SELF-SELECTION INTO LONG-DISTANCE COMMUTING ON EARNINGS AND LATENT CHARACTERISTICS

Sergii Troshchenkov

University of Milan

GRAPE (Group for Research in APplied Economics)

sergii.troshchenkov@unimi.it

Olle Westerlund

Department of Economics

Umeå University, Sweden

olle.westerlund@econ.umu.se

JEL codes: R 40, J61, J24

1 Introduction

Long distance commuting to work (LDC) has become a significant phenomenon in most developed countries, partially as a substitute for migration (Sultana & Weber, 2007; Sandow & Westin, 2010a). Previous studies have demonstrated that commuting options positively affect the job matching process, mitigate regional disparities, and satisfy labor demand of growing agglomerations and "grease wheels" of local economies (Edwardsson, 2000; Hazans, 2004; Lundholm, 2010). There is substantial literature on various socioeconomic and health outcomes of commuting (Koslowsky et al., 1996; Sandow, 2008 Gottholseder et al., 2009; Hansson et al., 2011; Lingren et al., 2014). A small but growing literature also study the effects of commuting on income and labor supply (Rouwendahl, 1998; Rouwendahl, 1999; Manning, 2003; Rouwendal, 2004; Mulalic & Pilegaard 2010; Rupert et al., 2012).

Findings show that migrants as well as long distance commuters are non-randomly drawn from the total population and also from the total work force (e.g. Greenwood, 1985; Tunali, 2000; Eliasson et al., 2003). Increased availability of highly informative longitudinal micro data sets has improved the ability of researchers to control for confounders, but unmeasured characteristics of individuals may still lead to misleading results. Latent characteristics associated with selection into LDC and correlated with the outcome of interest is (e.g. earnings) is likely to cause bias in estimated effects. Therefore, studies on observational data usually use various econometric techniques to address this problem. In most cases, the issues of main interest are whether and how selectivity on unobserved characteristics into the group of migrants or commuters emerges and how this is related to ex-post income or other outcomes. Research on urban commuting indicates substantial non-random selectivity in the residential market, i.e. self-selectivity in observed residential location (Wasmer & Zenou 2002; Murata & Thisse, 2005; Wasmer & Zenou 2006; Borck et al., 2008;. Rupert & Wasmer, 2009) .

Knowledge on the nature of selectivity is important for interpretation of empirical results. This study contributes to the previous literature on selection into commuting in three ways. First, the main interest lies with selection on latent characteristics correlated with pre-LDC earnings instead of earnings after commuting. The rationale for this is that the decision to engage in LDC is taken concurrently or before commuting is initiated. Second, two potential dimensions of selection are considered: 1) one based on unmeasured traits associated with pre-LDC earnings, 2)

one based on measured earnings before start of long-distance commuting. Following Heckman (1979), self-selection is based on unmeasured traits of commuters. The selection on unobserved characteristics is positive when individuals who exhibit latent propensities to engage in commuting are characterized by unobserved attributes that result in higher earnings. By contrast, negative self-selection is present when latent characteristics are associated with unexpected higher probability of LDC and lower earnings conditional on observed attributes, (e.g. Ihlanfelt, 1988, Zaiceva, 2006).

Another line of research has studied selection in terms of measured earnings previous to mobility in terms of commuting or migration; i.e positive selection means that mobile workers come from the upper part of the ex-ante earnings distribution. Negative selection occurs if long distance commuters (or migrants) are predominantly drawn from the lower part of the earnings distribution (e.g Öhman et al., 2003; Gabriel and Schmitz, 1995; Finnie, 2001). In the present study, both these aspects of selection are captured within a single model allowing testable hypotheses about selection based on both observed earnings and unmeasured attributes of potential long-distance commuters. Another contribution is that we use highly informative longitudinal population register data covering the whole population in Sweden. It includes detailed socioeconomic information and geocoding in 100 meter squares of individual's workplaces and places of residence.

Our main results indicate that men who commute over longer distances are negatively selected from the ex-ante earnings distribution, i.e. LDC:s tend to have lower earnings than expected conditional on *observed* characteristics as measured the year before engaging in long-distance commuting. But at the same time, our findings indicate that individuals with *unobserved* traits associated with higher earnings are also more likely to engage in long-distance commuting. A possible interpretation is that the lower earnings before commuting reflects labour market mismatch, and unobserved traits associated with higher incomes also correlate with higher expected return of spatially extended job-search. The results for females are similar, although the estimated negative selection on earnings before start of LDC is not statistically significant different from zero.

The next section presents the econometric model and the parameters of main interest. Section 3 presents data, model specifications and descriptive statistics. Section 4 gives the results followed by summary and conclusions in Section 5.

2 Earnings, commuting and self-selectivity

Following the modelling of migration in Tunali (2000) and Nakosteen et al. (2008), self-selection into long-distance commuting is assumed to be based on observed and unobserved (latent) characteristics. These may affect earnings as well as the probability of long-distance commuting. Here, for a sample of employees, LDC is defined as a change of work place involving transition from short-distance commute to long-distance commuting. Similar to most studies on migration, long-distance commuting is measured as a dichotomous outcome based on an explicit criteria, here a commuting distance of at least 50 km. This will be discussed further in Section 4.

Individuals are observed at consecutive points in time. Selectivity is assumed to be manifested through two sources. One is observations of individual's earnings at the time the decision to start long-distance commuting or not is taken (first period). The other is the correlation between unobserved heterogeneity affecting first period earnings and unobserved heterogeneity affecting the probability of engage in long-distance commuting in the second period.

In the first period individuals are employed at an initial work place in a specified initial geographical location and consider employment options at other work places. Earnings of individual *i* in the first period are:

$$y_i = \beta' x_i + \varepsilon_i \tag{1}$$

where y_i denotes earnings, x_i is a vector of explanatory variables, and β is a vector of unknown coefficients to be estimated. The random error term ε is normally distributed with zero mean and variance σ_ε^2.

In the second period, the individual chose to become a long-distance commuter or not. The two outcomes are c_i = 1 if long-distance commuting is chosen, and c_i = 0, otherwise.

Back in the first period, the individual evaluates future employment options and the expected income during the second period is:

$$y_i' = y_i + \omega_i \tag{2}$$

The term ω_i is adjustment for influence of latent characteristics, from the individual's point of view representing the expected increase in earnings.

The self-selection mechanism can be expressed in terms of the difference in expected outcomes of alternative choices, here the difference between expected earnings if long-distance commuting is chosen, and expected earnings in the alternative case :

$$E(y_i'|c_i = 1) - E(y_i'|c_i = 0) = E(\omega_i|c_i = 1) - E(\omega_i|c_i = 0) \tag{3}$$

Individuals who start long-distance commuting are then self-selected on unobserved characteristics that influence expected earnings as evaluated in the first period. For example, unobserved ability may be associated with systematic "positive" or "negative selection", depending on the level of ability of long distance commuters. A "positive" self-selection on unobserved traits may be due to high ability associated with high earnings in any location of work places, but also associated with additional earnings premium in case of accepting a job offer involving long-distance commuting. A "negative" selection could stem from low ability associated with low earnings at any work place but at the same time combined with relative high expected pay off if accepting long-commutes,

Long-distance commuters are also systematically self-selected on observed attributes traits measured in available data. The nature of selectivity is an empirical question. High earnings in the first period may signal good job matches, high opportunity costs for jobs search and therefore low job search intensity and low incentives for job mobility. The probability of selecting long-distance commuting would then be negatively associated with earnings. On the other hand, high

earnings could be associated with specific skills to be matched to few job openings on regionally "thin" labor markets (e.g. Manning, 2003). Long-distance commuters to a new work place can therefore be positively selected with respect to observed earnings in the first period.

The econometric model to be estimated considers systematic self-selection on observable characteristics as well as on unobserved heterogeneity. Let c_i^* indicate the latent propensity of individual's to engage in long-distance commuting. This option is chosen when $c_i^* > 0$. The joint model of earnings (eq. 1) and LDC is

$$\begin{aligned} y_i &= \beta' x_i + \varepsilon_i \\ c_i^* &= \alpha y_i + \delta' z + \omega_i \end{aligned} \tag{4}$$

where z_i is a vector of measured characteristics, δ a vector of coefficient parameters , and ω_i is the individual specific error term in the equation for expected second period earnings (eq. 2). [1] All variables with the exception of LDC status are measured in the initial period assuming that potential self-selection is reflected by the decision to start LDC and not necessarily in subsequent events.

A dichotomous variable (c_i) indicate whether the individual is observed as a long-distance commuter in the second period or not, and relate to the unobserved latent propensity for LDC (c_i*) as:

$$c_i = \begin{cases} 1 \; if \; c_i^* < 0 \\ 0 \; if \; c_i^* \geq 0 \end{cases} \tag{5}$$

The error terms ε_i and ω_i are assumed to be bivariate normally distributed with zero means, variances σ^2 and 1, respectively, and covariance $\sigma_{\varepsilon\omega}$.

The hypotheses of self-selectivity on pre-commuting earnings can be tested through estimates of α in eq 3. Conditional on other characteristics, long distance commuters represent a positive

[1] Following the adoption of Hausman and Wise (1979) used in Nakosteen et al. (2008).

selection on measured earnings if $\alpha > 0$ and a negative selection if $\alpha < 0$. Self-selectivity on unobserved heterogeneity is reflected by the estimated covariance $\sigma_{\varepsilon\omega}$. Higher ability or other unobserved traits associated with higher earnings may also be associated with higher propensity for LDC, i.e. $\sigma_{\varepsilon\omega} > 0$,an indication of a positive selection on unobserved traits. A negative covariance reflects negative selection due to latent characteristics from the conditional earnings distribution (initial period) into LDC.

Any combination of positive or negative selection indicated by the two parameters are possible, e.g. a negative selection on initial period earnings does not rule out a positive or a negative selection on unobserved traits. For example, individuals in the lower part of the earnings distribution may have unobserved traits reflected in unexpected high earnings (conditional on observed characteristics) and at the same time unexplained high propensity for LDC.

Equations (1) and (3) are estimated jointly stating the model as two reduced form equations. The reduced form commuting equation is obtained by substitution for y_i in eq (3)

$$c_i^* = \alpha\beta' X_i + \delta' z_i + \widetilde{\omega}_i \tag{6}$$

where

$$\widetilde{\omega}_i = \alpha\varepsilon_i + \omega_i$$

The error terms $\widetilde{\omega}_i$ and ε_i are assumed to follow a bivariate normal distribution with zero means and covariance Cov $(\varepsilon_i, \widetilde{\omega}_i) = \alpha\sigma_\varepsilon^2 + \sigma_{\varepsilon\omega}$

Let $g(\varepsilon_i)$ denote the unconditional density of the error term in equation (1) for income in the decision period prior to realization of LDC or not (a function of $y_i - \beta' x_i$), and $f(\widetilde{\omega}_i)$ the conditional density function of long-distance commuting (a function of $\alpha, \beta' X_i, \delta' z_i, \sigma_{\varepsilon,}\sigma_{\varepsilon w,}$).[2] The likelihood function for the sample of *C* long-distance commuters and *N* individuals in the non-LDC group is:

$$L = \prod_{i=1}^{C} f(\widetilde{\omega}_i)|\varepsilon_i, c_i = 1) \cdot g(\varepsilon_i) \cdot \prod_{i=1}^{N} f(\widetilde{\omega}_i)|\varepsilon_i, c_i = 0) \cdot g(\varepsilon_i) \tag{7}$$

[2] See Nakosteen et al 2008, p 773 and 774 for details.

Estimates of equations (1) and (3) together with the variance and covariance parameters are obtained by maximising *L*.

As stated previously, the present model of selection into long-distance commuting on observed and unobserved characteristics is an application of the migration model in Nakosteen et al. (2008) which in turn is a partial adaption of Tunali (2000). A basic presumption is that individuals self-select for mobility based on expected outcomes in terms of earnings. These expectations are partially dependent on latent attributes unobserved of the researcher. In Tunali's model, selection on unobserved heterogeneity is present if, conditional on migrant status, the means of those attributes differ between movers and stayers. Here, self-selection on unobserved traits is captured by the covariance between random error terms in equations (1) and (3), which carry information on endogenous selection in addition to the selection on measured earnings.

3 Data and empirical model

We use longitudinal data from Swedish population registers administered by Statistics Sweden. Apart from the precise geocoded information on place of residence and work place, data provide detailed information on the individual's personal characteristics and labour market outcomes.

The main interest of this study lies with the determinants of spatial labor mobility in terms of changing location of individual's employment. Commuting distances of the stock of employed convey very limited information on labour mobility in terms of *changing* spatial allocation of labour supply. The overwhelming share of the total stock of employed (commuters) does not change work place from one year to another. Therefore, we sample from the inflow of new employees to all work places in Sweden. The sample consists of individuals who in 2007 were of age 20-64, employed or registered as unemployed in 2007, and who became employed at a new work place in 2008. It represents the lion share of total flow of external hirings to all work places in the economy as measured between two time points, November 2007 and November 2008.[3] Individuals who change location of work places within the same firm are also included. Cases with missing information on their place of residence in 2007 or 2008 and cases with missing information on work place in 2008 are excluded. Students and individuals staying at home with parental benefits are also excluded because of uncertainties regarding labour force status and identification of change of work place.[4] The sample includes 392 818 individuals, 206 281 men and 186 537 women.

Long distance commuting (LDC) is defined as commuting over 50 km between place of residence and work place. Distance is measured as Euclidian distance based on co-ordinates (100 square meters precision). Individuals starting long-distance commutes amounted to 17 449 men (7.7 percent) and 10 580 women (5.3 percent).

In the earnings equation we control for age, education, foreign citizenship, sector of employment, and median of earnings in the region of residence. Covariates in the commuting equation measure individual's earnings in the first period (2007), marital status, age, education,

[3] It represents an understatement of total matches because multiple changes of individuals work place is not observed.

[4] We also excluded individuals with commuting distances over 50 km in 2007, because of uncertainties regarding discrimination between new and old work places. Robustness checks using a sample without these restrictions does not affect our main results and our conclusions. Results available on request from the authors.

foreign citizenship, and distance weighted access to jobs. Descriptive statistics by gender and commuting status are presented in Table 1.

Table 1.Sample means and standard deviations, long-distance commuters and short-distance commuters.

Variable	*Males*		*Females*	
	LDC	*Non-LDC*	*LDC*	*Non-LDC*
Log Earnings	*7.94*	*7.957*	*7.745*	*7.756*
	(0.555)	*(0.646)*	*(0.551)*	*(0.643)*
LDC	*0.077*		*0.053*	
	(0.267)		*(0.220)*	
Age	*42.312*	*40.476*	*44.952*	*41.397*
	(12.916)	*(12.814)*	*(12.452)*	*(13.011)*
Age squared	*1957.214*	*1802.561*	*2175.756*	*1883.033*
	(1095.073)	*(1071.235)*	*(1070.456)*	*(1089.432)*
Foreign	*0.122*	*0.093*	*0.125*	*0.099*
	(0.327)	*(0.291)*	*(0.331)*	*(0.299)*
Accessibility	*369123.2*	*245467.9*	*392355.1*	*270879.5*
	(451154.3)	*(359408.3)*	*(452774.9)*	*(379375.4)*
Post-secondary education <2	*0.066*	*0.074*	*0.041*	*0.051*
	(0.249)	*(0.262)*	*(0.199)*	*(0.221)*
Post-secondary education >2	*0.231*	*0.278*	*0.384*	*0.426*
	(0.421)	*(0.448)*	*(0.486)*	*(0.494)*
MSc or PhD	*0.015*	*0.021*	*0.009*	*0.021*
	(0.122)	*(0.143)*	*(0.098)*	*(0.146)*
Married	*0.405*	*0.376*	*0.461*	*0.358*
	(0.49)	*(0.484)*	*(0.498)*	*(0.479)*
Single mother/father	*0.029*	*0.027*	*0.1*	*0.085*
	(0.17)	*(0.162)*	*(0.3)*	*(0.279)*
Living with parents	*0.069*	*0.105*	*0.032*	*0.079*
	(0.253)	*(0.306)*	*(0.176)*	*(0.271)*
Children	*0.052*	*0.05*	*0.064*	*0.0526*
	(0.223)	*(0.218)*	*(0.244)*	*(0.223)*
Manufacture	*0.243*	*0.14*	*0.08*	*0.063*
	(0.429)	*(0.347)*	*(0.271)*	*(0.243)*
Construction	*0.105*	*0.121*	*0.014*	*0.015*
	(0.307)	*(0.327)*	*(0.118)*	*(0.122)*
Retail	*0.22*	*0.259*	*0.159*	*0.214*
	(0.414)	*(0.438)*	*(0.366)*	*(0.41)*
Private services	*0.086*	*0.091*	*0.056*	*0.073*
	(0.281)	*(0.288)*	*(0.23)*	*(0.261)*
Log Median of earnings in LA	*7.967*	*7.965*	*7.97*	*7.966*
	(0.052)	*(0.054)*	*(0.051)*	*(0.055)*
n	*17449*	*207863*	*10580*	*187837*

Comparison of the means of previous earnings between LDC:s and non-LDC:s suggest a selection into LDC from the upper part of the unconditional earnings distribution. In line with expectations, Table 1 indicates that earnings in the previous year are higher for non-commuters than commuters.

The labor force participation rate of females in Sweden is high and previous studies suggest different commuting and earnings patterns for males and females (Albrecht et al., 2001; Sandow 2008; Lundholm, 2010; Sandow & Westin, 2010a). We therefore estimate the earnings and commuting equations separately by gender. Earnings and commuting are assumed to be determined by individual and regional characteristics. The variables in the earnings equation includes age, educational attainment, sector of employment, nationality and regional wage level. The age variable capture individual experience, productivity and life course effects. Educational dummy variables are additional indicators of human capital affecting earnings and pay-off from commuting. The reference category is educational attainment of secondary school of three years or less. Variation in regional wage levels are controlled for by a variable measuring the median of earnings in the local labour market area where the work place is located. A set of dummy variables captures earnings differences by sector of employment. Nationality is a dummy variable indicating individuals of Swedish origin. Table 2 gives the specifications of the empirical counterparts to equations 1 and 3.

Table 2. Covariates in the earnings and commuting equations

Earnings equation	*Commuting equation*
Nationality	*Nationality*
Age	*Age*
	Age squared
Age squared	*Accessibility*
Education	*Education*
	Family status
	Presence of children
	Previous earning
Regional wage level	*Regional wage level*
Sector of employment	*Sector of employment*

The commuting equation includes covariates measuring previous earnings, age, education, marital status, presence of children, sector of employment, nationality, regional wage level, and

regional accessibility to jobs. Previous studies show systematic influence of age, education, marital status and family characteristics on commuting. (Bartel & Lichtenberg, 1987; Borsch-Supan, 1990; van Ham et al., 2001; Sandow & Westin, 2010b; Lingren et al., 2014). Regional labour market tightness and regions attractiveness for commuters is captured by the median wage level in the region where the workplace is located. Conditional on place of residence (ex-ante), spatial accessibility to jobs will affect the probability of long distance commuting. Following Eliasson et al. (2003), the accessibility measure was defined as a discounted sum of all jobs discounted by the distance between population centers of labor markets and individuals place of residence.[5] Systematic differences in commuting distances by industry are captured by a set of dummy variables indicating individuals sector of employment.

4 Results

The joint estimation of the earnings and commuting equation is carried out using maximum likelihood (MLE). The parameters of main interest are α as indicator of selectivity on (observed) earnings, and the covariance σ_{ew}. The latter indicating association between unobserved heterogeneity that correlates with earnings and unobserved heterogeneity correlating with the probability of long-distance commuting.

Males

The estimation results for males are given in Table 3. The earnings equation estimates are (qualitative) in line with expectations. Age, education and the regional wage level are positively correlated with earnings, the indicated concave age/earnings profile is also as expected. The point estimates suggest that earnings increase by age up to a turning point at about 50 years of age. The results also signal a significant premium of education. Relative to the baseline category of individuals with an educational attainment at the secondary level, the increase in earnings ranges from 21% for individuals with tertiary level education shorter that two years and up to 65% for males with a Master or PhD degree. Individuals of non-Swedish origin have lower estimated earnings, about -25% relatively to Swedish natives.

[5] Access is measured as $\sum Ejdij^{-\alpha}$ where E_j is all jobs in region *j* and$dij^{-\alpha}$ is the distance decay function with distance measured as distance between labour market region of residence (*i*) and labour market region of work place and α is the distance decay parameter.

The coefficient estimates of the commuting equation indicate a negative and statistically significant selection into long-distance commuting on earnings ($\hat{\alpha}$ = - 0,6345, $|t|$= 6,73). The estimate of the covariance parameter σ_{ew} is positive and significant ($\hat{\sigma}_{\varepsilon\omega}$=0,8753, $|t|$= 8,27). Thus, while the LDC:s are systematically selected from the lower part of the (unconditional) earnings distribution, there is an indication of a positive selection into long-distance commuting on latent characteristics affecting earnings (a positive selection from the conditional earnings distribution).

Table 2. MLE-results for the earning and commuting equations. Sample of males. No exclusion restrictions on sample selection.

Variables	*Earning equation*		*Commuting equation*	
	Coefficient	*Std. error*	*Coefficient*	*Std. error*
Previous earning (α)			*-0,6345****	*0,0943*
Age	*0,0591****	*0,0012*	*0,0530****	*0,0075*
Squared Age	*-0,0006****	*0,00001*	*-0,0006****	*0,00008*
Nationality	*-0,2553****	*0,0074*	*-0,3352****	*0,035*
Post-gymnasium level of education<2	*0,2150****	*0,0088*	*0,4243****	*0,0354*
Post-gymnasium level of education>2	*0,3198****	*0,0057*	*0,5556****	*0,034*
University level of education	*0,6526****	*0,0165*	*0,9651****	*0,0803*
Married			*-0,0262*	*0,018*
Single mother/father			*-0,0063*	*0,0453*
Living with parents			*0,355****	*0,0266*
Children			*-0,0011*	*0,0334*
Regional wage	*1,2079****	*0,0482*	*4,6128****	*0,2808*
Manufacture	*0,2014****	*0,0059*	*-0,5228****	*0,0267*
Construction	*0,1422****	*0,0066*	*0,1155****	*0,0267*
Retail	*0,0556***	*0,0066*	*0,1486****	*0,0201*
Private services	*0,2054****	*0,0121*	*0,091****	*0,0292*
Accessibility			*-1,13e-06****	*3,39e-08*
Constant	*-3,088****	*0,0165*	*-35,0406****	*2,0934*
Sigma (σ_{ew})	*0,8753**** 0,1059*			
Number of observations	*225312*			

Asterisks indicate significance level
Significance level: ""p<0.05, "**"p<0.01, "***" p<0.001*
The standard errors are heteroscedasticity robust and clustered at the individual level

Coefficients on linear and squared terms of age suggest a concave profile of age with an estimated turning point between 45 and 50 years.[6] The probability of commuting increases also with the level of education. Presence of a spouse/partner and/or children decreases probability of long-distance commuting in comparison to single men. Curiously enough, individuals living with parents tend to be more mobile than single individuals without children.[7] Moreover, presence of children is not associated with lower probability of LDC among men. Sector of new job correlates with the probability of commuting, possibly reflecting variation in spatial workplace distribution across sectors of employment. The reference category is individuals who received a job in the public sector. The estimated suggest higher probability of LDC for employed in the other sectors except manufacturing.

Turning to the covariates measuring regional attributes, the estimates are indicative of a positive association between regional wages and commuting. A higher regional wage level in the work place region increases the attractiveness of work places as commuting destinations. Accessibility is (unexpectedly) negatively associated with LDC. One possible explanation is that the attractiveness of neighboring labor markets is offset by the spatial distance, or that regional access is higher in densely populated areas with higher density of jobs within shorter commuting distances.

Females

The estimates for females (Table 3) indicate similar selectivity on observed and latent characteristics as for males.

[6] This is substantially different from the age profile for the stock of commuters between functional labour market regions in Sweden as reported in Eliasson et al, 2007, Diagram 4.17, p 160. Descriptive averages indicate maximum frequency of LDC at ages in the early 20:s (about 11 %) and then a monotonic decreasing frequency of commuting down to about 7% at ages 50-55. This may reflect that sampling from the flow into employment and commuting may yield quite different samples than sampling from the observed stock of commuters.

[7] This may be due to a measurement error where observations of singles may in fact be cohabitants without children, or a cohabitating parent living with partners who are not registered as a parent to children living in the household. Cohabitants are registered as cohabitants only if cohabitating adults have at least one child in common in the household.

Table 3. MLE-results for the earning and commuting equations. Sample of females. No exclusion restrictionson sample selection.

Variables	*Earning equation*		*Commuting equation*	
	Coefficient	*Std. error*	*Coefficient*	*Std. error*
Previous earning (α)			*-0,4388****	*0,1184*
Age	*0,0379****	*0,0012*	*0,0119*	*0,0078*
Squared Age	*-0,0003****	*0,00001*	*-0,0002***	*0,00008*
Nationality	*-0,1269****	*0,0074*	*-0,1845****	*0,0343*
Post-gymnasium level of education<2	*0,1868****	*0,0116*	*0,4645****	*0,0496*
Post-gymnasium level of education>2	*0,3125****	*0,0051*	*0,5079****	*0,0414*
University level of education	*0,6977****	*0,0266*	*1,4693****	*0,1077*
Married			*-0,2918****	*0,0235*
Single mother/father			*-0,1725****	*0,0367*
Living with parents			*0,6219****	*0,0416*
Children			*-0,0408****	*0,0420*
Regional wage	*1,3491****	*0,0493*	*3,4008****	*0,3948*
Manufacture	*0,1886****	*0,0084*	*-0,1179***	*0,0423*
Construction	*0,1623****	*0,01*	*0,2149***	*0,0776*
Retail	*0,0328****	*0,0074*	*0,3334****	*0,0261*
Private services	*0,2238****	*0,0125*	*0,4442****	*0,0413*
Accessibility			*-1,02e-06****	*4,52e-08*
Constant	*-3,9639*	*0,3898*	*-26,768****	*2,9132*
Sigma (σ_{ew})	*0,6537**** *0,1366*			
Number of observations	*198417*			

Significance level: ""p<0.05, "**"p<0.01, "***" p<0.001*
The standard errors are heteroskedasticity robust and clustered at the individual level

Selection into long-distance commuting on observed earnings is negative ($\hat{\alpha}$= - 0,4388, |*t*= 3,71|) and the estimated covariance parameter is positive (σ_{ew} =0.6537). As for the sample of men, the results for women show systematic influence of latent characteristics associated with lower earnings, but at the same time a higher probability of LDC for individuals with higher than expected earnings conditional on observed traits.

Estimated coefficients of linear and squared term of age suggest a concave age-earning profile of female workers and earnings increases with education. The estimated difference in earning between individuals having a Masters or a PhD degree vis-à-vis the reference category of individuals with less than 12 years of schooling is 69%. Female workers of foreign origin receive

on average 12% less than natives. Sector of employment also plays significant role in determining the level of earnings. Relatively to the public sector, the results suggest higher earnings in all other sectors: manufacturing 18%; construction 16%; retailing 3%; and private services 22%. Again, the regional wage level at the place of work is associated with higher earnings of individuals.

Also in line with the results for males, probability of LDC is concave in age and increases with level of education. Non-natives have lower probability of long-distance commuting and the regional wage level in the local labor market area of the work place seems to attract a larger share of LDC:s and the coefficient on access is negative. The estimates indicating relationship between sector of employment and LDC also show a similar pattern for females as for the corresponding results for the sample of men. In contrast to the results for men, having children is associated with a lower likelihood of commuting over longer distance among women.

In sum, the results demonstrate that long-distance commuters systematically self-selects from the lower part of the income distribution. It is in line with the prediction of our model which suggests that past income negatively affects the probability of commuting. Also, latent characteristics that affect earnings are positively correlated with the probability of LDC.

5 Robustness checks

To verify the robustness of estimated parameters of main interest (α and σ_{ew}), the empirical model was re-estimated using different definitions of long-distance commuting, a less restrictive sampling criteria, and by using an extended set of covariates.

Our definition of LDC > 50 km Euclidian distance, approximately corresponding to > [55-70] km road distance and at least 45 minutes one way travel, is meant to define a subsample of commuters who experience significant monetary and non-monetary losses associated with commuting. Following previous studies that analyse commuting behavior of individuals, we test 40, 50 and 60 kilometers as the threshold for definition of LDC (Mulalic & Pilegaard 2010; Sandow & Westin 2010b; Eliasson et al.; 2003, Manning 2003).

Based on our baseline specification of the model, Table 4 gives the results from stability checks with respect to different cutoffs for commuting distance defining LDC. The indicated negative selection on observed earnings into commuting is confirmed and the general pattern is that the negative selection increases with commuting distance. The results confirm our previous findings of a positive correlation between latent characteristics affecting earnings and the probability of LDC.

Table 4. Estimates by alternative criteria for definition of LDC

Variable	*Male sample*			*Female sample*		
	40 km	*50 km*	*60 km*	*40 km*	*50 km*	*60 km*
Previous earning (α)	*-0.5169****	*-0.6345****	*-0.4279****	*-0.2886****	*-0.4388****	*-0.5507****
	(0.0848)	*(0.0943)*	*(0.0596)*	*(0.1131)*	*(0.1184)*	*(0.1205)*
Sigma σ_{ew}	*0.7017****	*0.8753****	*0.7563****	*0.4354****	*0.6537****	*0.7881****
	(0.0951)	*(0.1059)*	*(0.0753)*	*(0.1289)*	*(0.1365)*	*(0.1407)*

Robust standard errors within parenthesis.
*Significance levels: * p<0.05, **p<0.01, *** p<0.001.*
The standard errors are clustered at the individual level.

The extended sample includes all individuals who find a full time job in 2008 regardless of previous commuting experience. The estimation results using our baseline model specification and alternative definitions of LDC are presented in Table 5. The point estimates α remain negative and increases with distance defining LDC. But they are now smaller in magnitude and not significantly different from zero for the sample of females. The estimates of the covariance parameter are positive statistically significant as before although they are smaller in magnitude.

Table 5. Estimates using alternative sampling criteria and alternative definitions of LDC.

Variable	*Male sample*			*Female sample*		
	40 km	*50 km*	*60 km*	*40 km*	*50 km*	*60 km*
Previous earning (α)	*-0.053**	*-0.068***	*-0.071***	*-0.072*	*-0.063*	*-0.007*
	(0.029)	*(0.032)*	*(0.036)*	*(0.052)*	*(0.059)*	*(0.059)*
Sigma (σ_{ew})	*0.178***	*0.206****	*0.213****	*0.216****	*0.216****	*0.140**
	(0.039)	*(0.044)*	*(0.048)*	*(0.068)*	*(0.077)*	*(0.085)*

Robust standard errors within parenthesis.
*Significance levels: * p<0.05, **p<0.01, *** p<0.001.*
The standard errors are clustered at the individual level.

Regarding the extended model specification, we included the family patterns characteristics such as marital status and presence of children in order to capture variation in labor supply in the earning equation. In addition, regional dummies were included into the earnings and commuting equations to control for different regional specific characteristics not reflected by the variables measuring regional wages and regional access.[8,9] Using the baseline sample criteria, and allowing for different thresholds of distance defining LDC, the estimates from the extended model specification are presented in Table 6. Again, the point estimates of alpha are negative but statistically significant only in two cases out of five. The covariance parameter is positive and statistically significant in most cases.

Table 6. Estimates for the extended model specification and alternative definitions of LDC

Variable	*Male sample*			*Female sample*		
	40 km	*50 km*	*60 km*	*40*	*50*	*60*
Previous earning (α)	*-0.0212*	*-0.2109**	*-0.2889****	*-0.0216*	*-0,0926*	*N/A*
	(0.0837)	*(0.0927)*	*(0.1029)*	*(0.0558)*	*(0,1655)*	
Sigma σ_{ew}	*0.1152*	*0.3659****	*0.4994****	*0.1608**	*0,2658**	*N/A*
	(0.092)	*(0.1026)*	*(0.114)*	*(0.0731)*	*(0,1865)*	

Standard errors are in parentheses below the main coefficients
Asterisks indicate significance level
Significance level: ""p<0.05, "**"p<0.01, "***" p<0.001*
The standard errors are heteroskedasticity robust and clustered at the individual level

The general impression of the robustness checks is that the estimated parameters of main interest are relatively robust for the sample of males as compared with the sample of females. The signs

[8] The regional dummies were aggregated on the NUTS2 level according to Nomenclature des Unites Territoriale Statistique (NUTS) classification of European Union.
[9] The results for the extended specification do not include the estimates on 60 km threshold determining LDC in the female sample due to the difficulties with convergence of the female sample.

of estimated parameters remain the same and they are statistically significant in most cases. Selection into LDC on observed earnings is negative and unobserved heterogeneity affecting earnings is positively correlated with unobserved factors affecting the probability of LDC. The magnitude of estimated parameters varies, moderately by different definitions of LDC but decreases substantially when using less restrictive sampling criteria. The results for females are more sensitive, especially for using different sampling criteria and extended model specification.

6 Summary and discussion

This study deals with non-random selection into long-distance commuting to work on observed and unobserved individual characteristics.

Using Swedish population register data we estimate self-selection into long-distance commuting on earnings and selection on latent characteristics affecting earnings and the probability of matching with a job involving a long commute. Earnings and latent characteristics affecting earnings are measured the year before commuting is observed, i.e. approximately at the time when the decision to commute is made.

Our findings indicate that long-distance commuters are negatively selected on earnings the year before they start to commute to their new work places. However, individuals with latent characteristics associated with higher than expected income have also a higher than expected probability of engage in long-distance commuting. Selection on earnings and latent characteristics show the same pattern for both women and men. However, the results for women are considerably less robust than the results for men.

The negative association between earnings and propensity for long-distance commuting may reflect that commuting is preferred to migration because of spatial variation in costs for housing. Commuting of high income specialists facing thin regional labor markets seems to be of less importance quantitatively. However, recent entrants to the labor market with high education may be found in the lower part of the income distribution and, because of thin regional labor markets for specialists, commuting may be necessary for matching their skills with higher paid jobs.

Conditional on observed characteristics, the positive correlation between unobserved traits affecting earnings and probability of long-distance commuting speaks against job mismatch as an

explanation for commuting to a new work place. A more plausible explanation is that heterogeneity in unobserved traits reflects individual ability associated with job search conducted with higher intensity, efficiency and over larger geographical areas.

To identify exact mechanisms for the observed positive selection on latent characteristics, further research using more direct measures of individual heterogeneity in cognitive and non-cognitive skills and measurement of different aspects of the job matching process is needed. For example, latent characteristics associated with higher than expected earnings (ex-ante) and higher probability of long-distance commuting, can yield even higher earnings (ex-post). Comparisons of how different dimensions of unobserved heterogeneity are associated with earnings (measured ex-ante and ex-post) may perhaps provide evidence on whether job mismatch on latent characteristics is a major explanation to long-distance commuting or not.

References

Albrecht, J., Bjorklund, A. and Vroman, S., (2003). Is there a glass ceiling in Sweden? *Journal of Labor Economics* 2003; 21(1), 145-177.

Bartel A. P. and Lichtenberg F. R. (1987).The comparative advantage of educated workers in implementing new technologies. *Review of Economics and Statistics* 69. pp 1-11

Borck, R. and Wrede, M., 2008. Commuting subsidies with two transport modes. *Journal of Urban Economics*, *63*(3), pp.841-848.

Borsch-Supan A. (1990) Education and its double-edged impact on mobility. *Economics of Education. Review* 9. pp.39-53

Edvardsson, I.R., Heikkila, E., Johansson, M., Persson, L.O., Stambol, L.S., (2000). Competitive Capital. Performance of Local Labor Markets – An International Comparison Based on Gross-stream Data. *Working Paper 2000:7. Nordregio, Stockholm*.

Finnie, Ross. (2001). "The Effects of Interprovincial Migration on Individuals' Earnings: Panel Model Estimates for Canada," *Research Paper No. 163. Business and Labor Market Analysis Division, Analytical Studies Branch. Statistics, Canada (October).*

Gabriel, P. E., & Schmitz, S. (1995). Favorable self-selection and the internal migration of young white males in the United States. *Journal of Human Resources*, 460-471.

Greenwood, M. J. (1985). Human migration: Theory, models, and empirical studies. *Journal of regional Science*, *25*(4), 521-544.

Gottholmseder, G., Nowotny, K., Pruckner, G. J. and Theurl, E. (2009). Stress perception and commuting, *Health Economics*, 18, pp. 559–576

Ham, M. Van, Mulder, C.H. & Hooimeijer, P., (2001). Spatial flexibility in job mobility macrolevel opportunities and microlevel restrictions. *Environment and Planning A*, 33(5), pp.921–940.

Hansson, E., Mattisson, K., Björk, J., Östergren, P.O. and Jakobsson, K., (2011). Relationship between commuting and health outcomes in a cross-sectional population survey in southern Sweden. *BMC public health*, *11*(1), p.1.

Hausman, J. A., & Wise, D. A. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica: Journal of the Econometric Society*, 455-473.

Eliasson, K., Lindgren, U. & Westerlund, O., (2003) Geographical Labor Mobility: Migration or Commuting? *Regional Studies*, 37(8), pp.827–837.

Hazans, M., (2004). Does Commuting Reduce Wage Disparities? *Growth and Change*, 35(3), pp.360–390.

Ihlanfeldt, K. R., (1988). "Intra-metropolitan variation in earnings and labor market discrimination." *Southern Economic Journal*, 55: 123-140

Koslowsky, M., Aizer, A. and Krausz, M., (1996). Stressor and personal variables in the commuting experience. *International Journal of Manpower*, 17(3), pp.4-14.

Lundholm, E., (2010) Interregional Migration Propensity and Labor Market Size in Sweden, 1970–2001. *Regional Studies*, 44(4), pp.455–464.

Manning, A., (2003) The real thin theory: monopsony in modern labor markets. *Labor Economics*, 10, pp.105–131.

Mulalic, I. & Pilegaard, N., (2010). Wages and Commuting : Quasi-Natural Experiments ' Evidence from Firms that relocate.

Murata, Y. and Thisse, J.F., (2005). A simple model of economic geography à la Helpman–Tabuchi. *Journal of Urban Economics*, *58*(1), pp.137-155.

Nakosteen, R.A., Westerlund, O. and Zimmer, M., (2008). Migration and self-selection: measured earnings and latent characteristics. *Journal of Regional Science*, *48*(4), pp.769-788.

Rouwendahl, J., (1998). Search theory, Spatial Labor Markets and commuting. *Journal of urban Economics*, 43, pp.1–22.

Rouwendahl, J., (1999). Spatial job search and commuting distances. *Regional sciense and urban economics*, 29 (491-517).

Rouwendal, J., (2004). Search Theory and Commuting Behavior. *Growth and Change*, 35(3), pp.391–418.

Rupert, P. & Wasmer, E., (2009). Housing and labor market: time to move and aggregate unemployment.

Ruppert, P., Stancanelli, E. and Wasmer, E., (2009). Commuting, wages and bargaining power. *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, pp.201-220.

Sandow, E., (2008). Commuting behaviour in sparsely populated areas: evidences from northern Sweden. *Journal of Transport Geography*, 16, pp.14–27.

Sandow, E., Westerlund, O. and Lindgren, U., (2014). Is your commute killing you? On the mortality risks of long-distance commuting. *Environment and planning A*, *46*(6), pp.1496-1516.

Sandow, E. & Westin, K., (2010). Preferences for commuting in sparsely populated areas.

Sandow, E. & Westin, K., (2010b). The persevering commuter - Duration of long-distance commuting. *Transportation Research Part A: Policy and Practice*, 44(6), pp.433–445.

Tunali, I., (2000). Rationality of migration. *International Economic Review*,*41*(4), pp.893-920.

Wasmer, E. & Zenou, Y., (2002). Does City Structure Affect Job Search and Welfare? *Journal of Urban Economics*, 51(3), pp.515–541.

Wasmer, E. & Zenou, Y., (2006). Equilibrium search unemployment with explicit spatial frictions B. , 13, pp.143–165.

Weber, J., & Sultana, S. (2007). Journey-to-Work Patterns in the Age of Sprawl: Evidence from Two Midsize Southern Metropolitan Areas*. *The Professional Geographer*, *59*(2), 193-208.

Westerlund, O. (1997). Employment opportunities, wages and international migration in Sweden 1970–1989. *Journal of regional science*, *37*(1), 55-73.

Zaiceva, A. (2006). Reconciling the estimates of potential migration into the enlarged European Union.

Öhman, M. & Lindgren U., (2003). "Who is the long-distance commuter? Patterns and driving forces in Sweden." *Cybergeo: European Journal of Geography*.