

# Nonparametric Identification and Estimation in a Generalized Roy Model

by

Patrick Bayer  
Duke University and NBER

Shakeeb Khan  
Duke University

Christopher Timmins  
Duke University

July 2006

PRELIMINARY AND INCOMPLETE

## **Abstract**

The classic Roy (1951) model characterizes the optimal choice of occupation given wage offers from two or more sectors. Heckman and Honore (1990) demonstrate that any cross-sectional dataset (i.e., the observed distribution of wages in each sector and the probabilities that each sector is chosen) can be rationalized by underlying population wage distributions in which wage offers are distributed independently across sectors. We show in this paper that, under relatively innocuous assumptions (i.e., wage distributions have finite lower supports or wage distributions do not depend upon origins in a migratory setting), it is possible to non-parametrically identify a generalized Roy model that includes a non-pecuniary component of utility associated with each alternative. These results apply even in empirical settings characterized by many alternative sectors and without the need for additional covariates. The resulting generalized Roy model is of interest in classic labor economics applications of the Roy model. More generally, such a model can be used to study any setting where individuals choose among discrete treatments or behaviors and where the researcher observes both the distribution of choice probabilities and the distribution of treatment outcomes.

## 1. Introduction

Because of the general nature of the problem that it describes, the Roy (1951) model has received extensive attention in the economics literature. In its initial form (and in many subsequent applications), the Roy model has been used to characterize the optimal choice of occupation or economic sector given wage offers from two or more sectors. In responding to their idiosyncratic wage offers, individuals select the sector that provides the greatest returns. This non-random selection generally implies that the underlying population wage distributions differ significantly from the wage distributions that are observed conditional on the observed choice of sector.

In a well-known study of the inference problem that arises in the classic Roy model, Heckman and Honore (1990) provide a number of key identification results. Most relevant for our paper is their central *non-identification* result. In particular, they demonstrate that any cross-sectional dataset (i.e., the observed distribution of wages in each sector and the probabilities that each sector is chosen) generated by agents behaving according to Roy model sorting can be rationalized by an underlying population wage distribution in which wage offers are distributed independently across sectors.<sup>1</sup> Thus, in a single cross-section, the correlation of wage offers across sectors is unidentified.

A common interpretation of this important result is that, without parametric restrictions or the availability of covariates, all of the useful content of a cross-sectional dataset with observed choice probabilities and wage distributions is absorbed in a restrictive specification that imposes independence. While this is certainly true *within* the pure Roy model, we will demonstrate in this paper that it is in fact possible to glean additional information from such a dataset – information that can be used to identify a more general version of the Roy model that allows for a common non-pecuniary component of utility associated with each sector.

The generalization of the Roy model to allow for non-pecuniary aspects of the decision problem is of interest in many applications. In the classic application to choice of occupation or sector, this non-pecuniary component of utility might capture variation in the amenity value or risk of injury across sectors. Likewise, in modeling the choice of a geographic labor market in which to locate, this component of utility would capture variation in amenities and housing prices across cities. More generally, such a model might be used to study any setting where individuals choose among discrete treatments or behaviors, where the researcher observes both the

---

<sup>1</sup> More specifically, Heckman and Honore argue that one can recover a different set of underlying population wage distributions for any assumed correlation in wage draws, and that there is no additional information remaining in the data.

distribution of choice probabilities and the distribution of treatment outcomes, and where individuals care about more than just the treatment outcome in making their decisions. In studying the choice of health behaviors, hospitals, or medical treatment, for example, the relevant outcome might be surviving days or the number of sick days, while the non-pecuniary component of utility might capture the enjoyment associated with a behavior (such as smoking) or variation in side-effects across various treatment outcomes. Or, in the study of school choice, the relevant outcome might be an achievement score, while other factors affecting the choice of school might be included as part of a separate component of utility unrelated to achievement score outcomes.

To keep the exposition of our results as parsimonious as possible, we assume that, as in Heckman and Honore (1990), the available data characterize a single cross-section of choice probabilities and observed wage distributions. We begin by considering the nonparametric identification of the model in a relatively demanding setting in which (i) the set of available choices is large and (ii) covariates are not available to the researcher.<sup>2</sup> The objects to be identified are the population wage distributions and a common non-pecuniary utility associated with each choice. Given the key result of Heckman and Honore (1990), we assume that wages are distributed independently across sectors. While this assumption can certainly be relaxed in more general data environments that have been studied in the existing literature, our goal is to focus on the simplest data environment possible that allows for a non-pecuniary component and many alternatives and thus we maintain independence throughout our formal analysis.

We prove identification under two sets of assumptions and provide a corresponding estimator in each case. For our first identification proof, we require only that the wage (outcome) distribution in each sector has a finite lower point of support. Given this assumption (which is likely to be innocuous in most settings), we demonstrate that the difference in the minimum order-statistic for any two alternatives exactly identifies the difference in the non-pecuniary value of those choices. Intuitively, that the minimum order statistic identifies the difference in the non-pecuniary component of utility follows directly from the observation that no individual will choose a less-preferred choice (on the basis of non-pecuniary considerations) unless the wage offered there exceeds this threshold. Thus, the minimum wage observed in the less-preferred sector should be exactly the minimum wage observed in the more-preferred sector plus the

---

<sup>2</sup> The related problem when covariates are available has been studied extensively in the literature, especially in the binomial choice problem (see, for example, Heckman (1990)). While circumstances where alternative strategies (e.g., “identification at infinity”) are applicable (e.g., the binomial choice setting), such strategies are difficult in practice to extend to the multinomial choice setting. There, one would need access to many covariates that, when set to alternative values, would always cause individuals to choose each of the available choices.

difference in non-pecuniary components.<sup>3</sup> Having identified the non-pecuniary component of utility, it is then straightforward to back-out the independent population wage distributions using transformed versions of the observed conditional wages distributions for each sector.

The corresponding estimator that we propose is straightforward and follows the identification proof directly. In particular, we first use the minimum order statistic to recover consistent estimates of the non-pecuniary component of utility. We then transform the wage distributions by the estimated non-pecuniary component of utility and use a simple Kaplan-Meier (1958) estimator to recover the independent population wage distributions.

While this estimator works very well in controlled data environments, relying exclusively on differences in minimum order statistics to identify the non-pecuniary component of utility naturally raises concerns about robustness to more general data generating processes. In particular, the difference in minimum order statistics is likely to provide a very noisy measure of non-pecuniary differences when wages are affected by measurement error or the underlying decision problem includes some form of uncertainty about the future stream of wages. As a result, we consider a second set of formal identifying assumptions.

Our second identification proof is based on two key assumptions. First, we assume that information is available for (at least) two subsets of the population that differ in their non-pecuniary valuation of the set of alternative sectors. In the application that we present below, for example, we consider the choice of regional labor market; in that context, moving costs (broadly defined) naturally imply that birth region affects the non-pecuniary value. In this way we exploit the fact that wage offers are likely to be similar for individuals with similar characteristics from neighboring regions while the non-pecuniary value of residing in these regions will vary significantly with an individual's birthplace. Given variation in non-pecuniary valuations across population subsets, we prove identification given a second assumption that we refer to as commonality, i.e., that a common wage distribution characterizes wage offers for all individuals regardless of birthplace. Given this assumption, we prove that both the non-pecuniary components of utility for each population subset and the overall population wage distributions are identified.

In this case, some intuition for why the model is identified by the commonality assumption can again be gained by referring back to the original Heckman and Honore (1990) proof described above. Without non-pecuniary components of utility, the observed conditional

---

<sup>3</sup> Notice that *within* the pure Roy model, the minimum order-statistics would be identical for all choices and, as a result, the full empirical content of the data would in fact be absorbed by a specification that included independent population wage distributions as in Heckman and Honore (1990).

wage distributions and choice probabilities map uniquely to a set of independent population wage distributions. With at least two subsets of the population that differ in their non-pecuniary valuations of alternatives, however, the resulting population wage distributions that would reconcile the two subsets of the data would differ. What our identification proof ensures is that the identical population wage distributions for each subset can only be reconciled at the true values of the non-pecuniary components of utility for each population subset.

Estimation of this model follows directly from the identification proof. As we show below, it is possible to write a system of equations based on the observed conditional wage distributions that must equal zero identically at the true values of the non-pecuniary parameters of the utility function for each population subset. These equations thus serve as natural moments for a minimum distance estimator. We simply search over the non-pecuniary utility parameters until the corresponding objective function is minimized.

### ***1.1 Previous Literature***

The study of the nonparametric identification of a generalized Roy model with many alternatives and non-pecuniary components of utility has been sparse in the existing literature. Dahl (2001) proposes a multinomial version of the estimator developed in the binomial context in Ahn and Powell (1994). This extension relies on the key assumption that a non-parametric selection correction term can be based on the first-best choice probability. But, this assumption is not made on the primitives of the model and it is not clear that this assumption holds in any case that is not degenerate – i.e., any case in which non-random selection of sectors actually occurs.

Applied work on spatial sorting behavior based on wages and non-pecuniary benefits has been more common in the literature. For example, Falaris (1987) and Davies, Greenwood, and Li (2001) study the determinants of migration decisions in Venezuela and the US, respectively. Falaris applies Lee’s (1983) generalized polychotomous choice model to control for non-random selection bias in conditional wage distributions, while Davies, Greenwood, and Li essentially ignore it. The entire literature on wage-hedonics, beginning with Roback (1982), has similarly ignored this problem. In those papers, wage and housing price gradients across cities are used to back-out the value of urban amenities. Wage distributions conditional on non-random selection into cities are typically used to calculate the first of these gradients, leading to biased estimates.

### ***1.2 Paper Outline***

To provide evidence on the performance of our proposed estimators, we conduct an extensive Monte Carlo analysis. We demonstrate the performance of our second estimator under

ideal circumstances and also a pair of mis-specifications. We then apply our estimator to US Census data to study the effect of spatial sorting on returns to a college education. We conclude by discussing the potential for extensions. Two issues are of particular importance – (i) relaxing the independence of wage distributions, (ii) considering non-degenerate distributions of tastes, (iii) minimum wages/participation decision

The remainder of the paper is organized as follows. Section 2 introduces the generalized Roy model that we study, proves identification in the case in which the wages (outcomes) are assumed to have a finite lower point of support, and develops a corresponding estimator. Section 3 proves identification under the alternative assumptions that information is available for multiple subsets of the population with different non-pecuniary valuation of sectors but with a common wage distribution, and develops a corresponding estimator. Section 4 provides the asymptotic theory related to both estimators. Section 5 provides our Monte Carlo results and Section 6 provides results for US Census data. Section 7 concludes.

## **2. Identification and Estimation – Finite Lower Support**

We begin our analysis by describing the generalized Roy model and data environment that we study. We then prove identification under two separate sets of assumptions. The first case is characterized by the assumption that the distribution of the endogenously determined payoffs (e.g., wages in a classic Roy model) has a finite lower support. In this case, our corresponding estimation strategy is transparent and easily applied. Our second set of assumptions, described in Section 3, is applicable in situations where a finite lower support cannot be assumed, or where the minimum order statistic provides a noisy measure of the lower bound. In both cases, we first prove identification with a simple model describing the sorting of individuals from a single origin location into one of two destinations ( $k = 1, 2$ ). We indicate the wage earned by an individual should he choose to settle in locations #1 and #2 as  $\omega_1$  and  $\omega_2$ , respectively. In contrast to the classic Roy model, where sorting is simply across employment sectors and driven entirely by pecuniary compensation, we model sorting in a geographic context where the individual’s location decision depends in part on his wage draw in each location, but also on non-wage determinants of utility specific to a particular location, which we label as “tastes”.<sup>4</sup> Utility from choosing to settle in location  $k$  is given by the sum of wages ( $\omega_k$ ) and tastes ( $\tau_k$ ):

---

<sup>4</sup> Tastes would certainly include natural amenities and local public goods associated with the destination location. In addition, they may include “migration costs”; i.e., costs specific to someone moving from a particular origin to a particular location. In a narrow sense, these costs would be comprised of re-location expenditures. In broader terms, these costs would likely involve the psychological costs of living far from

$$(1) \quad U_k = \omega_k + \tau_k$$

Without loss of generality, we normalize  $\tau_1 = 0$ .<sup>5</sup> The goal of our exercise is to recover estimates of  $\tau_2$ ,  $f_1(\omega_1)$ , and  $f_2(\omega_2)$  (i.e., the taste parameter associated with location #2 and the unconditional wage distributions in each location). The difficulty arises from the fact that we only see (i) wage distributions conditional upon optimal sorting behavior, and (ii) an indicator of which location an individual chooses.

### 2.1 Identification of Tastes Based on a Finite Lower Support

Our first approach uses only the conditional wage distributions and an indicator of location choice to recover  $\tau_2$ ,  $f_1(\omega_1)$ , and  $f_2(\omega_2)$  according to the following argument based on minimum order statistics. For an individual  $i$ , we only observe  $\omega_{2,i}$  if:

$$(2) \quad \omega_{2,i} + \tau_2 \geq \omega_{1,i}$$

and we only observe  $\omega_{1,i}$  if:

$$(3) \quad \omega_{2,i} + \tau_2 < \omega_{1,i}$$

Denote the smallest wage (i.e., the minimum order statistic) that we observe from someone choosing to settle in location #1 or #2 by  $\underline{\omega}_1$  and  $\underline{\omega}_2$ , respectively. Assuming that  $f_1(\omega_1)$  and  $f_2(\omega_2)$  have finite lower points of supports (denoted by  $\omega_1^*$  and  $\omega_2^*$ , respectively), we know that the smallest value of  $\omega_1$  that we could ever see given that individuals maximize utility:

$$(4) \quad \begin{array}{ll} \underline{\omega}_1 = \omega_1^* & \text{if } \omega_1^* > \omega_2^* + \tau_2 \\ \underline{\omega}_1 = \omega_2^* + \tau_2 & \text{if } \omega_1^* \leq \omega_2^* + \tau_2 \end{array}$$

---

one's birth location. 2000 Census data indicate that a majority of US household heads live in the narrowly defined region in which they were born. [Bayer, Keohane, and Timmins (2006)]

<sup>5</sup> As in all random-utility frameworks, utility is only identified up to an additive constant. This requires some sort of a normalization, which we use to eliminate one of the  $\tau$ 's from the two-destination example. In the more general  $N \times N$  case, we estimate  $(N-1)$   $\tau$ 's for each of the  $N$  origins.

Similarly, the smallest value of  $\omega_2$  that we could ever see would be:

$$(5) \quad \begin{aligned} \underline{\omega}_2 &= \omega_2^* && \text{if} && \omega_1^* \leq \omega_2^* + \tau_2 \\ \underline{\omega}_2 &= \omega_1^* - \tau_2 && \text{if} && \omega_1^* > \omega_2^* + \tau_2 \end{aligned}$$

In order to make sense of (4) and (5), define the following two cases:

$$(6) \quad \begin{aligned} A: & \omega_1^* > \omega_2^* + \tau_2 \\ B: & \omega_1^* \leq \omega_2^* + \tau_2 \end{aligned}$$

We are not able to tell whether case A or B prevails in the data without recovering an estimate of  $\tau_2$ . Conveniently, we are able to recover an estimate of  $\tau_2$  in either case. In particular:

$$(7) \quad \tau_2 = \underline{\omega}_1 - \underline{\omega}_2$$

regardless of which case we are in. Equation (7) therefore describes our first estimator of  $\tau_2$  in the simplest 1 x 2 case.

## 2.2 Identification of $f_1(\omega_1)$ and $f_2(\omega_2)$ with Kaplan-Meier

Having recovered an estimate of  $\tau_2$ , it is a simple matter to recover  $f_1(\omega_1)$  and  $f_2(\omega_2)$  by employing a variation of the Kaplan-Meier (1958) procedure typically used in competing-risks models. The Kaplan-Meier procedure has the interpretation as a nonparametric maximum likelihood estimator of a censored distribution, and has been proven to be asymptotically normally distributed- see, e.g. Gill(1980).

Our variation will be to apply the Kaplan Meier procedure to draws from  $\omega_1 + \tau_1$ , where  $\tau_1$  can be estimated using the proposed procedure. In particular, we estimate  $f_1(\omega_1)$  by first creating a new data vector which corresponds to only those values of utilities (i.e.,  $\omega_1 + \tau_1$ ) that are “uncensored” for destination #1 (i.e., observed for individuals who optimally chose destination #1). Note that, because we were able to recover tastes with equation (7), we can treat utility (i.e., the sum of wages and tastes) as observed for the remainder of the exercise – our only goal is to recover its unconditional distribution, from which we can recover the unconditional distribution of wages. This vector of utilities will be of smaller dimension than the vector of all utilities, which includes draws for individuals who chose destination #1 or destination #2.

To implement the Kaplan Meier procedure, we first sort this vector of uncensored utility observations from highest to lowest. We next create a “risk set” vector. In survival analysis setting, where survival times are usually right censored, the risk set corresponds to observations known to have not failed by the time in question. This vector will be of the same dimension as the vector of uncensored utilities, and will be constituted entirely of integers. Its first element will correspond to the number of observations in the original data set (i.e., both censored and uncensored) that are less than or equal to the first element in the newly created (sorted) vector – in other words, the number of observations in the original data set that are less than or equal to the largest uncensored observation. Analogously, the second component of the risk set vector is the number of observations in the original data set that are less than or equal to the second largest uncensored observation.

With this risk set vector, we estimate the c.d.f. of the uncensored random variable  $U_2 = \omega_2 + \tau_2$  as follows. Assume we want to estimate this c.d.f. at a point  $x$ . Call this parameter  $S$ . First set the initial estimate of  $S$  to 1. If the first component of the sorted uncensored vector is bigger than  $x$ , multiply  $S$  by  $[1 - (1/R_1)]$ , where  $R_1$  is the first component in the risk set vector created above. This is the new estimate of  $S$ . Proceed analogously to the second component of the risk set vector ( $R_2$ ). If the second component of the sorted uncensored vector is bigger than  $x$ , multiply the existing estimate of  $S$  by  $[1 - (1/R_2)]$ . This is the new estimate of  $S$ . We keep repeating this procedure until we have gone through the entire risk set vector. The resulting value of  $S$  is the Kaplan-Meier estimate of the c.d.f. of  $U_2$  at  $x$ . In the final step, we simply deduct our estimate of  $\tau_2$  from utility  $U_2$  at each point in the support of its distribution. The resulting distribution is a non-parametric representation of  $f_2(\omega_2)$ . We then repeat this process in order to recover  $f_1(\omega_1)$ , recalling that  $\tau_1$  had been normalized to zero.

Note that a portion of the unconditional distribution for one of these two locations will necessarily be censored. Suppose we are in case A, where  $\omega_1^*$  is large relative to  $\omega_2^* + \tau_2$ . We are therefore able to observe the complete distribution  $f_1(\omega_1)$ , beginning with  $\underline{\omega}_1 = \omega_1^*$ . We are, however, unable to observe  $f_2(\omega_2)$  to the left of  $\underline{\omega}_2 = \omega_1^* - \tau_2 > \omega_2^*$ . While we are unable to determine the shape of the distribution  $f_2(\omega_2)$  between  $\omega_2^*$  and  $\underline{\omega}_2$  in the above case, we are able to bound from above the value of  $\omega_2^*$  (i.e., the lower point of support for the censored distribution). In particular, knowing that  $\omega_1^* = \underline{\omega}_1$ , we know that  $\omega_2^* < \underline{\omega}_1 - \tau_2$ . We are unable

to determine more about the shape of the distribution  $f_2(\omega_2)$  between  $\omega_2^*$  and  $\underline{\omega}_2$  without resorting to parametric assumptions.

### 2.3 The 1 x 3 Case – Finite Lower Support

The theory used to describe the 1 x 2 case scales up naturally to the any number of potential origin and destination locations; we illustrate here the 1 x 3 case. With more than two potential destination locations, the estimation algorithm is easily explained with the help of some additional notation. Consider the following three-location system with wages for individual  $i$  denoted by  $\omega_{1,i}$ ,  $\omega_{2,i}$ , and  $\omega_{3,i}$ . We denote the lower supports of each location's wage distribution by  $\omega_1^*$ ,  $\omega_2^*$ , and  $\omega_3^*$ , and model individuals coming only from origin location #1. We therefore normalize  $\tau_1 = 0$ . For individual  $k$ , we observe  $w_k$ , where:

$$(8) \quad \begin{aligned} w_i = & \omega_{1,i} I[\omega_{1,i} > \max(\omega_{2,i} + \tau_2, \omega_{3,i} + \tau_3)] + \\ & \omega_{2,i} I[\omega_{2,i} + \tau_2 > \max(\omega_{1,i}, \omega_{3,i} + \tau_3)] + \\ & \omega_{3,i} I[\omega_{3,i} + \tau_3 > \max(\omega_{1,i}, \omega_{2,i} + \tau_2)] \end{aligned}$$

We also observe an indicator corresponding to which destination location individual  $i$  has moved – i.e.,  $d_{1,i}$ ,  $d_{2,i}$ , and  $d_{3,i}$ . We note that, under convex supports for all random variables and assuming finite lower support points  $(\omega_1^*, \omega_2^*, \omega_3^*)$ , we have the following:

$$(9) \quad \underline{w}_1 = \min(w_i \mid d_{1,i} = 1) = \max(\omega_1^*, \omega_2^* + \tau_2, \omega_3^* + \tau_3)$$

$$(10) \quad \underline{w}_2 = \min(w_i \mid d_{2,i} = 1) = \max(\omega_1^*, \omega_2^* + \tau_2, \omega_3^* + \tau_3) - \tau_2$$

$$(11) \quad \underline{w}_3 = \min(w_i \mid d_{3,i} = 1) = \max(\omega_1^*, \omega_2^* + \tau_2, \omega_3^* + \tau_3) - \tau_3$$

Notice that  $\tau_3$  is given by  $(\underline{w}_1 - \underline{w}_3)$ , while  $\tau_2$  is given by  $(\underline{w}_1 - \underline{w}_2)$ . We can then proceed to use the Kaplan-Meier technique (applied to utilities as opposed to wages) to recover the unconditional wage distributions in the same manner as is described above.

### 3. Identification and Estimation – Unbounded Support

While clean and transparent, there are two practical problems with the technique outlined in Section 2. First, the payoff variable in question may not naturally have a finite lower support (e.g., theory might dictate using the natural log of wages in the utility function). Second, the minimum order statistic can be a very noisy measure of  $\underline{w}_1$ ,  $\underline{w}_2$ , or  $\underline{w}_3$ , as defined in equations (9) – (11).<sup>6</sup> Unless one has tremendous confidence in the estimate of the minimum order statistic, that noise will be translated directly through to the estimates of the taste parameters and, subsequently, on to the Kaplan-Meier estimates.

As an alternative, we propose in this section an estimator that employs data from the full distribution of conditional wages. Importantly, this approach is valid for an unbounded support.<sup>7</sup> With that flexibility, however, comes the need for an additional identification assumption. In particular, we begin by showing that, without an additional assumption,  $\tau_2$ ,  $f_1(\omega_1)$ , and  $f_2(\omega_2)$  are not identified. This negative proof, however, reveals just how easily identification can be achieved by exploiting the assumption of “commonality” described in Section 3.2.

### 3.1 Non-Identification in the 1 x 2 Case

We begin with a simple model of individuals sorting over two locations, indexed by 1 and 2. We assume for simplicity that the individuals are from location 1, and we therefore normalize their taste for staying there to zero ( $\tau_1 = 0$ ). Our interest is therefore in recovering estimates of  $\tau_2$ ,  $f_1(\omega_1)$ , and  $f_2(\omega_2)$ .

We define a variable  $d_i$ , which functions as an indicator that individual  $i$  remained in his origin location:

$$(12) \quad d_i = I[\omega_{1,i} > \omega_{2,i} + \tau_2]$$

Using this indicator, we can write down an expression for individual  $i$ 's observed wage:

$$(13) \quad w_i = d_i \omega_{1,i} + (1 - d_i) \omega_{2,i}$$

---

<sup>6</sup> For example, the bottom 2-3% of wage observations in the US Census data used for our empirical application in Section 7 are implausibly low (i.e., less than 50¢ per hour).

<sup>7</sup> In practice, this means that poorly measured data in the lower tail of the wage distribution will not have a significant impact on the estimation algorithm, whereas it can have severe effects on the minimum order statistic approach.

i.e., the individual receives his draw from location #1 if it was utility maximizing to stay there. Next, define the following joint probability distributions, both of which are easily observed in the data:

$$\begin{aligned} \Psi_1(t) &= P(d_i = 1, w_i \leq t) \\ (14) \quad \Psi_2(t) &= P(d_i = 0, w_i \leq t) \end{aligned}$$

We will also work with the derivatives of these expressions, which we denote by:

$$\begin{aligned} \psi_1(t) &= \frac{\partial}{\partial t} P(d_i = 1, w_i \leq t) \\ (15) \quad \psi_2(t) &= \frac{\partial}{\partial t} P(d_i = 0, w_i \leq t) \end{aligned}$$

Focusing on the expression for  $\Psi_1(t)$ , we can re-write it as

$$\begin{aligned} \Psi_1(t) &= P(d_i = 1, w_i \leq t) \\ (16) \quad &= P(\omega_{1,i} > \omega_{2,i} + \tau_2, \omega_{1,i} \leq t) = P(\omega_{1,i} - \tau_2 > \omega_{2,i}, \omega_{1,i} \leq t) \\ &= \int_{-\infty}^t f_1(\omega_1) d\omega_1 \int_{-\infty}^{\omega_1 - \tau_2} f_2(\omega_2) d\omega_2 = \int_{-\infty}^t f_1(\omega_1) F_2(\omega_1 - \tau_2) d\omega_1 \end{aligned}$$

This means that we can define  $\psi_1(t)$  as follows:

$$(17) \quad \psi_1(t) = \frac{\partial}{\partial t} \int_{-\infty}^t f_1(\omega_1) F_2(\omega_1 - \tau_2) d\omega_1 = f_1(t) F_2(t - \tau_2)$$

An analogous argument defines  $\psi_2(t)$ :

$$(18) \quad \psi_2(t) = \frac{\partial}{\partial t} \int_{-\infty}^t f_2(\omega_2) F_1(\omega_2 + \tau_2) d\omega_2 = f_2(t) F_1(t + \tau_2)$$

Going back to the final integral in equation (16) and carrying out integration-by-parts yields:

$$(19) \quad \Psi_1(t) = \int_{-\infty}^t f_1(\omega_1) F_2(\omega_1 - \tau_2) d\omega_1 = F_1(t) F_2(t - \tau_2) - \int_{-\infty}^t F_1(s) f_2(s - \tau_2) ds$$

Performing a change of variables  $u = s - \tau_2$ , equation (19) becomes:

$$(20) \quad \Psi_1(t) = F_1(t) F_2(t - \tau_2) - \int_{-\infty}^{t - \tau_2} F_1(u + \tau_2) f_2(u) du$$

Next, we use the expressions for  $\psi_1(t)$  and  $\psi_2(t)$  defined in (17) and (18) to re-write equation (20) as follows:

$$(21) \quad \Psi_1(t) = \frac{F_1(t) \psi_1(t)}{f_1(t)} - \int_{-\infty}^{t - \tau_2} \psi_2(u) du$$

Noting that the second integral in (21) is simply  $\Psi_2(t - \tau_2)$ , we can solve for the distribution of  $\omega_1$  as a function of  $\tau_2$ :

$$(22) \quad \lambda_1(t) = \frac{f_1(t)}{F_1(t)} = \frac{\psi_1(t)}{\Psi_1(t) + \Psi_2(t - \tau_2)}$$

where  $\lambda_1(t)$  is a function of the unconditional wage distribution in location #1. (22) is a single equation in two unknowns ( $\lambda_1(t)$  and  $\tau_2$ ) for a particular value of  $t$ , and it is therefore not surprising that we cannot identify both of these values without making an additional assumption. One solution would involve making a parametric assumption about  $F_1(t)$ . For example, assuming  $F_1(t) \sim N(\mu_1, \sigma_1^2)$  would reduce the equation to three parameters. The number of parameters would not increase, however, as one considered the expression evaluated at different values of  $t$ .

By forcing the equation to hold for many values of  $t$ , we would have more equations than unknowns and could identify the model's parameters.

In the following section, we show how the assumption of commonality can be used to non-parametrically recover  $\lambda_1(t)$  and  $\tau_2$ .

### 3.2 Identification via Commonality in the 2 x 2 Case

Consider now the case of individuals born into one of two locations (again indexed by 1 and 2), who decide where to reside based on the maximization of utility. This introduces the need for additional notation – we use a superscript to indicate origin location and a subscript to indicate destination location.

The dummy variable indicating that an individual originating in location #1 chooses to stay in that location is given by:

$$(23) \quad d_i^1 = I[\omega_{1,i}^1 > \omega_{2,i}^1 + \tau_2^1]$$

while the indicator that an individual originating in location #2 chooses not to migrate is given by:

$$(24) \quad d_i^2 = I[\omega_{2,i}^2 > \omega_{1,i}^2 + \tau_1^2]$$

As before, we normalize the taste parameter for those choosing not to migrate to zero (i.e.,  $\tau_1^1 = \tau_2^2 = 0$ ). With these indicators, we can now write the expression for the observed wage of an individual  $i$  who originates in location #1:

$$(25) \quad w_i^1 = d_i^1 \omega_{1,i}^1 + (1 - d_i^1) \omega_{2,i}^1$$

Based on these definitions for  $d$  and  $w$ , we define the following expressions analogously to the previous sub-section:

$$(26) \quad \begin{aligned} \Psi_1^1(t) &= P(d_i^1 = 1, w_i^1 \leq t) & \Psi_2^1(t) &= P(d_i^1 = 0, w_i^1 \leq t) \\ \Psi_1^2(t) &= P(d_i^2 = 0, w_i^2 \leq t) & \Psi_2^2(t) &= P(d_i^2 = 1, w_i^2 \leq t) \end{aligned}$$

Continuing in a manner similar to the previous sub-section, we can use equation (26) to derive the following four expressions:

$$(27) \quad \lambda_1^1(t) = \frac{f_1^1(t)}{F_1^1(t)} = \frac{\psi_1^1(t)}{\Psi_1^1(t) + \Psi_2^1(t - \tau_2^1)}$$

$$(28) \quad \lambda_2^1(t) = \frac{f_2^1(t)}{F_2^1(t)} = \frac{\psi_2^1(t)}{\Psi_2^1(t) + \Psi_1^1(t + \tau_2^1)}$$

$$(29) \quad \lambda_1^2(t) = \frac{f_1^2(t)}{F_1^2(t)} = \frac{\psi_1^2(t)}{\Psi_1^2(t) + \Psi_2^2(t + \tau_1^2)}$$

$$(30) \quad \lambda_2^2(t) = \frac{f_2^2(t)}{F_2^2(t)} = \frac{\psi_2^2(t)}{\Psi_2^2(t) + \Psi_1^2(t - \tau_1^2)}$$

By itself, the expansion of the 1 x 2 case to the 2 x 2 case does nothing to help with identification. It does, however, allow us to introduce an additional assumption – commonality. Under the assumption of commonality,  $\lambda_1^1(t) = \lambda_1^2(t)$  and  $\lambda_2^1(t) = \lambda_2^2(t) \forall t$ . Under this assumption, we can re-write equations (27)-(30) as the following two equations:

$$(31) \quad \lambda_1^1(t) = \frac{\psi_1^1(t)}{\Psi_1^1(t) + \Psi_2^1(t - \tau_2^1)} = \frac{\psi_1^2(t)}{\Psi_1^2(t) + \Psi_2^2(t + \tau_1^2)} = \lambda_1^2(t)$$

$$(32) \quad \lambda_2^1(t) = \frac{\psi_2^1(t)}{\Psi_2^1(t) + \Psi_1^1(t + \tau_2^1)} = \frac{\psi_2^2(t)}{\Psi_2^2(t) + \Psi_1^2(t - \tau_1^2)} = \lambda_2^2(t)$$

Estimation proceeds by forming minimum distance criterion functions based on equations (31) and (32):

$$(33) \quad \lambda_1^1(t; \tau_2^1) - \lambda_1^2(t; \tau_1^2) = 0$$

$$(34) \quad \lambda_2^1(t; \tau_2^1) - \lambda_2^2(t; \tau_1^2) = 0$$

and then relying on the properties of M-estimators to recover  $\tau_2^1$  and  $\tau_1^2$ . [Davidson and MacKinnon (1993)] We then use these taste parameters along with a Kaplan-Meier procedure to recover estimates of  $f_1(\omega_1)$  and  $f_2(\omega_2)$  as described in Section 2.2.

We now provide sufficient conditions for identification and estimation of the taste parameters in the 2 x 2 setting with commonality. We begin by rearranging the expressions (31) and (32):

$$(35) \quad \Psi_2^1(t - \tau_2^1)\psi_1^2(t) - \psi_1^1(t)\Psi_2^2(t + \tau_1^2) = \Psi_1^2(t)\psi_1^1(t) - \Psi_1^1(t)\psi_1^2(t) = H(t)$$

$$(36) \quad \Psi_1^1(t + \tau_2^1)\psi_2^2(t) - \psi_2^1(t)\Psi_1^2(t - \tau_1^2) = \Psi_2^2(t)\psi_2^1(t) - \Psi_2^1(t)\psi_2^2(t) = J(t)$$

Note that the right-hand-side of each of these expressions is an observable function of the data for a particular value of  $t$ . Our identification result begins with the following lemma:

**Lemma 1:** *At the true parameter values  $(\tau_2^{1*}, \tau_1^{2*})$ , we have*

$$(37) \quad \begin{aligned} & \left( \Psi_2^1(t - \tau_2^{1*})\psi_1^2(t) - \psi_1^1(t)\Psi_2^2(t + \tau_1^{2*}) - H(t) \right)^2 + \\ & \left( \Psi_1^1(t + \tau_2^{1*})\psi_2^2(t) - \psi_2^1(t)\Psi_1^2(t - \tau_1^{2*}) - J(t) \right)^2 = 0 \end{aligned}$$

$\forall t \in \mathfrak{R}$  in the intersection of the supports of  $\psi_1^1(t)$ ,  $\psi_2^1(t)$ ,  $\psi_1^2(t)$ , and  $\psi_2^2(t)$ .

This is simply a re-statement of our minimum distance criterion function described above. We will now show that, for each set of values of the taste parameters different from  $(\tau_2^{1*}, \tau_1^{2*})$ , denoted by  $(\tilde{\tau}_2^1, \tilde{\tau}_1^2)$ , we must have:

$$(38) \quad \left( \Psi_2^1(t - \tilde{\tau}_2^1) \psi_1^2(t) - \psi_1^1(t) \Psi_2^2(t + \tilde{\tau}_1^2) - H(t) \right)^2 + \left( \Psi_1^1(t + \tilde{\tau}_2^1) \psi_2^2(t) - \psi_2^1(t) \Psi_1^2(t - \tilde{\tau}_1^2) - J(t) \right)^2 > 0$$

for some  $t \in \mathfrak{R}$  in the intersection of the supports of  $\psi_1^1(t)$ ,  $\psi_2^1(t)$ ,  $\psi_1^2(t)$ , and  $\psi_2^2(t)$ . To prove this result, first note that if  $\tilde{\tau}_2^1 = \tau_2^1 *$ , then only  $\tilde{\tau}_1^2 = \tau_1^2 *$  will make equation (37) hold, by the monotonicity of the conditional c.d.f.'s that make-up that expression. By a similar argument, if  $\tilde{\tau}_1^2 = \tau_1^2 *$ , then  $\tilde{\tau}_2^1 = \tau_2^1 *$  in order for equation (37) to hold. Therefore, we need only consider the case in which  $\tilde{\tau}_2^1 \neq \tau_2^1 *$  and  $\tilde{\tau}_1^2 \neq \tau_1^2 *$ . I.e., is it possible that an imposter pair  $(\tilde{\tau}_2^1, \tilde{\tau}_1^2)$  could satisfy equation (37)?

Consider the following condition which we argue will be sufficient to rule out this possibility:

$$(39) \quad \psi_2^1(t - \tau_2^1 *) \psi_1^2(t) \psi_2^1(t) \psi_1^2(t - \tau_1^2 *) \neq \psi_1^1(t + \tau_2^1 *) \psi_2^2(t) \psi_1^1(t) \psi_2^2(t + \tau_1^2 *)$$

for some  $t \in \mathfrak{R}$  in the intersection of the supports of  $\psi_1^1(t)$ ,  $\psi_2^1(t)$ ,  $\psi_1^2(t)$ , and  $\psi_2^2(t)$ . This condition has a simple interpretation – i.e., that the Jacobian matrix associated with equations (35) and (36) is non-singular. There are situations in which this condition will not hold; for example, when the two conditional wage distributions are identical and  $\tau_2^1 = -\tau_1^2$ .<sup>8</sup> We consider this to be a pathological case.

To establish the sufficiency of the above condition for identification, consider a local linearization of equations (35) and (36) around the true values of  $\tau_2^1$  and  $\tau_1^2$  and evaluated at  $t$ . For any pair of perturbations,  $\Delta_2^1$  and  $\Delta_1^2$ , we require the net effect on the left-hand-side of each equation to be zero (since  $H(t)$  and  $J(t)$  are functions of only  $t$ ).

$$(40) \quad \psi_2^1(t - \tau_2^1 *) \psi_1^2(t) \Delta_2^1 + \psi_1^1(t) \psi_2^2(t + \tau_1^2 *) \Delta_1^2 = 0$$

---

<sup>8</sup> This would be the case if we took a single location and arbitrarily divided it into two locations with the exact same wage distributions and amenities. This condition therefore places a practical constraint on the level of geographic precision at which we can apply our estimator – i.e., at the level at which we can observe different spatial wage distributions.

$$(41) \quad \psi_1^1(t + \tau_2^{1*})\psi_2^2(t)\Delta_2^1 - \psi_2^1(t)\psi_1^2(t - \tau_1^{2*})\Delta_1^2 = 0$$

If condition (39) holds, then the only solution to these expressions is given by  $\Delta_2^1 = \Delta_1^2 = 0$ , implying that no imposter values of  $(\tilde{\tau}_2^1, \tilde{\tau}_1^2)$  could satisfy the system. Of course, we do not know the value(s) of  $t$  where equation (39) holds, requiring that we evaluate our minimum distance estimator at all available values of  $t$ . In doing so, we are restricted to only using values of  $t \in \mathfrak{R}$  in the intersection of the supports of  $\psi_1^1(t)$ ,  $\psi_2^1(t)$ ,  $\psi_1^2(t)$ , and  $\psi_2^2(t)$ . Without any overlap, this identification strategy is not applicable.

#### 4. Asymptotics

*[to be added]*

#### 5. Monte Carlo Simulations

We illustrate the properties of the unbounded support estimator with a series of Monte Carlo simulations.<sup>9</sup> The first two sets of simulations are designed to show how the estimator performs given that the assumptions used for identification are true – i.e., commonality of the unconditional wage distributions irrespective of the origin location and independence of wage draws across destination locations. In the final two Monte Carlo simulations, we relax the assumption of global commonality and independence in the data generating process. In particular, we show how the researcher can still learn a lot about tastes and unconditional wage distributions even without global commonality, as long as the researcher is willing to impose some structure on tastes and the pattern of commonality. Finally, we show how the model’s predictive powers degenerate when correlation is built into the wage draws. Adapting the estimator to handle the case of correlated draws is a focus of our continuing research.

Simulation results describe the bias and mean squared error (MSE) of the taste parameters, the medians, and 75<sup>th</sup> percentiles of the estimated unconditional wage distributions. We vary (i) the number of locations (which serve both as origins and destinations) and (ii) the

---

<sup>9</sup> For the sake of parsimony, we do not present any Monte Carlo results for the finite lower bound estimator. With high quality data and if the finite lower bound assumption is not violated, this estimator perfectly replicates the underlying distributions and taste parameters (results available upon request). With a small sample size or measurement error (i.e., situations one would expect to come across in many applications), however, the finite lower support estimator does not perform well. We therefore focus our attention on the unbounded support estimator in small sample simulations.

number of simulated observations from each origin location. We begin with a simple 2 x 2 case. Simulated individual  $i$  receives an i.i.d. wage draw in locations #1 and #2 from the following distributions:

$$(42) \quad f_1(\omega_{i,1}^j) \sim N(0, 1) \quad f_2(\omega_{i,2}^j) \sim N(0.35, 1)$$

Under commonality, these distributions are the same irrespective of birth location  $j = 1, 2$ . We use the following matrix of taste parameters:

$$(43) \quad \begin{bmatrix} \tau_1^1 & \tau_2^1 \\ \tau_1^2 & \tau_2^2 \end{bmatrix} = \begin{bmatrix} 0 & -0.6 \\ -0.3 & 0 \end{bmatrix}$$

We denote the  $\tau^{\text{th}}$  percentile of the distribution for destination  $k$  by  $\rho_k^\tau$  (i.e.,  $\rho_1^{0.5} = 0$  and  $\rho_1^{0.5} = 0.35$ ). The 75<sup>th</sup> percentiles of these distributions are given by  $\rho_1^{0.75} = 0.674$  and  $\rho_1^{0.75} = 1.024$ , respectively. We simulate the behavior of either  $N = 2,500$  or  $N = 10,000$  individuals per origin location. We then use the (true) assumption of commonality of the unconditional distributions irrespective of the origin location to form minimum distance criterion functions at 100 values of wages where there is significant overlap across all of the conditional wage distributions. Figure 1 describes the four conditional distributions taken from the first of 1,000 Monte Carlo simulations. Dashed lines indicate the region over which there is significant overlap to form moment conditions.<sup>10</sup> Table 1 summarizes the results of 1,000 Monte Carlo simulation runs. The bias is low for every parameter, even when  $N = 2,500$ . The most notable advantage of the larger sample size  $N = 10,000$  is a reduction in mean squared error (e.g., from 0.124 to 0.001 for the estimates of  $\tau_2^1$ ).

The next set of simulations illustrates what happens when we increase the dimension of the choice set from two to three. We keep the same distributions for destination locations 1 and 2, but introduce a third destination:

$$(44) \quad f_1(\omega_{i,1}^j) \sim N(0, 1) \quad f_2(\omega_{i,2}^j) \sim N(0.35, 1) \quad f_3(\omega_{i,3}^j) \sim N(0.75, 1)$$

---

<sup>10</sup> In determining this region, we take the product of the four conditional distributions at each point in their support. We then normalize that product so that its maximum value is 1, and then take all values of wages where the normalized product of the conditional distributions is greater than 0.1.

We similarly augment the matrix of taste parameters:

$$(45) \quad \begin{bmatrix} \tau_1^1 & \tau_2^1 & \tau_3^1 \\ \tau_1^2 & \tau_2^2 & \tau_3^2 \\ \tau_1^3 & \tau_2^3 & \tau_3^3 \end{bmatrix} = \begin{bmatrix} 0 & -0.6 & -0.8 \\ -0.3 & 0 & -0.5 \\ -0.1 & -0.25 & 0 \end{bmatrix}$$

The results of 1,000 Monte Carlo simulations are summarized in Table 2. Not surprisingly, increasing the dimension of the choice set without increasing the sample size reduces the precision of the estimates, both in terms of bias and mean squared error. This is, in large part, a result of a reduction in the size of the region of  $t$  over which all conditional distributions overlap. In most cases, increasing the sample size to  $N = 10,000$  reduces the bias, and in every case it dramatically reduces the MSE.

In the third set of simulations, we demonstrate what happens when the data generating process does not satisfy global commonality. In particular, we draw wages from the following set of birth-and-destination-specific distributions:

$$(46) \quad \begin{array}{lll} f_1^1(\omega_{i,1}^1) \sim N(0, 1) & f_2^1(\omega_{i,2}^1) \sim N(0.1, 1) & f_3^1(\omega_{i,3}^1) \sim N(0.75, 1) \\ f_1^2(\omega_{i,1}^2) \sim N(0, 1) & f_2^2(\omega_{i,2}^2) \sim N(0.35, 1) & f_3^2(\omega_{i,3}^2) \sim N(0.4, 1) \\ f_1^3(\omega_{i,1}^3) \sim N(-0.1, 1) & f_2^3(\omega_{i,2}^3) \sim N(0.35, 1) & f_3^3(\omega_{i,3}^3) \sim N(0.75, 1) \end{array}$$

For destination #1, unconditional wage distributions are now common for origins #1 and #2. For destination #2, they are common for origins #2 and #3. For destination #3, they are common for origins #1 and #3. The Monte Carlo simulations assume (incorrectly) that commonality holds across all origins.

We use a new matrix of taste parameters:

$$(47) \quad \begin{bmatrix} \tau_1^1 & \tau_2^1 & \tau_3^1 \\ \tau_1^2 & \tau_2^2 & \tau_3^2 \\ \tau_1^3 & \tau_2^3 & \tau_3^3 \end{bmatrix} = \begin{bmatrix} 0 & -0.6 & -0.8 \\ -0.3 & 0 & -0.8 \\ -0.3 & -0.6 & 0 \end{bmatrix}$$

I.e., rather than global commonality in wages, we instead impose “taste commonality” on the data generating process – conditional upon being a migrant, the non-pecuniary payoff of locating in a particular destination does not depend upon an individual’s origin.

Results that incorrectly assume global commonality in wages are described in Table 3. While the model does a reasonably good job at recovering the unconditional wage distributions, there are large biases in the estimates of the taste parameters. Table 4 describes the results of another set of Monte Carlo experiments that assume that the researcher knows the exact pattern of commonality between the destination regions. In particular, criterion functions are only formed for destination #1 using origins #1 and #2, for destination #2 using origins #2 and #3, and for destination #3 using origins #1 and #3. Of course, this leaves us with fewer moments in the minimum distance estimator, and the researcher is therefore required to restrict the model in some way. We assume that the researcher correctly restricts the model to exhibit commonality in tastes, so that we only need to recover  $(\tau_1, \tau_2, \tau_3)$ . The estimator recovers these parameters very precisely – biases and mean squared errors drop to practically zero. This set of Monte Carlo simulations demonstrates that blindly assuming global commonality can indeed lead to biased estimates when it is, in fact, not true. Of course, there are no remaining degrees of freedom with which to test global commonality within our model. However, if the researcher has some outside information describing which unconditional wage distributions are likely to be common (e.g., adjacent states), that information can be used to her advantage.

In the final set of Monte Carlo experiments, we demonstrate how the estimator fails when the independence assumption is violated. Independence is a common assumption in most sorting models, but could be relaxed with the proper data – e.g., observing the same individual in many choice situations, as in Kennan and Walker (2005). The model in this paper is applicable to situations where that is not possible (i.e., cross-sectional data), but in current research we look to extend the techniques developed here to estimating the degree of correlation in wage draws.

Table 5 describes the results of two simulations. Both use 10,000 individuals from each of two origin locations. The first modifies the data generating process to include a correlation of 0.25 between an individual’s wage draws in each location, while the second uses a correlation of 0.50. The bias rises dramatically for all estimates when correlation equals 0.25, relative to its value when the data generating process exhibits independence. The bias continues to rise (roughly proportionally) with further increases in the correlation. This clearly highlights a shortcoming in our estimator (and in most of the literature on spatial sorting) and motivates our further research on the use of panel data.

## 6. Empirical Application: Measuring the Returns to College Education

In order to demonstrate the estimator in an actual empirical setting, we examine a question similar to that posed by Dahl (2001) – i.e., what are the returns to a college education (relative to graduating from high school) before and after controlling for the non-random spatial sorting of workers across the United States? The results of the basic Roy model (1951) suggest that such sorting shifts the means of the (observed) conditional wage distributions up from their (unobserved) unconditional values. Whether spatial sorting increases or reduces the estimated returns to a college education will depend upon whether this shift is proportionally bigger for high school or college educated individuals. If, for example, college educated individuals were more mobile and, hence, more able to migrate in response to favorable idiosyncratic wage draws, we would expect spatial sorting to create an upward bias in the estimated returns to a college education. Whether or not this is the case (and how big is the resulting bias) is an empirical question.

In order to answer that question, we use data extracted from the 1990 US Census 1% microsample, available from the IPUMS ([www.ipums.org](http://www.ipums.org)). Specifically, we consider a sample of 325,084 high school graduates taken from each of four regions of the United States (70,449 from the Northeast, 108,623 from the Midwest, 105,775 from the South, and 40,237 from the West).<sup>11</sup> We also consider a sample of 254,256 college graduates (72,493 from the Northeast, 78,554 from the Midwest, 67,517 from the South, and 35,692 from the West). We use only data describing male household heads less than 35 years of age.<sup>12</sup> For each individual, we observe annual wage and region of birth.<sup>13</sup> Tables 6 (a) and (b) summarize the long-run migration probabilities observed in the data for high school and college graduates, respectively. In particular, each row indicates the birth region while each column indicates the region in which the individual is observed in the 1990 Census. Each entry describes the fraction of individuals originating in the row birth region who are found to be living in the column destination region.

---

<sup>11</sup> Northeast = {CT, ME, MA, NH, RI, VT, NJ, NY, PA}, Midwest = {IL, IN, MI, OH, WI, IA, KS, MN, MO, NE, SD, ND}, South = {DE, DC, FL, GA, MD, NC, SC, VA, WV, AL, KY, MS, TN, AR, LA, OK, TX}, West = {AZ, CO, ID, MT, NV, NM, UT, CA, OR, WA}.

<sup>12</sup> We use only household heads because we assume they are more likely to have made their own geographic location decision, and we use only individuals less than 35 years of age as they are more likely to have recently migrated. Older individuals may have migrated further in the past in response to different wage or amenity distributions.

<sup>13</sup> We calculate hourly wage by taking reported annual income from wages and dividing by average hours worked per week times the number of weeks worked. Number of weeks worked is a categorical variable. We assign values at the middle of each category. We drop any individuals reporting an hourly wage less than \$2 or greater than \$60. The US Census describes both the individual's birth state as well as the PUMA in which he/she was living five years prior. We use birth state to define birth region, which becomes our measure of "origin location", but a similar analysis could be performed using location five years prior as the "origin", leading to a short-run measure of mobility cost.

83% of high-school graduates born in the Northeast are found to be living in the Northeast. The fraction of “stayers” is similarly high for other regions.<sup>14</sup>

Tables 7 (a) and (b) report the estimates of the taste parameters for high school and college graduates, respectively. The most striking aspect of all the estimates (for both education groups) is that they are all negative relative to the diagonal elements (which are normalized to zero) and large in magnitude relative to actual wages – it takes a significantly higher wage draw (between \$1.19 and \$5.05 per hour, depending upon education and region) to induce someone to leave their birth region.<sup>15</sup> The magnitude of this effect is not, however, uniform. High school graduates from the South, for example, face a lower non-wage barrier to migration than their counterparts from other regions. This can be a result of either their enjoying desirable amenities from these other regions, or their facing a low cost to migrating – our model is not able to disentangle these without further parameterizing the taste parameter.<sup>16</sup> The opposite effect can be seen for college educated individuals born in the Northeast when deciding whether to migrate to the Midwest. That strong negative effect is, however, diminished for the same individuals should they consider migrating to the South.

Figure 1 illustrates the observed conditional (solid line) and estimated unconditional (dashed line) wage distributions for high school graduates residing in the Northeast. The latter is recovered by applying the Kaplan-Meier procedure to data on high school graduates born in the Northeast. Given the estimated taste parameters, these individuals are likely to be the least censored, allowing us to recover nearly all of the unconditional distribution.<sup>17</sup> There is a noticeable downward shift in the wage distribution once one has controlled for non-random spatial sorting. Table 8 describes this shift for all origin regions and both education groups, both

---

<sup>14</sup> Note that the fraction of “stayers” would be smaller if we had used a finer geographic division (e.g., states), but would still constitute a clear plurality.

<sup>15</sup> Here we use the term “significantly” in the colloquial sense – the numbers are big relative to the average wage earned by these individuals. In future drafts, formulae for the asymptotic distributions of these estimates will be available to calculate standard errors.

<sup>16</sup> It is also possible that the low estimates for those born in the South reflect a violation of the commonality assumption. If, for example, those born in the South actually received wages in the Northeast from a lower distribution than those born elsewhere, the estimator (assuming commonality) would try to explain why so many individuals had migrated by giving them a low migration cost. In future drafts, we plan to conduct this empirical exercise at a higher spatial frequency (e.g., states). Doing so will allow us to relax the global commonality assumption that we use here (e.g., considering commonality only amongst adjacent states).

<sup>17</sup> For example, we would only ever expect to see college educated individuals from the West region living in the Northeast if they received an especially high wage draw (i.e., big enough to overcome their taste parameter of -4.70). We could not recover the lower tail of the Northeast wage distribution using these individuals. Since we assume commonality in recovering the taste parameters anyway, it is consistent to use only individuals from the best origin location in recovering our Kaplan-Meier estimates. There is, of course, a loss of efficiency from not using data on individuals from other origins, and we are currently working on a Kaplan-Meier procedure that would allow us to incorporate those data.

at the median and the 75<sup>th</sup> percentile of the wage distribution. There is a non-trivial downward shift in all distributions. Table 9 uses the results from Table 8 to calculate the effect of correcting for non-random spatial sorting on the estimated returns to a college education. We calculate those returns at both the median and 75<sup>th</sup> percentile by taking the percentage increase in the wage going from the high school to the college wage distribution. For example, using the wages observed in the raw data for individuals living in the Northeast, the returns to a college education would be  $(19.35-12.51)/12.51 = 54.7\%$ .<sup>18</sup> Controlling for non-random spatial sorting, this return falls to 31.5%. The drop in returns is similar for the Midwest, and smaller for the South and West. This likely reflects the fact that even college graduates from the South and West are relatively immobile (compared to high school graduates), whereas college graduates from the Northeast and Midwest are much more mobile than their high school graduate counterparts, increasing the selection bias. There does not appear to be any clear trend in the size of the selection bias at various points in the wage distribution (i.e., at the median versus the 75<sup>th</sup> percentile), highlighting the importance of the non-parametric approach we take to recovering unconditional wage distributions.

## 7. Conclusion

*[to be added]*

## References

- Ahn, H. and J.L. Powell (1993) "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics*. 58:3-29.
- Bayer, P., N. Keohane, and C. Timmins (2006) "Migration and Hedonic Valuation: The Case of Air Quality." NBER Working Paper.
- Borjas, George (1987) "Self-Selection and the Earnings of Immigrants." *American Economic Review*. 77:531-553.
- Dahl, Gordon (2001) "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets." *Econometrica*. 70(6):2367-2420.
- Davies, Greenwood and Li (2001) "A Conditional Logit Approach to U.S. State-to-State Migration." *Journal of Regional Science*. 41(2):337-360.

---

<sup>18</sup> These returns are similarly in magnitude, but slightly larger than those found in the raw data used by Dahl (2001). We expect these conditional returns to fall slightly as we refine this application and condition on more individual attributes.

Falaris, Evangelos (1987) "A Nested Logit Migration Model with Selectivity." *International Economic Review*. 28:429-43.

Heckman, James (1990) "Varieties of Selection Bias." *American Economic Review*. 80(2):313-18.

Heckman, James and Bo Honore (1990) "The Empirical Content of the Roy Model." *Econometrica*. 58:1121-49.

Kaplan, E.L. and P. Meier (1958) "Nonparametric Estimation from Incomplete Data." *Journal of the American Statistical Association*. 53:457-481.

Kennan, John and James Walker (2005) "The Effect of Expected Income on Individual Migration Decisions." NBER Working Paper 9585.

Roback, Jennifer (1982) "Wages, Rents, and the Quality of Life." *Journal of Political Economy*. 90:1257-78.

Roy, A.D. (1951) "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers*. 3:135-146.

Table 1  
2 x 2 Monte Carlo Simulation Results

	N = 2,500		N = 10,000	
	Bias	MSE	Bias	MSE
$\tau_2^1$	$-7.3 \times 10^{-4}$	0.124	$2.6 \times 10^{-5}$	$7.1 \times 10^{-4}$
$\tau_1^2$	-0.029	0.128	$1.1 \times 10^{-4}$	$7.6 \times 10^{-4}$
$\rho_1^{0.5}$	$-5.6 \times 10^{-4}$	$6.5 \times 10^{-3}$	$-1.7 \times 10^{-4}$	$2.5 \times 10^{-4}$
$\rho_2^{0.5}$	$-3.1 \times 10^{-3}$	$2.5 \times 10^{-3}$	$-5.5 \times 10^{-4}$	$2.1 \times 10^{-4}$
$\rho_1^{0.75}$	$-5.4 \times 10^{-4}$	$2.5 \times 10^{-3}$	$-7.6 \times 10^{-4}$	$2.1 \times 10^{-4}$
$\rho_2^{0.75}$	$-1.9 \times 10^{-3}$	$1.1 \times 10^{-3}$	$-5.3 \times 10^{-4}$	$1.9 \times 10^{-4}$

Table 2  
3 x 3 Monte Carlo Simulation Results

	N = 2,500		N = 10,000	
	Bias	MSE	Bias	MSE
$\tau_2^1$	-0.004	0.053	$-7.4 \times 10^{-3}$	0.013
$\tau_3^1$	0.036	0.086	$5.3 \times 10^{-3}$	0.014
$\tau_1^2$	0.002	0.055	$4.3 \times 10^{-3}$	$6.8 \times 10^{-3}$
$\tau_3^2$	0.042	0.085	0.015	0.021
$\tau_1^3$	-0.099	0.264	-0.026	0.068
$\tau_2^3$	-0.097	0.270	-0.029	0.076
$\rho_1^{0.5}$	$2.4 \times 10^{-4}$	0.021	$-2.1 \times 10^{-3}$	$5.3 \times 10^{-3}$
$\rho_2^{0.5}$	$4.6 \times 10^{-3}$	0.017	$3.9 \times 10^{-3}$	$2.0 \times 10^{-3}$
$\rho_3^{0.5}$	-0.020	0.010	$-5.1 \times 10^{-3}$	$2.3 \times 10^{-3}$
$\rho_1^{0.75}$	$3.9 \times 10^{-3}$	$2.8 \times 10^{-3}$	$1.0 \times 10^{-4}$	$4.9 \times 10^{-4}$
$\rho_2^{0.75}$	$7.1 \times 10^{-3}$	$3.5 \times 10^{-3}$	$2.4 \times 10^{-3}$	$9.3 \times 10^{-4}$
$\rho_3^{0.75}$	$-4.7 \times 10^{-3}$	$3.9 \times 10^{-3}$	$-1.4 \times 10^{-3}$	$5.7 \times 10^{-4}$

Table 3  
 3 x 3 Monte Carlo Simulation Results  
 Violation of Global Commonality, N = 10,000

	Bias	MSE
$\tau_2^1$	-0.224	0.056
$\tau_3^1$	-0.069	$8.6 \times 10^{-3}$
$\tau_1^2$	-0.063	0.016
$\tau_3^2$	-0.259	0.074
$\tau_1^3$	-0.080	0.017
$\tau_2^3$	0.023	0.010
$\rho_1^{0.5}$	-0.082	$8.4 \times 10^{-3}$
$\rho_2^{0.5}$	-0.054	$3.8 \times 10^{-3}$
$\rho_3^{0.5}$	$-4.4 \times 10^{-3}$	$3.4 \times 10^{-4}$
$\rho_1^{0.75}$	-0.026	$1.1 \times 10^{-3}$
$\rho_2^{0.75}$	-0.015	$4.4 \times 10^{-4}$
$\rho_3^{0.75}$	$-1.2 \times 10^{-3}$	$2.1 \times 10^{-4}$

Table 4  
 3 x 3 Monte Carlo Simulation Results  
 Using Pairwise Commonality and Correctly Assuming  
 Global “Taste Commonality”, N = 10,000

	Bias	MSE
$\tau_1$	$-5.2 \times 10^{-3}$	$7.3 \times 10^{-4}$
$\tau_2$	$1.3 \times 10^{-3}$	$7.7 \times 10^{-4}$
$\tau_3$	$2.1 \times 10^{-3}$	$5.9 \times 10^{-4}$
$\rho_1^{0.5}$	$8.9 \times 10^{-4}$	$4.2 \times 10^{-4}$
$\rho_2^{0.5}$	$-9.9 \times 10^{-4}$	$2.5 \times 10^{-4}$
$\rho_3^{0.5}$	$-7.5 \times 10^{-4}$	$2.1 \times 10^{-4}$
$\rho_1^{0.75}$	$8.3 \times 10^{-5}$	$2.5 \times 10^{-4}$
$\rho_2^{0.75}$	$-4.9 \times 10^{-4}$	$2.0 \times 10^{-4}$
$\rho_3^{0.75}$	$-4.7 \times 10^{-4}$	$2.0 \times 10^{-4}$

Table 5  
 2 x 2 Monte Carlo Simulation Results  
 Violation of Independence Assumption, N = 10,000

	Correlation = 0.25		Correlation = 0.5	
	Bias	MSE	Bias	MSE
$\tau_2^1$	-0.077	0.023	-0.200	0.070
$\tau_1^2$	-0.117	0.052	-0.276	0.126
$\rho_1^{0.5}$	-0.101	0.011	-0.210	0.046
$\rho_2^{0.5}$	-0.062	0.004	-0.112	0.013
$\rho_1^{0.75}$	-0.050	$3.0 \times 10^{-3}$	-0.108	0.012
$\rho_2^{0.75}$	-0.027	$1.3 \times 10^{-3}$	-0.052	$2.9 \times 10^{-3}$

Table 6  
 Mobility Matrices, 1990 US Census Data  
 Fraction of those originating in birth region living in destination region

(a) High School Graduates

		Destination Region			
		Northeast	Midwest	South	West
Birth Region	Northeast	0.83	0.02	0.10	0.04
	Midwest	0.01	0.82	0.09	0.07
	South	0.03	0.07	0.86	0.05
	West	0.01	0.05	0.08	0.86

(b) College Graduates

		Destination Region			
		Northeast	Midwest	South	West
Birth Region	Northeast	0.65	0.06	0.19	0.10
	Midwest	0.05	0.62	0.17	0.16
	South	0.05	0.06	0.80	0.09
	West	0.03	0.05	0.11	0.80

Table 7  
Taste Parameter Estimates

(a) High School Graduates

		Destination Region			
		Northeast	Midwest	South	West
Birth Region	Northeast	0	-3.51	-3.37	-3.38
	Midwest	-3.48	0	-3.35	-3.28
	South	-1.37	-1.19	0	-1.20
	West	-3.47	-3.10	-3.34	0

(b) College Graduates

		Destination Region			
		Northeast	Midwest	South	West
Birth Region	Northeast	0	-5.05	-4.74	-4.96
	Midwest	-4.83	0	-4.56	-4.55
	South	-4.49	-4.27	0	-4.41
	West	-4.70	-4.48	-4.38	0

Table 8  
Wages by Education and Region, 1990 US Census  
Raw Data and Corrected for Spatial Selection

	High School				College			
	Median		75 <sup>th</sup> Percentile		Median		75 <sup>th</sup> Percentile	
	Raw Data	Selection Corrected	Raw Data	Selection Corrected	Raw Data	Selection Corrected	Raw Data	Selection Corrected
Northeast	12.51	10.44	16.66	14.53	19.35	13.73	26.07	19.62
Midwest	11.44	9.32	15.69	13.11	16.35	11.28	22.36	16.91
South	10.30	8.82	14.71	12.37	16.35	12.75	22.87	18.05
West	11.78	9.81	16.40	14.47	17.69	13.73	24.55	19.61

Table 9  
Percentage Returns to College Education

	Median		75 <sup>th</sup> Percentile	
	Raw Data	Selection Corrected	Raw Data	Selection Corrected
Northeast	54.7	31.5	56.5	35.0
Midwest	42.9	21.0	42.5	29.0
South	58.7	44.6	55.5	45.9
West	50.2	40.0	49.7	35.5

Figure 1  
2 x 2 Monte Carlo Simulation  
Sample Estimation Region

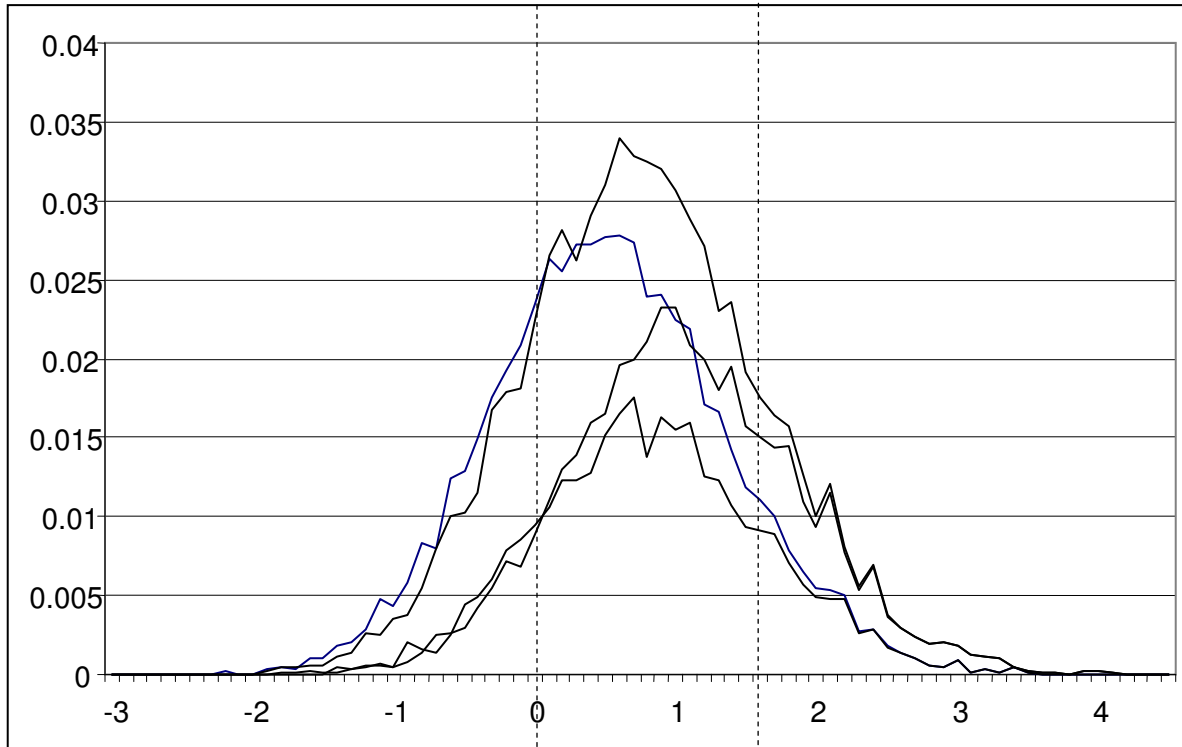


Figure 2  
High School Graduates (Northeast)

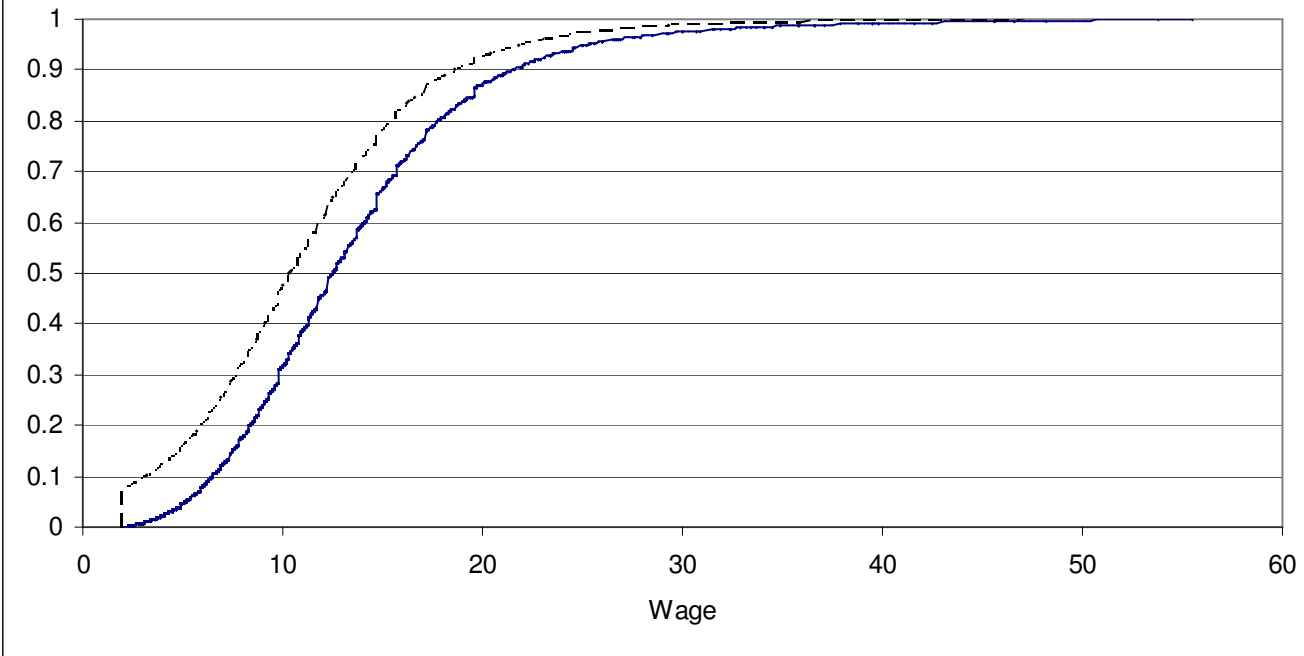


Figure 3  
College Graduates (Northeast)

