions that have been shown to result in emigrants being positively (negatively) self-selected in terms of their average earnings actually imply that the earnings distribution of emigrants first-order stochastically dominates (or is first-order stochastically dominated by) the earnings distribution of non-migrants. Our theoretical analysis also distinguishes between selection in observable and selection in unobservable characteristics.

Our empirical analysis uses the Danish full population administrative data to analyze the self-selection of emigrants, in terms of education, earnings and unobservable ability, measured by residuals from Mincerian earnings regressions. The results are in line with the theory; the migrants are better educated and both pre-emigration earnings and wage regression residuals of migrants stochastically dominate those of non-migrants over most of the support of the distributions. Consider, for example, the case of full-time workers aged 25-54. For 98 percent of men and 97 percent of women who migrate outside other Nordic countries the cumulative earnings distribution in the year before emigration

stochastically dominates that of non-migrants with a 95% confidence interval. The difference between the cumulative distributions is not statistically significantly different in either direction for the remaining 2 to 3 percent.

Decomposing the self-selection in total earnings into self-selection in observable characteristics and self-selection in unobservable characteristics (as measured by residuals from Mincerian wage regressions), reveals that unobserved abilities play the dominant role. For men, about 70 percent of the positive self-selection in pre-migration earnings is attributable to unobservable determinants of productivity. For women, the fraction is about 50 percent. This suggests that relying on counterfactual distributions, based on observed characteristics, would strongly underestimate positive self-selection. This result complements the Fernández-Huertas Moraga (2011) finding that counterfactual estimates also greatly underestimate the extent of negative selection in the pre-migration earnings of Mexicans who move to the United States. In short, the use of counterfactual earnings distributions based on observable characteristics greatly understate the true extent of selection in total earnings. Strong positive self-selection in residuals also suggests that unobserved abilities play a much bigger role in migration decisions than match quality.

Our findings also have implications for immigration policies. Receiving countries can only base their admission policies on skill variables that are observed, whereas much of the selection of immigrants is "hidden" in their unobserved characteristics. It can be expected that migrants will be self-selected in terms of unobserved characteristics even when admission restrictions are applied, and the self-selection among those fulfilling admission criteria can be expected to reflect relative skill prices. This raises a question about the effectiveness of point systems that are necessarily based on observable characteristics. The importance of relative skill prices is also supported by our separate analyses of self-selection of Danes migrating to the countries belonging to common European labor market (excluding other Nordic countries that have skill prices similar to Denmark) and not having any immigration restrictions, and the self-selection to the rest of the world. There is virtually no difference in the self-selection to these destination areas.

References

Abramitzky, R., L. Boustan, and K. Eriksson (2012). "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102(5): 1832-56.

Abramitzky, R. and F. Braggion (2006). "Migration and Human Capital: Self-Selection of Indentured Servants to the Americas." *Journal of Economic History* 66 (4): 882–905.

Araar, A. (2006). "Poverty, Inequality and Stochastic Dominance, Theory and Practice: The Case of Burkina Faso." Cahiers de recherché PMMA 2007-087, PEP-PMMA.

Araar, A. and J.Y. Duclos (2013). "User Manual for Stata Package DASP: Version 2.3." Université Laval PEP, CIRPÉE and World Bank.

Araar, A., J.Y. Duclos, M. Audet, and P. Makdissi (2009). "Testing for Pro-poorness of Growth, with an Application to Mexico." *Review of Income and Wealth* 55 (4): 853-881.

Arnold, B. C., R. Beaver, R. A. Groeneveld and W. Q. Meeker (1993). "The Nontruncated Marginal of a Truncated Bivariate Normal Distribution." *Psychometrika* 58 (3): 471-488.

Atkinson, A. B. (1987). "On the Measurement of Poverty." *Econometrica* 55: 749-764.

Boarini, R. and H. Strauss (2010). "What is the Private Return to Tertiary Education?: New Evidence from 21 OECD Countries." *OECD Journal: Economic Studies*, Vol. 2010/1.

Borjas, G.J. (1987). "Self-Selection and the Earnings of Immigrants." *American Economic Review* 77: 531-553.

Borjas, G. J. (2003). "The Labor Demand Curve is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market." *Quarterly Journal of Economics* 118(4): 1335–1374.

Bratsberg, B. (1995). "The Incidence of Non Return Among Foreign Students in the United States." *Economics of Education Review* 14(4): 373-384.

Chiquiar, D. and G.H. Hanson (2005). "International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the United States." *Journal of Political Economy* 113(2): 239-281.

Chow, K. V. (1989). "Statistical Inference for Stochastic Dominance: a Distribution Free Approach." Ph.D. Thesis, University of Alabama.

Cobb-Clark, D. (1993). "Immigrant Selectivity and Wages: The Evidence for Women." *American Economic Review* 83: 986-93.

Davidson R. and J.Y. Duclos (2000). "Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality." *Econometrica* 68: 1435-64.

Davidson R. and J.Y. Duclos (2013). "Testing for Restricted Stochastic Dominance." *Econometric Reviews* 32: 84-125.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach." *Econometrica* 64 (September): 1001–44.

Fernández-Huertas Moraga, J. (2011). ''New Evidence on Emigrant Selection.'' *Review of Economics and Statistics* 93(1): 72–96.

Ferrie, J. (1996). "A New Sample of Males Linked from the Public Use Micro Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules." *Historical Methods* 29: 141–56.

Grogger, J. and G. H. Hanson (2011). "Income Maximization and the Selection and Sorting of International Migrants." *Journal of Development Economics* 95(1): 42-57.

Hanushek E. A., G. Schwerdt, S. Wiederhold, and L. Woessmann (2015). "Returns to Skills around the World: Evidence from PIAAC." *European Economic Review* 73: 103-130.

Junge, M., M.D. Munk, and P. Poutvaara (2014). "International Migration of Couples." CESifo WP 4927.

Kaestner, R. and O. Malamud (2014). "Self-Selection and International Migration: New Evidence from Mexico," *The Review of Economics and Statistics,* 96(1): 78-71.

Kleven, H.J., C. Landais, E. Saez, and E. Schultz (2014). "Migration and Wage Effects of Taxing Top Earners: Evidence of the Foreigners' Tax Scheme in Denmark." *Quarterly Journal of Economics* 129: 333–78.

Leibbrandt, M., J. Levinsohn, and J. McCrary (2005). "Incomes in South Africa since the Fall of Apartheid." NBER working paper no. 11384.

Leshno, M. and H. Kevy (2002). "Preferred by All and Preferred by Most Decision Makers: Almost Stochastic Dominance." *Management Science* 48: 1074-1085.

Lundborg, P. (1991). "Determinants of Migration in the Nordic Labor Market." *The Scandinavian Journal of Economics* 93(3): 363-375.

Margo, R. A. (1990). *Race and Schooling in the South, 1880–1950: An Economic History*. Chicago: University of Chicago Press.

Pirttilä, J. (2004). "Is International Labour Mobility a Threat to the Welfare State? Evidence from Finland in the 1990's." *Finnish Economic Papers* 17(1): 18-34.

Sinn, H.-W. (1997). "The Selection Principle and Market Failure in Systems Competition." *Journal of Public Economics* 66: 247-274.

Thistle, P.D. (1993). "Negative Moments, Risk Aversion and Stochastic Dominance." *Journal of Financial and Quantitative Analysis* 28(2): 301-311.

United Nations, Department of Economic and Social Affairs (2013). Trends in International Migrant Stock: Migrants by Destination and Origin (United Nations database, POP/DB/MIG/Stock/Rev.2013).

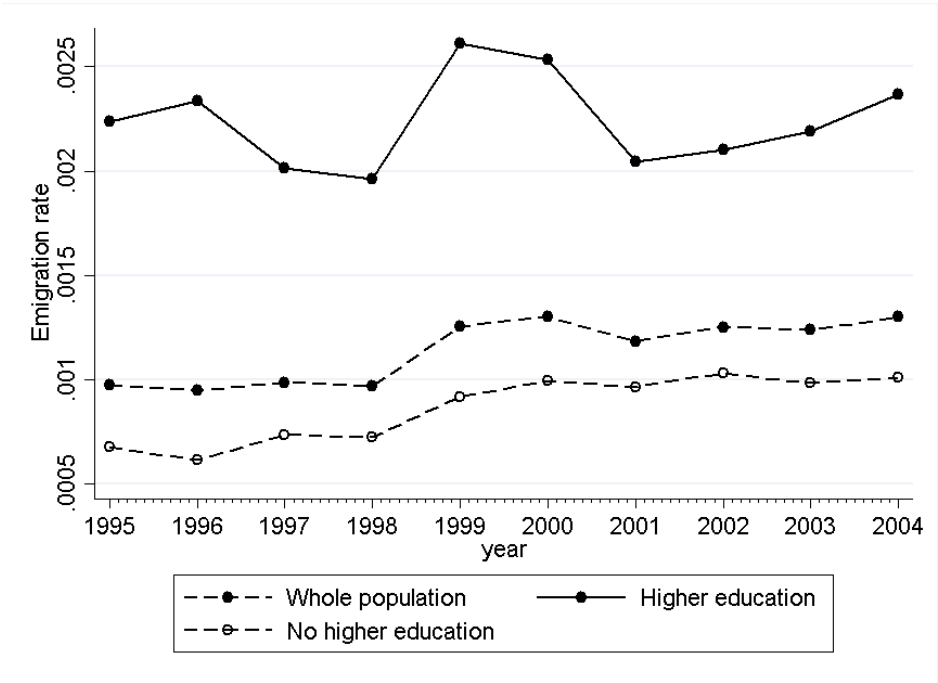
Wegge, S.A. (1999). "To Part or Not to Part: Emigration and Inheritance Institutions in Nineteenth-Century Hesse-Cassel." *Explorations in Economic History* 36 (1): 30–55.

Wegge, S.A. (2002). "Occupational Self-Selection of European Emigrants: Evidence from Nineteenth-Century Hesse-Cassel." *European Review of Economic History* 6 (3): 365–94.

Wildasin, D. E. (1991). "Income Redistribution in a Common Labor Market." *American Economic Review* 81 (4): 757-774.

Figure 1. Evolution of the emigration rate

a. Men



b. Women

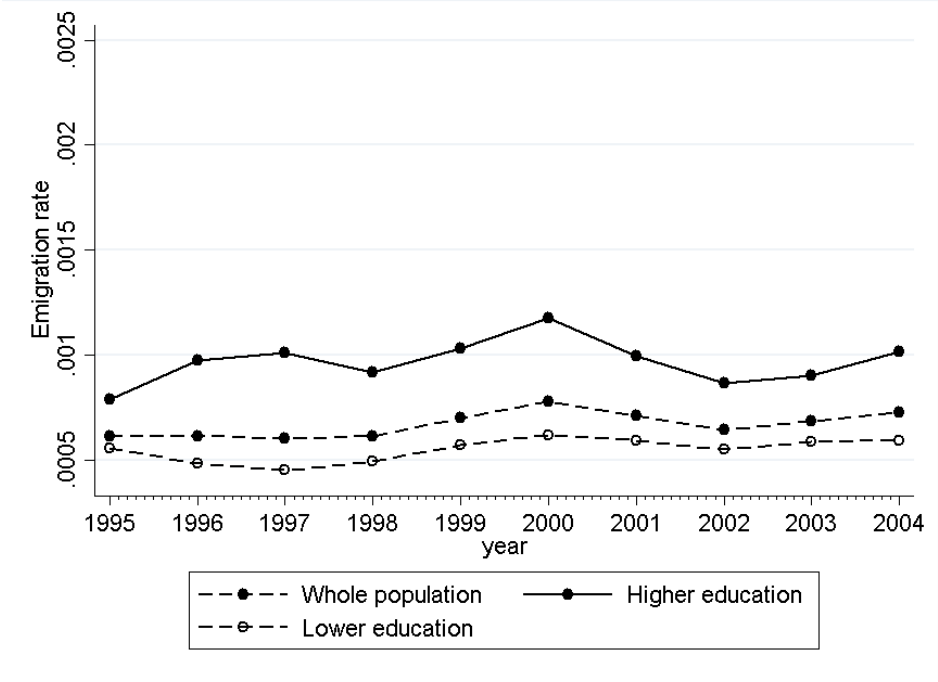
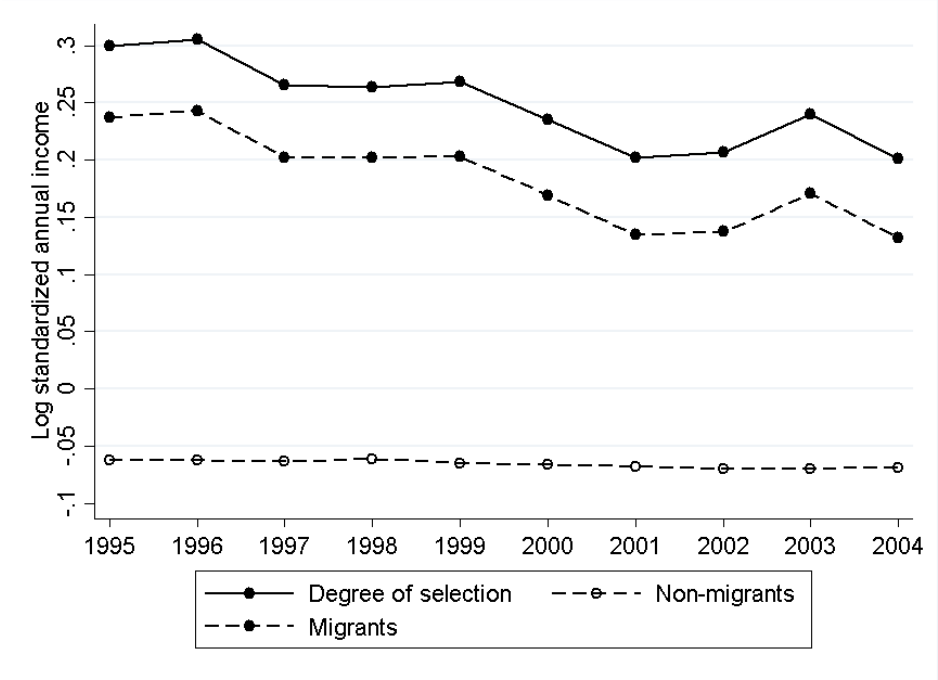


Figure 2. Evolution of the difference between average log standardized earnings of migrants and non-migrants

a. Men



b. Women

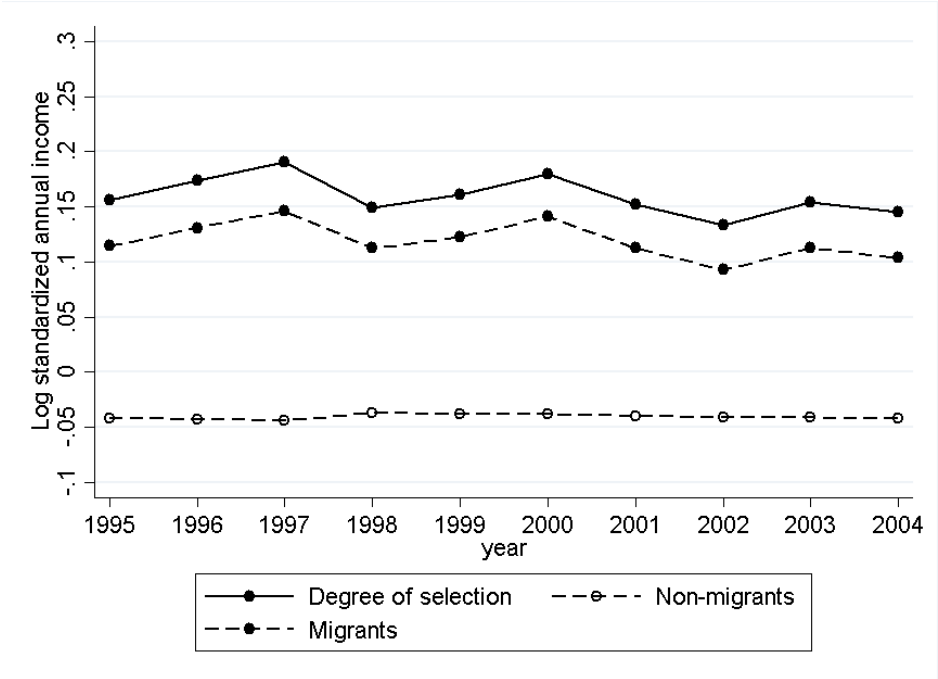
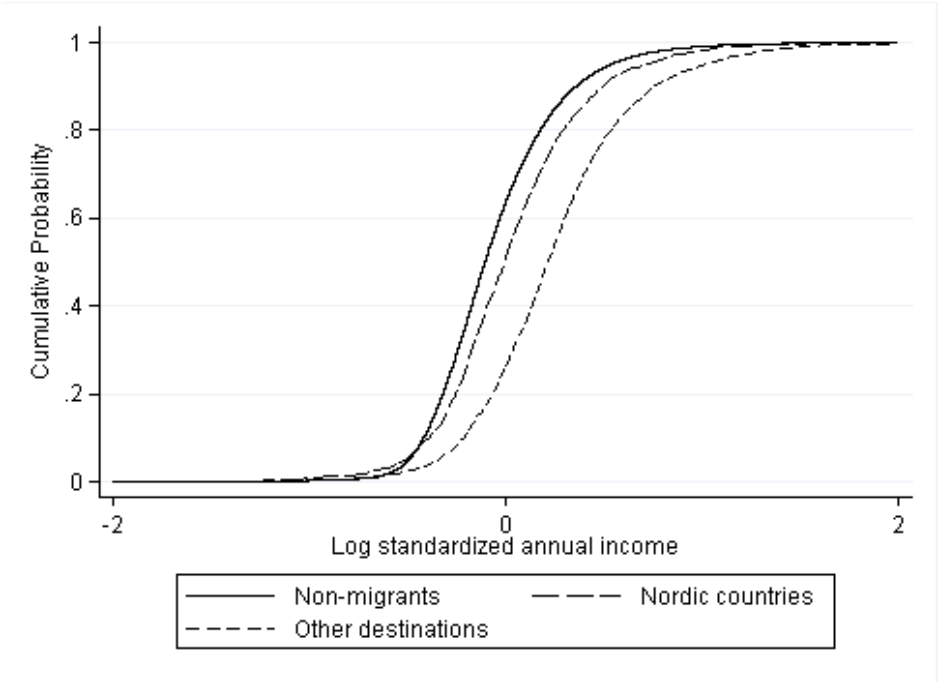


Figure 3. Distribution functions of standardized annual earnings for migrants and non-migrants

a. Men



b. Women

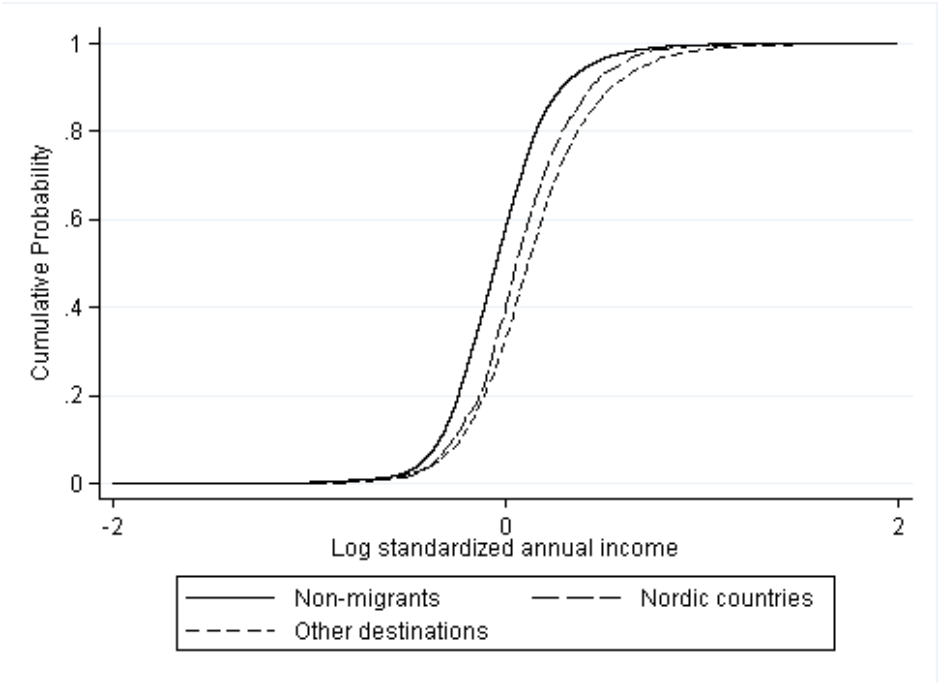
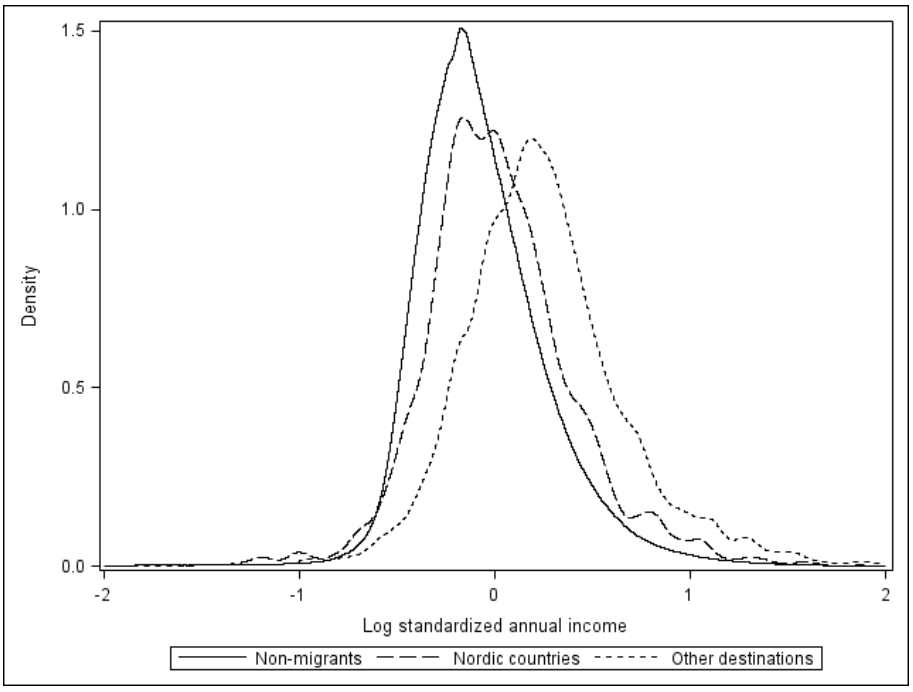


Figure 4. Density functions for standardized earnings for migrants and non-migrants

a. Men



b. Women

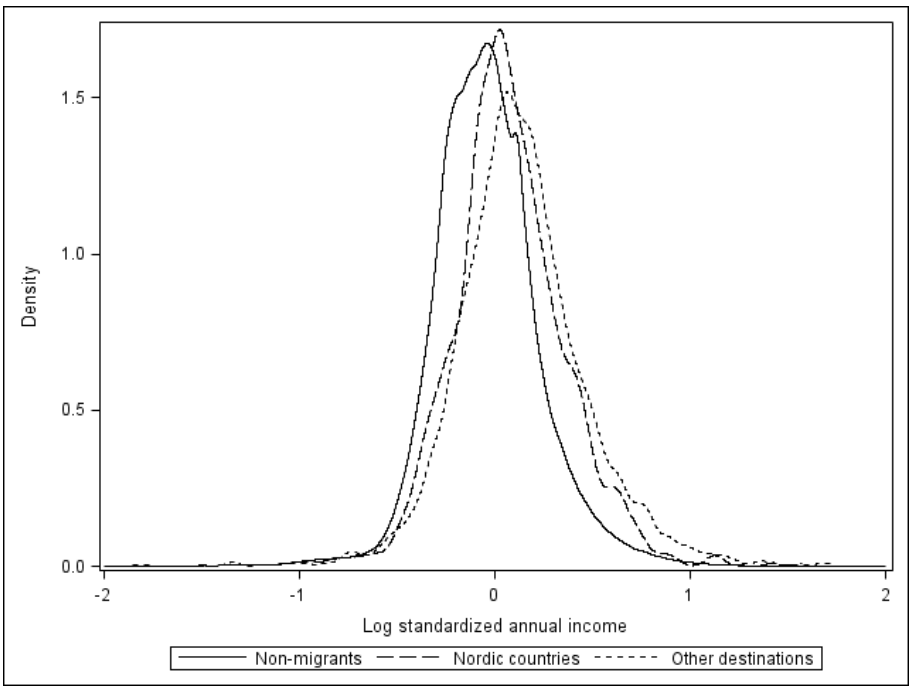
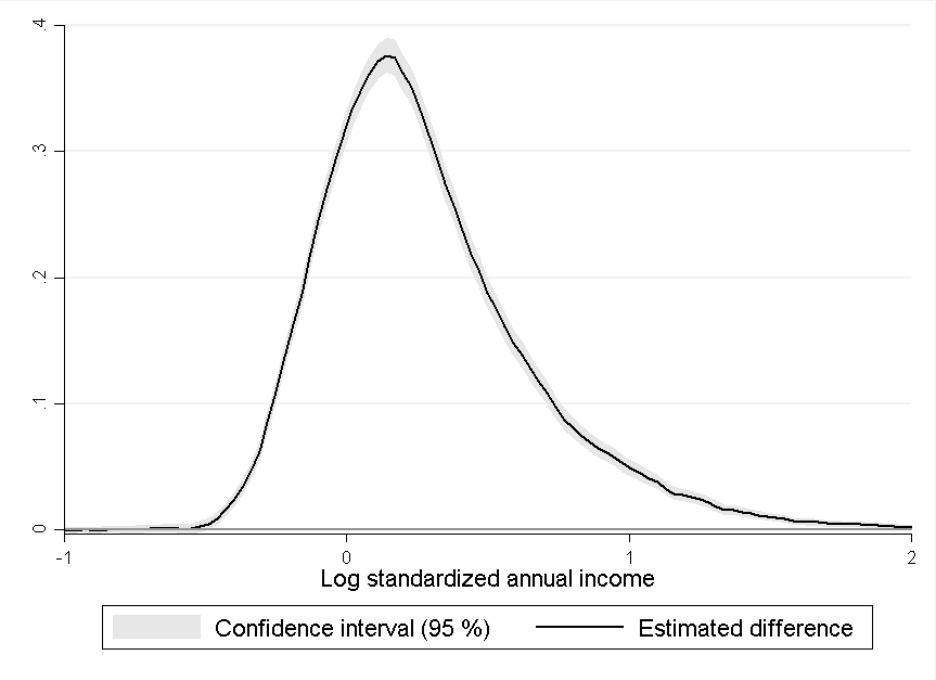


Figure 5. Difference of the cumulative distribution functions for pre-migration earnings between migrants moving outside Nordic countries and non-migrants

a. Men



b. Women

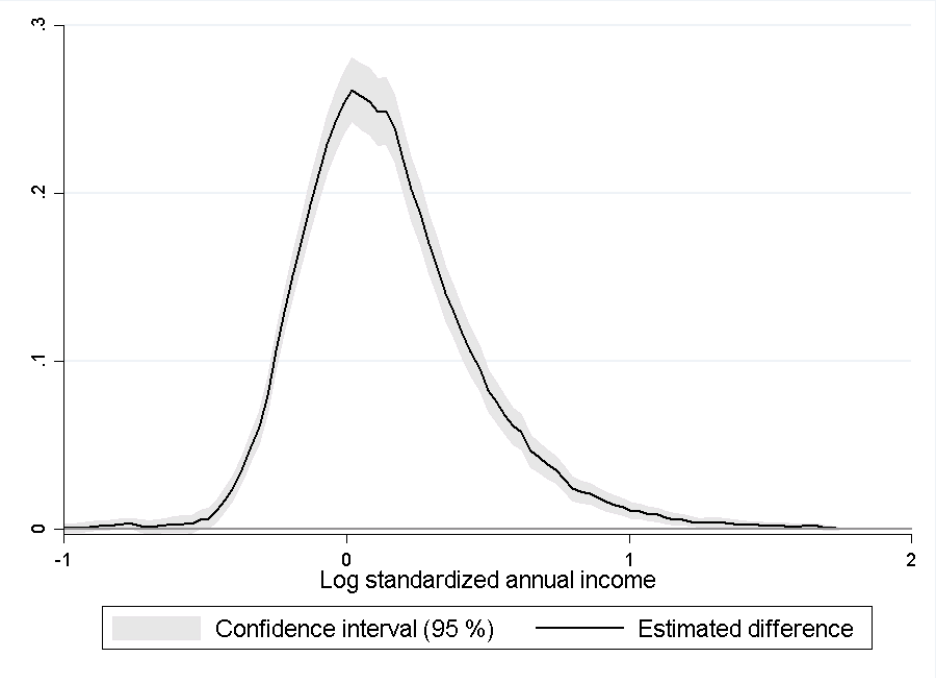
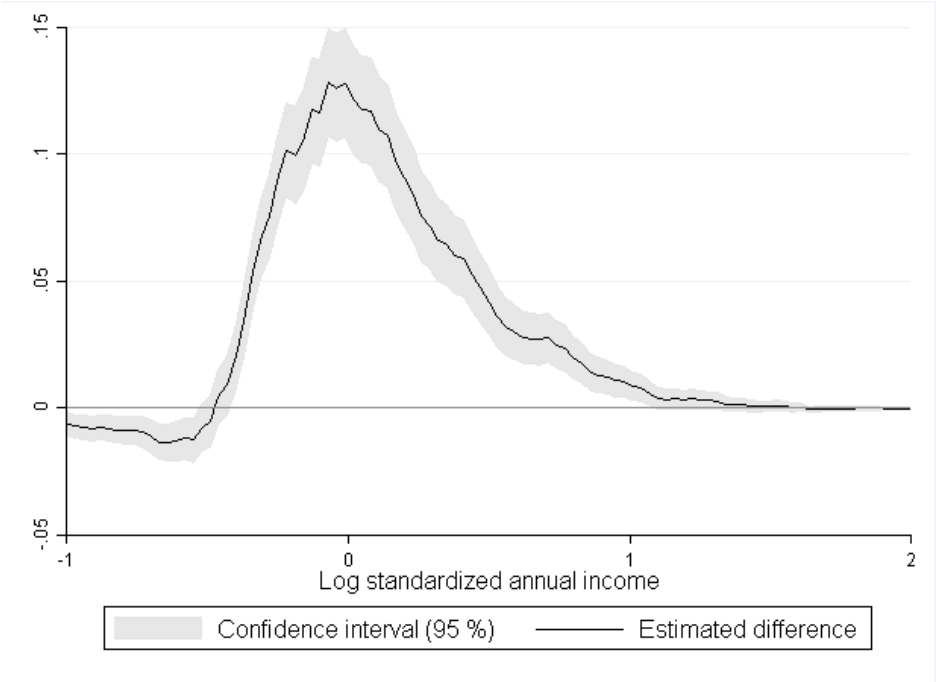


Figure 6. Difference of the cumulative distribution functions for pre-migration earnings of migrants going to other Nordic countries and non-migrants

a. Men



b. Women

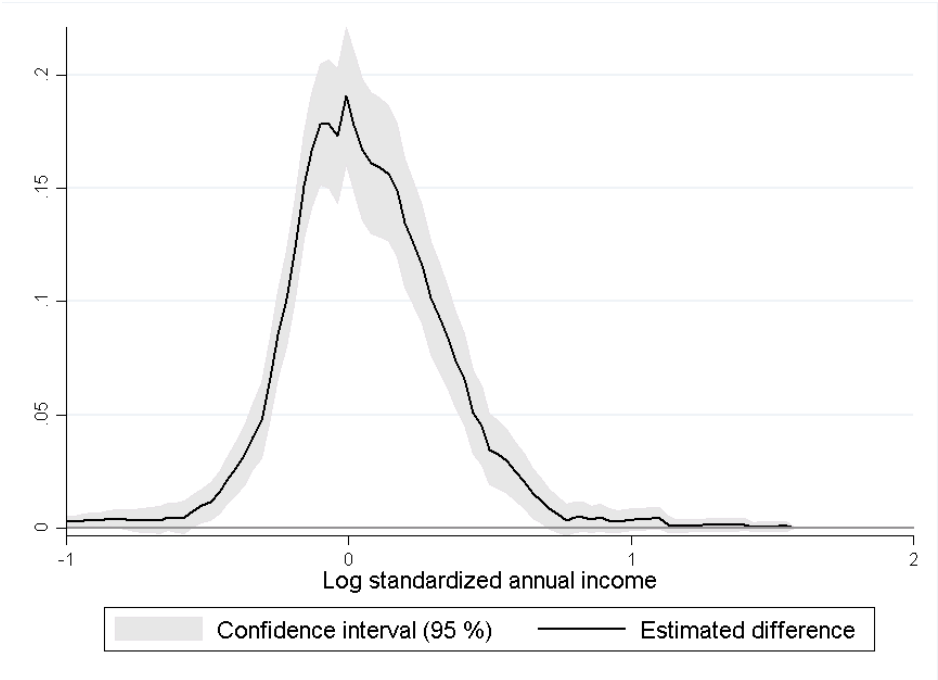
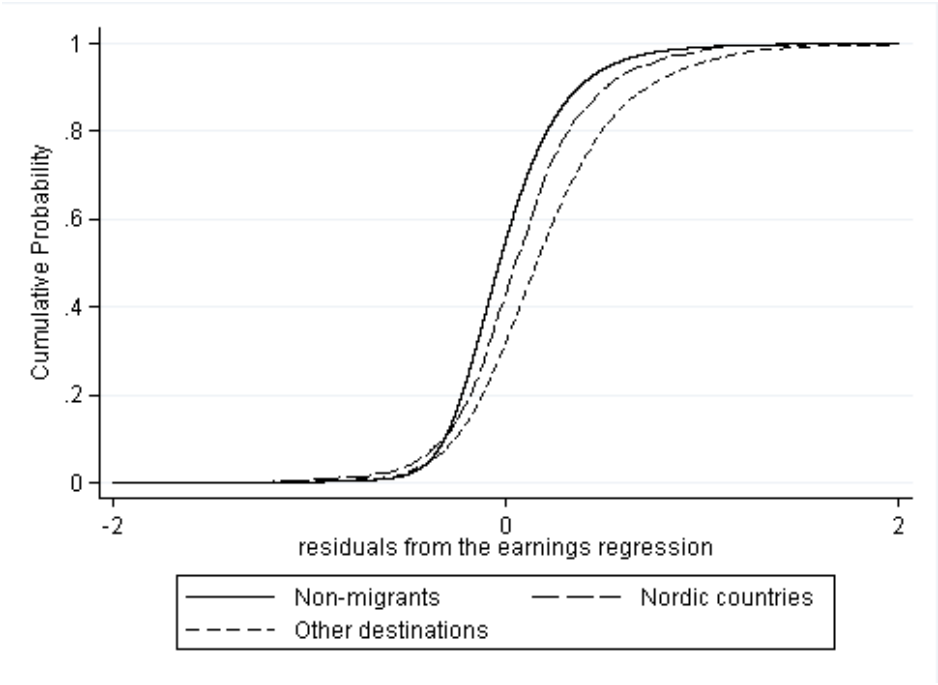


Figure 7. Distribution functions of residuals from earnings regression for migrants and non-migrants

a. Men



b. Women

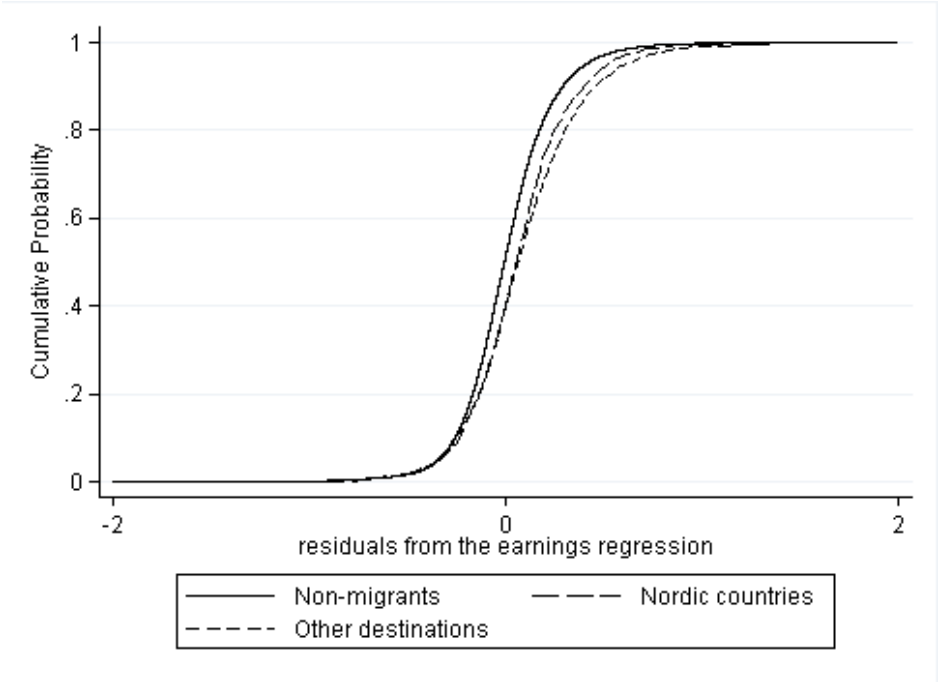
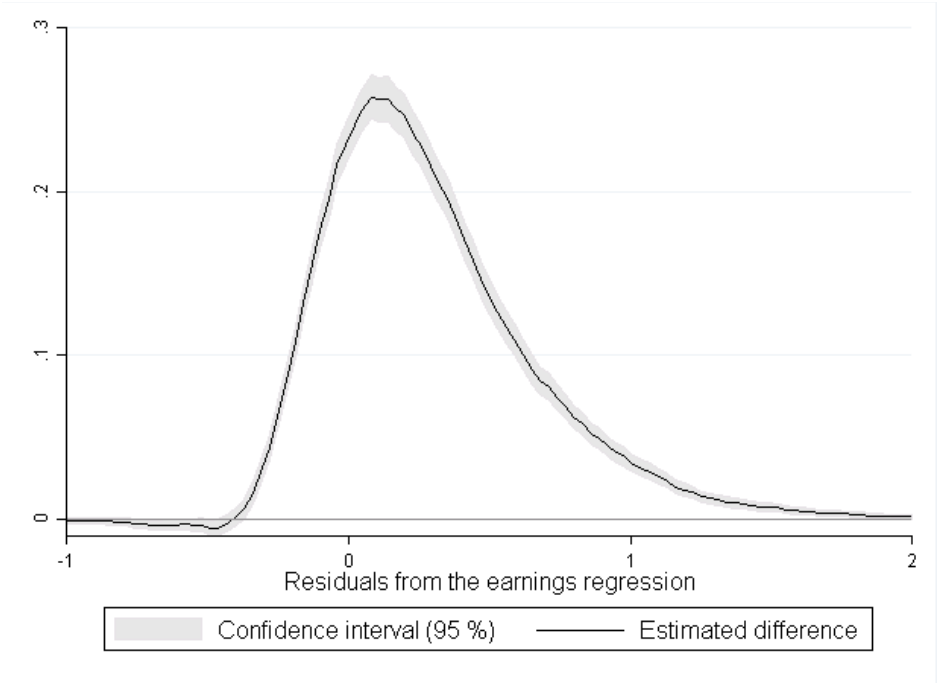


Figure 8. Difference of the cumulative distribution functions of residuals for migrants going outside other Nordic countries and non-migrants

a. Men



b. Women

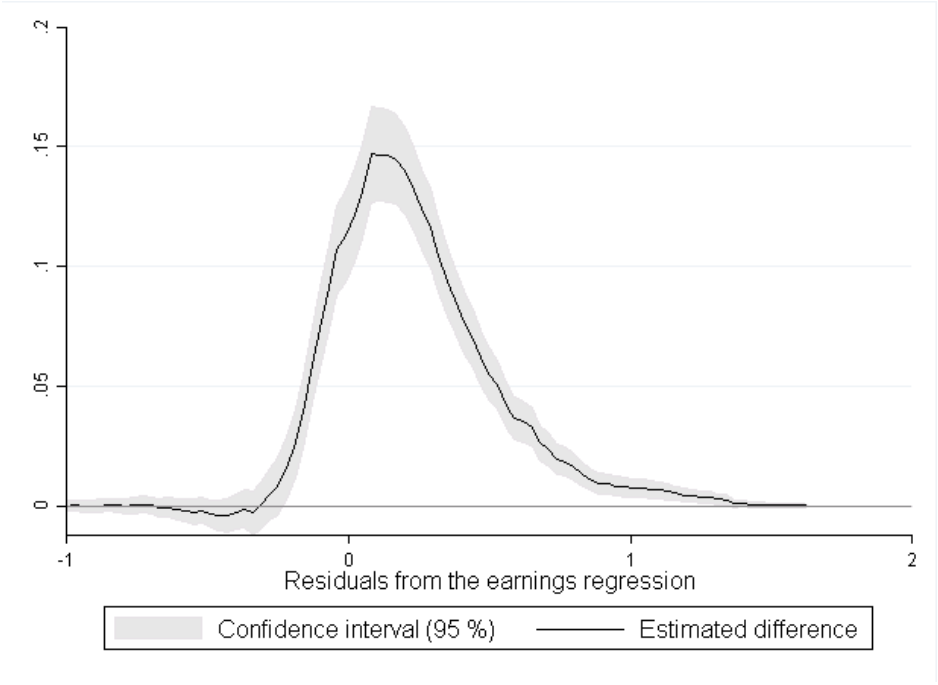
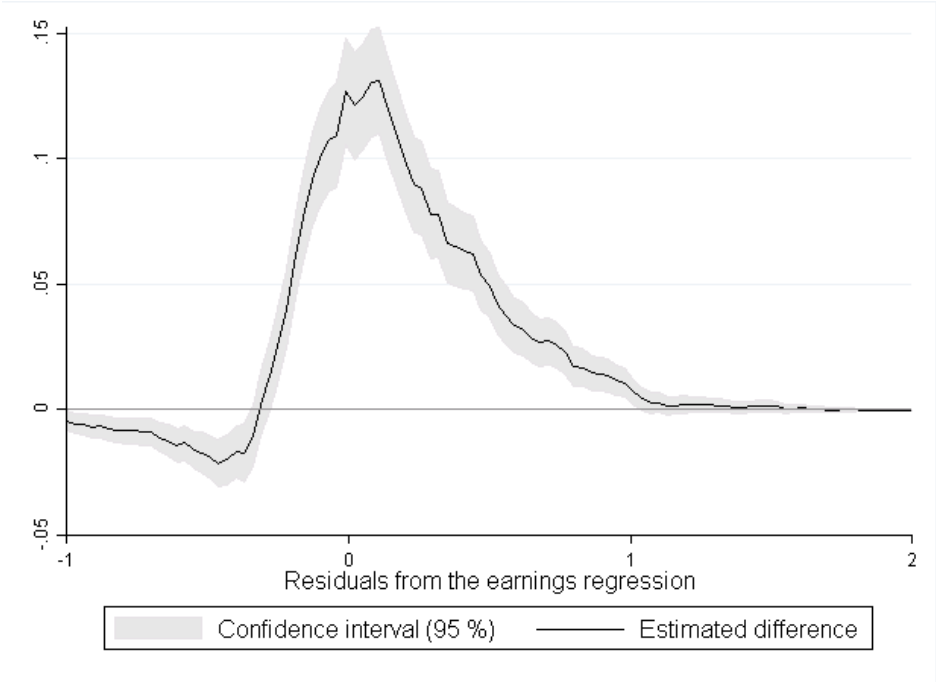


Figure 9. Difference of the cumulative distribution functions of residuals for migrants going to other Nordic countries and non-migrants

a. Men



b. Women

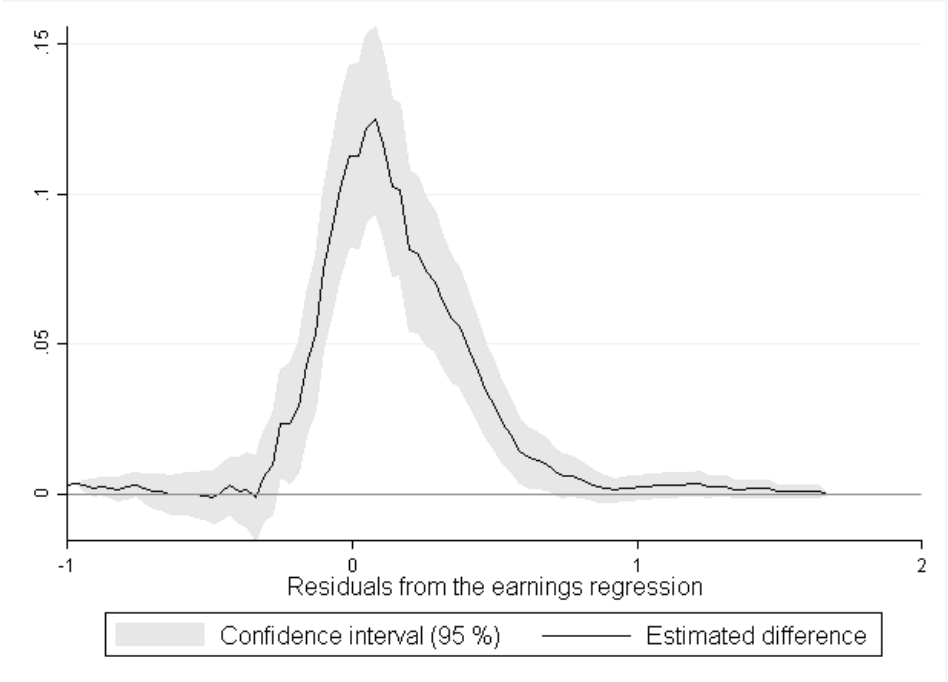
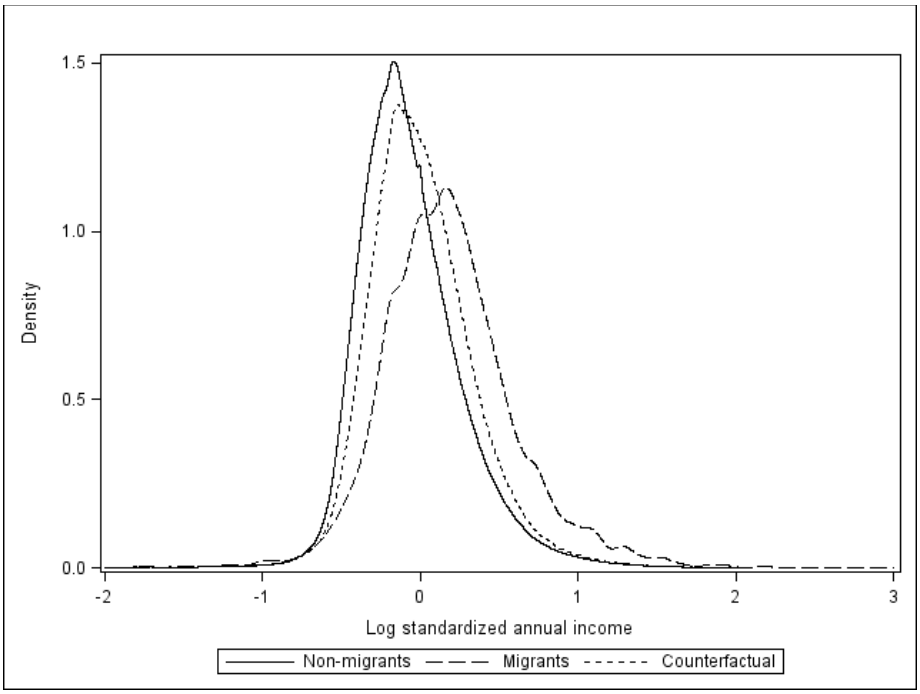


Figure 10. Counterfactual and actual densities of standardized gross earnings

a. Men



b. Women

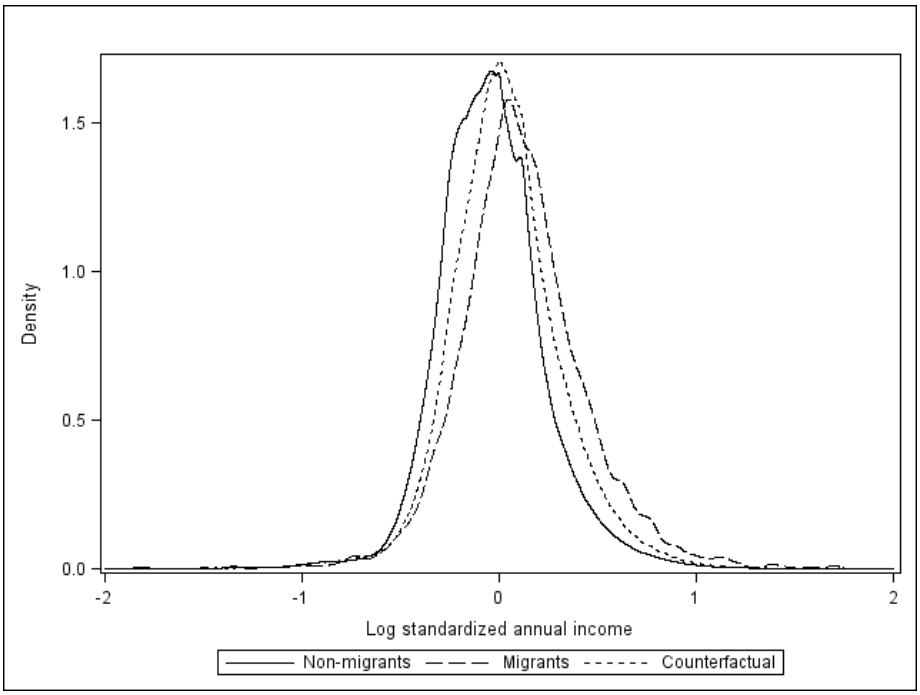
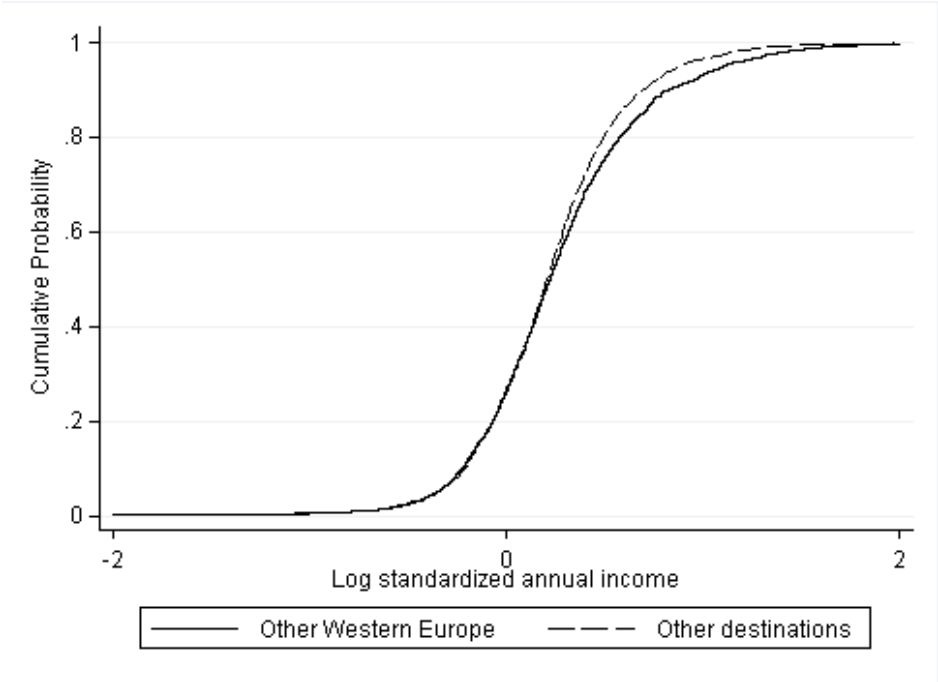


Figure 11. Distribution functions of annual gross earnings for migrants to the EU15 and Switzerland and migrants to other destinations

a. Men



b. Women

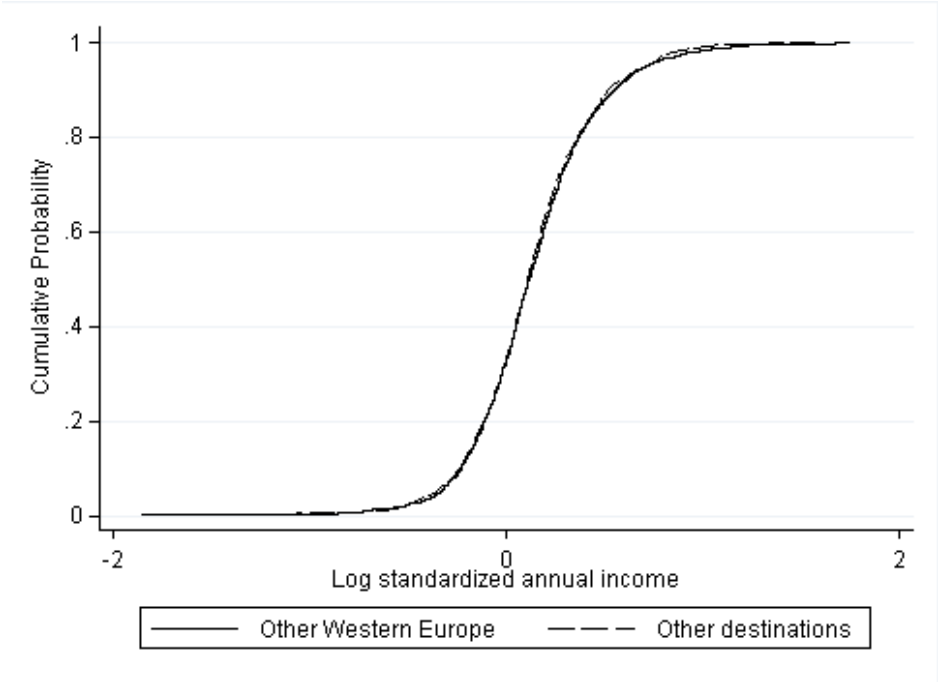


Table 1. Summary statistics

		Non-migrant men	Migrant men	Non-migrant women	Migrant women
Observations		6450665	7323	5163129	3436
Age					
	Average	39.8	33.0	40.2	35
	Median	40.0	35.4	40.0	33.0
Annual earnings in 2010 euros					
	Average	52725	68151	40299	46412
	Median	46675	57350	37976	42393
Standardized annual earnings					
	Average	1.0	1.3	1.0	1.2
	Median	0.9	1.2	0.95	1.1

Table 2. Numbers of migrants, by destination

	Men	Women
Sweden	1466	699
The United States	763	363
The United Kingdom	725	432
Norway	576	273
Germany	560	249
Spain	255	147
Switzerland	233	118
France	222	156
Other	2523	999

Table 3. Education levels of non-migrants and migrants going to Nordic countries or to other destinations

Education	Men			Women		
	Non-migrants	Nordic countries	Other destinations	Non-migrants	Nordic countries	Other destinations
Comprehensive school	21.4	19.8	8.3	21.5	15.7	8.9
High school	3.2	7.8	8.6	3.1	6.9	8.9
Vocational school	49.8	43.5	30.3	41.8	36.5	30.8
Advanced vocational	5.6	5.7	6.6	4.9	5.1	7.8
Bachelor or equivalent	12.2	11.6	20.6	23.3	22.9	25.4
Master's or equivalent	7.3	10.6	23.9	5.1	12.3	17.6
Doctoral or equivalent	0.5	1.0	1.7	0.2	0.7	0.7

Notes: The category "advanced vocational" includes all the tertiary education programs below the level of a Bachelor's program or equivalent. Programs on this level may be referred to for instance with such terms as community college education, advanced vocational training or associate degree.

Table 4. Summary of tests of stochastic dominance in distributions of standardized pre-migration earnings

Distributions being compared:	Percent of sample below lower bound		Percent of sample above upper bound	
	Migrants	Non-migrants	Migrants	Non-migrants
Migrants outside Nordic countries and non-migrants				
Male	2.0	3.4	0.1	0.0
Female	2.8	4.1	0.2	0.0
Migrants to Nordic countries and non-migrants				
Male	11.6	15.5	1.5	0.7
Female	2.6	4.1	2.6	1.3

Notes: Lower bound and upper bound refer to the range over which the difference of the cumulative distribution functions is significant at a 95 percent confidence level.

Table 5. Mincerian earnings regressions, by gender

	(1) men		(2) women	
	B	Se	B	se
Married	0.068***	(0.00)	-0.016***	(0.00)
Children	0.025***	(0.00)	-0.048***	(0.00)
High school	0.224***	(0.00)	0.190***	(0.00)
Vocational school	0.092***	(0.00)	0.089***	(0.00)
Advanced vocational	0.186***	(0.00)	0.198***	(0.00)
Bachelor	0.298***	(0.00)	0.225***	(0.00)
Master's	0.498***	(0.00)	0.536***	(0.00)
PhD	0.490***	(0.00)	0.622***	(0.00)
1996	0.020***	(0.00)	0.017***	(0.00)
1997	0.043***	(0.00)	0.041***	(0.00)
1998	0.078***	(0.00)	0.083***	(0.00)
1999	0.103***	(0.00)	0.112***	(0.00)
2000	0.141***	(0.00)	0.143***	(0.00)
2001	0.175***	(0.00)	0.175***	(0.00)
2002	0.207***	(0.00)	0.210***	(0.00)
2003	0.236***	(0.00)	0.235***	(0.00)
2004	0.252***	(0.00)	0.258***	(0.00)
Constant	12.131***	(0.00)	11.931***	(0.00)
Age fixed effects	Yes		Yes	
N	6470720		5173706	
R-squared	0.2597		0.3062	
*p<0.05, ** p<0.01, *** p<0.001				

The table reports OLS results for the log annual earnings. Individually clustered standard errors are in parentheses. Coefficients for the age dummies are not shown.

Table 6. Summary of tests of stochastic dominance in distributions of residuals

Distributions being compared:	Percent of sample below lower bound		Percent of sample above upper bound	
	Migrants	Non-migrants	Migrants	Non-migrants
Migrants outside Nordic countries and non-migrants				
Male	9.9	15.2	0.1	0.0
Female	19.6	24.7	0.4	0.0
Migrants to Nordic countries and non-migrants				
Male	13.4	15.2	2.0	0.9
Female	19.5	24.7	3.4	1.8

Notes: Lower bound and upper bound refer to the range over which the difference of the cumulative distribution functions is significant at a 95 percent confidence level.

Table 7. Logit estimates of the probability of emigration, by gender

	(1) men		(2) women	
	B	Se	B	se
Married	-0.110**	(0.04)	-0.191***	(0.05)
Children	-1.137***	(0.05)	-1.232***	(0.07)
Married*Children	0.460***	(0.07)	0.374***	(0.09)
High school	1.377***	(0.05)	1.158***	(0.08)
Vocational school	0.186***	(0.04)	0.159**	(0.06)
Advanced vocational	0.648***	(0.06)	0.714***	(0.08)
Bachelor	1.097***	(0.04)	0.581***	(0.06)
Master's	1.652***	(0.04)	1.444***	(0.07)
PhD	1.723***	(0.10)	1.655***	(0.21)
y1996	-0.032	(0.06)	-0.001	(0.08)
y1997	0.002	(0.06)	-0.016	(0.08)
y1998	-0.024	(0.06)	-0.001	(0.08)
y1999	0.230***	(0.05)	0.131	(0.08)
y2000	0.260***	(0.06)	0.238**	(0.09)
y2001	0.161**	(0.05)	0.146	(0.08)
y2002	0.208***	(0.05)	0.046	(0.08)
y2003	0.198***	(0.05)	0.112	(0.08)
y2004	0.246***	(0.05)	0.178*	(0.08)
Constant	-6.700***	(0.08)	-6.951***	(0.12)
N	6470720		5173706	
Pseudo R^2	0.0540		0.0557	
*p<0.05, ** p<0.01, *** p<0.001				

Notes: The table reports logit results for the long-term emigration. Individually clustered standard errors are in parentheses. Coefficients for the age fixed effects are not shown.

Table 8. Actual and counterfactual differences between the average log standardized earnings of migrants and non-migrants

	Men	Women
Non-migrant average	-0.065	-0.040
Estimated average for migrants	0.008	0.034
True average for migrants	0.180	0.117
True difference	0.245	0.157
Counterfactual difference	0.073	0.074
Share of the actual difference explained by observable characteristics, %	29.6	47.0