Discussion Papers No. 567, December 2008
Statistics Norway, Research Department

John K. Dagsvik, Torbjørn Hægeland and Arvid Raknerud

Estimating the Returns to Schooling: A Likelihood Approach Based on Normal Mixtures

Abstract:
In this paper we develop likelihood based methods for statistical inference in a joint system of equations for the choice of length of schooling and earnings. The model for schooling choice is assumed to be an ordered probit model, whereas the earnings equation contains variables that are flexible transformations of schooling and experience, with corresponding coefficients that are allowed to be heterogeneous across individuals. Under the assumption that the distribution of the random terms of the model can be expressed as a particular finite mixture of multinormal distributions, we show that the joint probability distribution for schooling and earnings can be expressed on closed form. In an application of our method on Norwegian data, we find that the mixed Gaussian model offers a substantial improvement in fit to the (heavy-tailed) empirical distribution of log-earnings compared to a multinormal benchmark model.

Keywords: Schooling choice, earnings equation, normal mixtures, treatment effects, self-selection, random coefficients, full information maximum likelihood

JEL classification: C31, I20, J30

Acknowledgement: Financial support from The Norwegian Research Council ("KUNI") is gratefully acknowledged.

Address: John K. Dagsvik, Statistics Norway, Research Department. E-mail: jda@ssb.no.
Torbjørn Hægeland, Statistics Norway, Research Department. E-mail: thd@ssb.no.
Arvid Raknerud, Statistics Norway, Research Department. E-mail: rak@ssb.no

Discussion Papers	comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1 Introduction

The relationship between schooling and earnings is one of the most frequently studied in empirical economics. A large number of these studies build upon versions of the earnings equation proposed by Mincer (1974). A key parameter in the Mincer earnings equation is the coefficient associated with years of schooling, intended to capture the effect on earnings differences caused by differences in schooling. However, to give a causal interpretation of the parameters in the earnings equation, one must take into account that the independent variable "years of schooling" is endogenous because it is the outcome of a choice variable. The endogeneity problem is related to the fact that the econometrician does not observe all factors that affect schooling choice. If some of these unobserved factors are correlated with unobservables in the earnings equation, OLS will produce biased estimates of the returns to schooling (ability bias).

Traditionally, ability bias is assumed to arise because of correlation between length of schooling and the additive error term in the earnings equation. If such correlation exists and is positive, it implies that people with high earnings capacity (irrespective of level of schooling) systematically choose a higher schooling level than people with low earnings capacity. In the literature, such heterogeneity is often termed "absolute advantage". Various econometric methods have been developed to deal with this problem, see Griliches (1977) for an overview of the early literature. Several more recent econometric studies have also taken into account that there may be heterogeneity not only associated with general earning capacity, but also associated with returns to schooling: Some individuals gain more from an extra year of schooling than others, cf. for example the theoretical model of Willis and

Rosen (1979). Heterogeneity of this sort is often termed "comparative advantage", and is typically dealt with by formulating a random coefficient model, in which the coefficient associated with years of schooling is allowed to vary across individuals according to some distribution function. If this random coefficient is correlated with the schooling variable or the additive error term in the earnings equation, then standard OLS estimates of returns to schooling will be biased.

To deal with ability bias and endogeneity of schooling, instrumental variable approaches have often been applied. As a result, there is now a substantial literature on how to interpret instrumental variable estimates in the case of heterogeneity in returns to schooling. See for example Angrist, Imbens and Rubin (1996), Wooldridge (2002) and Heckman and Vytlacil (2005). A somewhat closely related approach is the so-called two-stage- or control function approach. In this approach a choice-of-schooling equation is estimated in the first stage from which suitable variables are computed. In a second stage, these variables are used as additional regressors in the earnings equation, to account for the correlations between the schooling variable and the error terms, see Heckman (1979) and Garen (1984). Card (2001) gives an overview of these approaches to estimating earnings relations in the presence of individual heterogeneity in the returns to schooling.

In addition to the focus on different types of selection biases there has been a growing attention to the specification of the Mincer equation in the literature. One of the most important features of the Mincer equation is that log earnings is assumed to be linear in years of schooling, while another is the assumed separability between schooling and experience. Several papers, e.g. Heckman and Polachek (1974), Heckman, Lochner and Todd (2003) and Belzil (2007) have examined the validity of – and the consequences of relaxing – these and other functional form assumptions of the standard Mincer framework. A general finding is that some of the

simplifying assumptions are rejected, and hence that there is need for a framework that accommodates more flexibility.

Several authors have incorporated a structural, discrete choice dynamic programming approach to model schooling and related labor market decisions. Keane and Wolpin (1997) estimate a dynamic human capital investment model of schooling-, employment- and occupation-decisions, where skill heterogeneity and hence self-selection plays a role in all three choices. Belzil and Hansen (2002a) estimate a dynamic programming model where individuals differ in market and schooling ability, and relax the assumption of constant marginal returns to schooling. They find clear evidence of ability bias, and, perhaps more importantly, that the (log) wage-schooling relationship is highly non-linear, so that "...estimation methods that do not allow for a flexible estimation of the local returns to schooling will lead to unreliable estimates of both the local and the average return to schooling." Belzil and Hansen (2007) estimate a model with both absolute and comparative advantage (a correlated random coefficient wage regression model) within a dynamic programming framework. Belzil (2007) provides a thorough review of structural approaches to estimating the returns to schooling. He also compares this approach to the instrumental variables approach, and discusses commonalities and differences. On this latter point, see also Keane (2005).

The approach developed in this paper is, from the perspective of structural choice modelling, more modest than the dynamic programming setting. Specifically, in line with other works dating back to Cameron and Heckman (1998), we represent schooling choice by a simple stochastic index function that yields an ordered probit model. The idea of approximating individual schooling choices by a semi-structural model dates back to Cameron and Heckman (1998). The ordered model accounts for forward-looking behavior and unobserved heterogeneity. However, by essentially

modelling schooling choice as a static decision, an implicit assumption is that all future uncertainty is observed initially. Since the focus of our analysis is on the associated earnings relation, a structural dynamic schooling choice model does not seem necessary. In contrast, if the purpose is to analyze schooling choices per se, then a structural dynamic choice model is of interest.

In our study the earnings relation is rather general and flexible, both with respect to assumptions about the distribution of unobserved variables, functional forms, and the correlation structure of the random coefficients. As a result, we are able to allow for two types of self-selection into schooling; namely selection by "absolute advantage" (correlation between schooling and the additive error term in the earnings equation) and selection by "comparative advantage" (correlation between schooling and the random coefficients associated with the returns to schooling and experience).

Our approach offers several advantages over the traditional two-stage, control function approach. First, since estimation is carried out in one stage, we do not have to worry about biased estimates of the standard errors. Such biases may arise because of imputation of parameters estimated in the first-stage and because, conditional on the individual's choice, the error term is heteroscedastic. Second, our approach allows us to deal with non-linear transformations of earnings, schooling and experience that may contain unknown parameters (such as in Box-Cox transformations), as well as random components that can be represented as mixtures of normally distributed random variables. Third, our approach makes it easy to test interesting hypotheses by means of the likelihood ratio test, whereas in the two-stage method exact testing will be cumbersome.

Our framework has many similarities with Carneiro et al. (2003), who consider a setting with one probit model for the schooling choice and several measurement equations, with mixed multinormally distributed random components. Their estimation

strategy is based on a particular Bayesian approach which requires Markov Chain Monte Carlo methods. In contrast, we show that when the random components are mixed multinormally distributed one can express the corresponding likelihood function on closed form and derive explicit formulas for several types of treatment effects.

A key issue in the recent literature on returns to schooling (and more generally in the program evaluation literature), is the discussion of how key structural parameters associated with the returns to schooling can be identified. The strategy in the IV/experimentalist literature is to search for valid exclusion restrictions, where the excluded variables are the source of exogenous variation in the level of schooling. On the other hand, the structural literature relies more explicitly on parametric assumptions ("identification by functional form"). We emphasize that also within our framework interpretation of the results depends on exclusion restrictions, although such restrictions are not formally needed to obtain identification.

In an application of our method on Norwegian data, it is confirmed that selection effects due to unobservables are important when analyzing the returns to schooling. Specifically, we find a significant positive correlation between the error term of the schooling choice equation and the random coefficient of schooling in the earnings equation, and a significant negative correlation between the additive error term of the schooling choice equation and the additive error term of the earnings equation. Moreover, our study shows, similar to Heckman and Polachek (1974), that, for all practical purposes, the specification with logarithm of earnings fits the data best (within the class of Box-Cox transformations). Regarding the transformation of the independent variables, we find that piecewise linear functions of "length of schooling" and of "experience" give better fit and also substantially different results than generalized Box-Cox transformations (Box-Cox transformations with arbitrary

translations). While allowance for mixed normally distributed error terms is essential for obtaining a good fit to the empirical distributions of log earnings (given different levels of schooling), many of our results are quite robust with respect to the specification of the error distribution, including the estimated marginal returns to schooling as a function of years of schooling.

The rest of the paper is organized as follows. In Section 2 we present the modeling framework and derive several results that enable us to carry out empirical inferences. In Section 3 we present the empirical application, while Section 4 concludes the paper.

2 The modelling framework

In this section we specify the modelling framework for estimating the earnings equation and the choice of schooling relation. We first present a benchmark model with normally distributed error terms. We then extend this model to incorporate mixtures of normal distributions.

2.1 The basic model

We follow Cameron and Heckman (1998) in assuming a semi-structural probit model for the choice of length of schooling. From a choice theoretic perspective this model may be viewed a reduced form one, but it is semi-structural in the sense that it accounts for the hierarchical and discrete nature of the choice setting, in the presence of unobserved heterogeneity in preferences[1].

[1] The choice model only considers length of schooling and is silent about other potential important dimensions of the choice setting, such as type of schooling and occupational choice. Technically, this means that when we condition on length of schooling, type of schooling and occupation is exogeneous. This means that we implicitly make the (rather strong) assumption that, given the length of schooling, there is no self-selection into fields of study. While relaxing this assumption is outside the scope of the present paper, this is certainly an interesting topic for future research.

Let X^* be a latent index that represents the desired level of schooling on a continuous scale. The observed level of education, J, is a categorical variable with M possible categories; $J \in \{1, 2, .., M\}$. It is related to X^* through the relation

$$J = j \text{ iff } \mu_{j-1} < X^* < \mu_j \ , \ j = 1, ..., M, \tag{1}$$

where $\{\mu_j\}$ are unknown threshold values, except for $\mu_0 = -\infty$ and $\mu_M = \infty$. The variable J represents the choice of level of schooling as constrained by the institutional schooling system, whereas X^* represents the individual's preferences with regard to the level of schooling on a continuous scale. The threshold values $\{\mu_j\}$ determine the level of schooling in the institutional schooling system that corresponds to X^*.

Furthermore, we assume that

$$X^* = Z_1\gamma_1 + \varepsilon_1, \tag{2}$$

where Z_1 is a row-vector of exogenous variables affecting the individual's choice of schooling (typically family background variables describing the situation prior to the choice of schooling), ε_1 is a normally distributed random variable with zero mean and unit variance and γ_1 is a fixed, unknown coefficient vector. Thus, (1)-(2) specifies a standard ordered probit model for the discrete choice variable J.

Consider now the earnings equation. Let $T_1(X_1; \alpha_1)$ be a transformation of years of schooling, X_1, and $T_2(X_2; \alpha_2)$ a transformation of labor market experience, X_2. By experience we mean age minus years of schooling minus seven years, i.e., *potential* experience. Each of the transformations $T_1(X_1; \alpha_1)$ and $T_2(X_2; \alpha_2)$ may be a Box-Cox, polynomial, or spline function and possibly depend on unknown parameter vectors, α_1 and α_2, respectively. Our earnings equation is given by

$$(Y^\nu - 1)/\nu = T(X; \alpha)(\beta + \eta) + Z_2\gamma_2 + \varepsilon_2, \tag{3}$$

where ν is an unknown parameter to be estimated, $X = (X_1, X_2)$ and $T(X;\alpha) = (T_1(X_1;\alpha_1), T_2(X_2;\alpha_2))$. Moreover, $\eta = (\eta_1, \eta_2)'$ is a zero mean random coefficient vector, $\beta = (\beta_1, \beta_2)'$ is the corresponding fixed coefficient vector, Z_2 is a vector of exogenous variables which – in addition to the components of Z_1 – also may contain other variables affecting earnings, γ_2 is a vector of corresponding coefficients, and ε_2 is a zero mean random term.

Note that, with the usual convention that $(Y^\nu - 1)/\nu = \ln Y$ when $\nu = 0$, the dependent variable in (3) is a continuously differentiable transformation of Y. Also note that, through the random coefficient vector η, our model allows for heterogeneity in the coefficients of both schooling and experience. The vector of random terms $(\varepsilon_1, \varepsilon_2, \eta')$ is assumed to be multinormally distributed with zero mean and a general covariance matrix, apart from the conventional identifying restriction that ε_1 has unit variance. Even in the special case where the parameters ν and α are known (or given), one cannot estimate (3) by standard methods due to the fact that $T(X;\alpha)$ depends on ε_1, which may be correlated with both η and ε_2.[2]

Let $\zeta(X_1)$ be the function that assigns the schooling level that corresponds to X_1 years of schooling, i.e., $J = \zeta(X_1)$. If $\zeta(\cdot)$ is one-to-one, then $X_1 = \zeta^{-1}(J)$ and $X_2 = age - X_1 - 7$ (in this case $\zeta(\cdot)$ is, in fact, redundant). However, it may be useful to have a framework which allows a given level of schooling to cover several possible values for X_1. For example, one may want to assume (after initial exploration) that the self-selection is related to broader educational levels, such as short and long tertiary education, rather than actual years of schooling within these levels. For some specific years of schooling there may also be few observations. In our application in

[2] In the specification of the earnings relation given in (3), we have assumed separability between length of schooling and experience. Several papers, see, e.g., Heckman, Lochner and Todd (2008), have shown that this assumption may be unrealstic. In principle, it is possible also within our framework to incorporate transformations with interactions. However, since such an extension raises many new questions; e.g. with respect to functional form assumptions, how to incorporate heterogeneity in the interaction effects and interpretation of the results, we have decided to leave this problem aside for future research.

Section 3, the highest category of schooling ($j = 8$) covers the interval from 16 to 18 years of schooling. In that case, $\zeta(\cdot)$ is not one-to-one. However, the actual realization of X_1 *within the interval* is assumed to be exogenous in the sense that the distribution of X_1 conditional on J is independent of the random terms $\varepsilon_1, \varepsilon_2$ and η. Thus, in our application we ignore any selectivity issues related to the choice between, say, 16 and 17 years of schooling.

To denote the outcome of X given a particular level of schooling $J = j$, we use the notation X^j. Thus X^j may denote any value of X that is consistent with the choice $J = j$. Whereas X is an endogenous variable, X^j is exogenous. For example, given that $J = j$, X_2^j depends on age, which is exogenous.

Let Z denote the vector of *all* relevant exogenous variable of the model (including *age* and the variables in Z_1 and Z_2) and let

$$E(\varepsilon_1\varepsilon_2) = \theta,\ E(\varepsilon_1\eta_k) = \rho_k;\ k = 1, 2, \tag{4}$$

and $\rho = (\rho_1, \rho_2)'$. Then, we can write

$$\varepsilon_2 = \theta\varepsilon_1 + \widetilde{\varepsilon}_2,\ \eta = \rho\varepsilon_1 + \widetilde{\eta}, \tag{5}$$

where $\widetilde{\varepsilon}_2$ and $\widetilde{\eta}$ are independent of ε_1, with mean zero and a general covariance matrix, Σ. Let $\Phi(\cdot)$ denote the standard normal c.d.f. and $\phi(\cdot)$ the corresponding density. We have the following result:

Theorem 1 *Assume that* $(\varepsilon_1, \varepsilon_2, \eta')$ *is multinormally distributed with zero mean and let* Σ *be the covariance matrix of* $(\widetilde{\varepsilon}_2, \widetilde{\eta}')$*,*

$$g(T(X^j;\alpha))^2 = \left[\begin{pmatrix} 1, & T(X^j;\alpha) \end{pmatrix} \Sigma \begin{pmatrix} 1, & T(X^j;\alpha) \end{pmatrix}'\right] \tag{6}$$

and

$$\psi(T(X^j;\alpha))^2 = g(T(X^j;\alpha))^2 + (T(X^j;\alpha)\rho + \theta)^2. \tag{7}$$

If $f(y,j|Z)$ denotes the joint density of (Y,J) given Z, then

$$f(y,j|Z) =$$
$$\frac{y^{\nu-1}}{\psi(T(X^j;\alpha))}\phi\left(\frac{(y^\nu-1)/\nu - T(X^j;\alpha)\beta - Z_2\gamma_2}{\psi(T(X^j;\alpha))}\right)\times$$

$$\left\{\Phi\left(\left[\mu_j - Z_1\gamma_1 - \frac{((y^\nu-1)/\nu - T(X^j;\alpha)\beta - Z_2\gamma_2)\,(T(X^j;\alpha)\rho+\theta)}{\psi(T(X^j;\alpha))^2}\right]\frac{\psi(T(X^j;\alpha))}{g(T(X^j;\alpha))}\right)\right.$$

$$\left.-\Phi\left(\left[\mu_{j-1} - Z_1\gamma_1 - \frac{((y^\nu-1)/\nu - T(X^j;\alpha)\beta - Z_2\gamma_2)\,(T(X^j;\alpha)\rho+\theta)}{\psi(T(X^j;\alpha))^2}\right]\frac{\psi(T(X^j;\alpha))}{g(T(X^j;\alpha))}\right)\right\}. \tag{8}$$

The proof of the theorem is given in Appendix A.

Theorem 1 shows that the joint density of (Y,J) (conditional on Z) can be expressed on closed form by means of the normal c.d.f. and p.d.f. The first factor in (8) can be interpreted as the marginal distribution of Y when level of schooling is considered a fixed index (j) – that is, not as the outcome of the choice variable J. The second factor expresses the conditional distribution of J given Y.

The fact that one can express $f(y,j|Z)$ on closed form has several important advantages. First, it becomes easy to carry out maximum likelihood estimation and to perform statistical tests by means of the likelihood ratio statistic. Second, as we show in Corollary 2 below, it is easy to extend the model to the case where the distribution of the random components $(\varepsilon_1,\varepsilon_2,\eta')$ can be expressed as a particular finite mixture of multinormal distributions. Third, by utilizing the results in Corollary 3 below, several types of treatment effects commonly discussed in the literature, can be estimated.

2.2 Extension to normal mixtures

Similarly to Carneiro et al. (2003) we now consider the case where the distribution of the error term in the earnings equation is a finite mixture of normal distributions. Moreover, this distribution is allowed to depend on the chosen level of schooling, J. These extensions are highly relevant from an applied point of view. First, earnings data typically have heavy tails and may be skewed (also after applying appropriate transformations). Second, the shape of the earnings distribution may vary across different levels of schooling. Specifically, we assume in this section that the vector of error terms $(\varepsilon_1, \varepsilon_2, \eta')$ in the model analyzed in section 2.1 is replaced by $(\varepsilon_1(R), \varepsilon_2^J(R), \eta(R)')$, with

$$\varepsilon_2^J(R) = \kappa_{JR}\varepsilon_2(R) + \chi_{JR}, \tag{9}$$

where, for fixed j and r, κ_{jr} and χ_{jr} are unknown scale and location parameters, respectively, and $(\varepsilon_1(r), \varepsilon_2(r), \eta(r)')$ is an *i.i.d.* vector with the same ("standardized") mean-zero multivariate normal distribution as $(\varepsilon_1, \varepsilon_2, \eta')$ (specified in Section 2.1). Equations (2) and (3) in Section 2.1. are then replaced, respectively, by

$$X^* = Z_1\gamma_1 + \varepsilon_1(R) \tag{10}$$

and

$$(Y^\nu - 1)/\nu = T(X;\alpha)(\beta + \eta(R)) + Z_2\gamma_2 + \varepsilon_2^J(R). \tag{11}$$

The above specification means that $\varepsilon_2^J(R)$ is mixed Gaussian, whereas $(\varepsilon_1(R), \eta(R)')$ is multinormally distributed.[3]

Now let $f(y, j|Z, R = r)$ denote the joint density of earnings and chosen schooling level (Y, J), given Z and $R = r$. We then have the following result:

[3]Tecnically it is possible to allow both $\varepsilon_1(R)$ and $\eta(R)$ to be mixed Gaussian (similarly to $\varepsilon_2^J(R)$). But this extension is hardly interesting from an empirical point of view, as the data reveal little (if anything) about the shapes of the distributions of these variables.

Corollary 2 *Let* $(\varepsilon_1(r), \varepsilon_2(r), \eta(r)')$, $r = 1, 2, \ldots, Q$, *be i.i.d. multinormal random vectors with the same distribution as* $(\varepsilon_1, \varepsilon_2, \eta')$ *for every* r. *Let* R *be a multinomially distributed random variable, independent of* $(\varepsilon_1(r), \varepsilon_2(r), \eta(r)')$ *for each* r, *with* $P(R = r) = q_r$. *Assume that in the model in Section 2.1, (2) and (3) are replaced by (10) and (11), respectively, where* $\varepsilon_2^J(R)$ *is given by (9). Then*

$$f(y, j|Z, R = r) =$$

$$\frac{y^{\nu-1}}{\psi_{jr}(T(X^j;\alpha))}\phi\left(\frac{(y^\nu - 1)/\nu - T(X^j;\alpha)\beta - Z_2\gamma_2 - \chi_{jr}}{\psi_{jr}(T(X^j;\alpha))}\right) \times$$

$$\left\{\Phi\left(\left[\mu_j - Z_1\gamma_1 - \frac{\left((y^\nu - 1)/\nu - T(X^j;\alpha)\beta - Z_2\gamma_2 - \chi_{jr}\right)(T(X^j;\alpha)\rho + \theta)}{\psi_{jr}^2(T(X^j;\alpha))}\right]\frac{\psi_{jr}(T(X^j;\alpha))}{g_{jr}(T(X^j;\alpha))}\right)\right.$$

$$\left.-\Phi\left(\left[\mu_{j-1} - Z_1\gamma_1 - \frac{\left((y^\nu - 1)/\nu - T(X^j;\alpha)\beta - Z_2\gamma_2 - \chi_{jr}\right)(T(X^j;\alpha)\rho + \theta)}{\psi_{jr}^2(T(X^j;\alpha))}\right]\frac{\psi_{jr}(T(X^j;\alpha))}{g_{jr}(T(X^j;\alpha))}\right)\right\}, \tag{12}$$

where

$$g_{jr}(T(X^j;\alpha))^2 = \left[\left(\ 1,\ \ T(X^j;\alpha)\ \right)\Sigma_{jr}(r)\left(\ 1,\ \ T(X^j;\alpha)\ \right)'\right] \tag{13}$$

$$\Sigma_{jr}(r) = D_{jr}\Sigma D_{jr}; D_{jr} = \begin{bmatrix} \kappa_{jr} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{14}$$

$$\psi_{jr}(T(X^j;\alpha))^2 = g_{jr}(T(X^j;\alpha))^2 + (T(X^j;\alpha)\rho + \kappa_{jr}\theta)^2. \tag{15}$$

The proof of the Corollary is given in Appendix A.

Consequently, the joint density of (Y, J) conditional on Z can be expressed as

$$f(y, j|Z) = \sum_{r=1}^{Q} q_r f(y, j|Z, R = r).$$

We have thus shown that in the special case with only one outcome equation in addition to the choice equation, the likelihood function can be expressed on closed

form also in the case when the distribution of the the error term in the earnings equation is a finite mixture of normal distributions.

Our model can be seen as a (non-Bayesian) version of the model estimated in Carneiro et al. (2003) (see Section 7 and Appendix B in their paper). They consider (complicated) simulation based Bayesian inference in a model with several measurement or outcome equations (in addition to the schooling choice equation). The factor structure and the distributional assumptions they impose on the random terms, coincide with the distributional assumptions made in this paper when there is only one measurement equation (i.e., the earnings equation), except that we allow more flexibility by letting the parameters in the distribution of $\varepsilon_2^j(r)$ also to be specific for each level of schooling, j. This extension raises specific identification issues.

First, to obtain identification of the intercept in (11), we assume that

$$\sum_{r=1}^{Q} q_r \chi_{jr} = 0, \text{ for } j = 1, ..., M. \tag{16}$$

Thus, $E(\varepsilon_2^j(R)) = 0$ for every level of schooling, j. As a further identifying restriction we impose

$$\sum_{r=1}^{Q} q_r \kappa_{jr} = 1, \text{ for } j = 1, .., M. \tag{17}$$

As seen from Corollary 3 below, (17) has the important implication that $E(\varepsilon_2^k(R)|J = j)$ is independent of k. Thus, a person who actually chooses $J = j$, will by assumption have the same expected value of the additive error term $\varepsilon_2^k(R)$ at all (other) levels of schooling, k. The restriction (17) rules out that the idiosyncratic part of the marginal returns to schooling, which is assumed to be picked up by the random coefficient $\eta(R)$, may be confounded by a shift in the mean of the additive error term, leading to obvious problems of identifying and interpreting our model.

Corollary 3 *Under the assumptions of Corollary 2 and the restrictions in (16);*

$$E(\eta(R)|J=j) = -\rho\lambda(j) \tag{18}$$

and

$$E(\varepsilon_2^k(R)|J=j) = -\theta\left(\sum_{r=1}^{Q} q_r\kappa_{kr}\right)\lambda(j)\ ;\ j,k=1,..,M, \tag{19}$$

where

$$\lambda(j) = \frac{\phi(\mu_j - Z_1\gamma_1) - \phi(\mu_{j-1} - Z_1\gamma_1)}{\Phi(\mu_j - Z_1\gamma_1) - \Phi(\mu_{j-1} - Z_1\gamma_1)}. \tag{20}$$

The proof of Corollary 3 is given in Appendix A.

From Corollary 3 and the additional restriction (17), it follows that we can express (11) as

$$(Y^\nu - 1)/\nu = T(X;\alpha)\beta - \lambda(j)T(X;\alpha)\rho + Z_2\gamma_2 - \lambda(j)\theta + \varepsilon^*,\ j=1,..,M, \tag{21}$$

where

$$\varepsilon^* = \varepsilon_2^J(R) + \theta\lambda(j) + T(X;\alpha)(\eta(R) + \rho\lambda(j)),$$

and the error term, ε^*, has the property that $E(\varepsilon^*|J=j)=0$. Thus, if ν is known, it is possible to estimate β, ρ, γ_2 and θ consistently by a two-stage procedure in which $\lambda(j)$ is obtained in a first stage probit analysis using data on schooling choices, whereas the earnings equation is estimated in a second stage by (a possibly non-linear) least squares with $\lambda(j)$ and $\lambda(j)T(X;\alpha)$ as additional regressors. Note that the coefficients of these regressors do not depend on the mixing parameters. This is due to (17). If (17) is *not* imposed, the coefficient of $\lambda(j)$ becomes $-\theta\sum_{r=1}^{Q} q_r\kappa_{jr}$ and hence depends on j. Thus (17) ensures that the mixing parameters only affect the shape of the earnings distribution at different levels of schooling (variance, skewness, kurtosis, etc.), but not the causal effects of schooling. Also note that despite the fact that there are several endogenous unobservables in the earnings equation, i.e., $\varepsilon_2^J(R)$ and $\eta(R)$, only one "control function", $\lambda(j)$, is needed to control for selectivity bias.

2.3 Definition of treatment effects

When analyzing the implications of alternative schooling choices, it is of interest to calculate causal effects (treatment effects). The first is the average treatment effect, $ATE(x)$:

$$ATE(x) = \beta_1[T_1(x;\alpha_1) - T_1(x-1;\alpha_1)],$$

where the expression in the squared bracket is the change in the transformation of schooling, when years of schooling increases from $x-1$ to x. The second is the "effect of the treatment on the treated", $TT(x)$, given by

$$\begin{aligned} TT(x) &= (\beta_1 + E(\eta_1(R)|X_1 = x-1)))) \left[T_1(x;\alpha_1) - T_1(x-1;\alpha_1)\right] \\ &= (\beta_1 - \rho_1\lambda(\zeta(x-1))) \left[T_1(x;\alpha_1) - T_1(x-1;\alpha_1)\right], \end{aligned}$$

cf. (18), which has the interpretation of the marginal return of increasing years of schooling from $x-1$ to x for those who did in fact choose $X_1 = x-1$. Note that

$$E\left(\varepsilon_2^{\zeta(x)}(R) - \varepsilon_2^{\zeta(x-1)}(R)|J = \zeta(x-1)\right) = 0$$

due to (17) and Corollary 3, and hence does not enter the expression for $TT(x)$. The third effect is the observed differentials between levels of schooling, $OD(x)$:

$$\begin{aligned} OD(x) &= TT(x) + E(\varepsilon_2^{\zeta(x)}(R)|X_1 = x) - E(\varepsilon_2^{\zeta(x-1)}(R)|X_1 = x-1) \\ &= (\beta_1 - \rho_1\lambda(\zeta(x-1))) \left[T_1(x;\alpha_1) - T_1(x-1;\alpha_1)\right] - \theta(\lambda(\zeta(x)) - \lambda(\zeta(x-1))), \end{aligned}$$

cf. (18) and (19). This is the sum of (i) the average treatment effect, (ii) the average of the idiosyncratic marginal returns to schooling for the individuals with this level of schooling and (iii) the average idiosyncratic earnings level effect for the same individuals.

3 An empirical application on Norwegian data

3.1 Data and transformations

The data for this application are taken from the Norwegian system of register data, where individual information about essentially all Norwegian residents is gathered from a number of governmental administrative registers. Our sample is randomly drawn from the population of native-born males, who were born between 1952 and 1970, and who were living in Norway in both 1970 and 1997. The data contain information on years of schooling and type of education for each individual. The earnings equation sample is further restricted to full-time wage-earners, defined as individuals working 30 hours or more per week, leaving us with 29332 observations. Labor market experience is represented by potential experience, i.e., age minus years of schooling minus seven years. The earnings measure used is total annual taxable labor income. Because the earnings measure reflects annual earnings, observations where employment relationships started or terminated within the actual year were excluded. Holders of multiple jobs and individuals who have received labor market compensation or have participated in active labor market programs have been excluded. Family background information is taken from the National Census of the Population and Housing in 1970. A full list of variables with key summary statistics is given in Tables 2-4.

In our application the level of schooling is divided into eight groups, i.e., $J \in \{1, 2, .., 8\}$. Level 1 covers the interval $[7, 9]$ years of schooling, levels 2 to 7 correspond to 10-15 years, respectively, whereas level 8 covers the interval $[16, 18]$ years. The first category represents compulsory level of schooling (which was gradually increased from seven to nine years from the late 1950s to the early 1970s). The last category comprises longer tertiary education. We consider four types of transfor-

mation functions of schooling ($k = 1$) and experience ($k = 2$). Assume that x is an integer and let $[j/2]$ denote the integer value of $j/2$:

$$\begin{array}{ll} \text{Linear:} & T_k(x;\alpha_k) = x \\ \text{Quadratic:} & T_k(x;\alpha_k) = \left[(x+\alpha_{k,1})^2 - 1\right]/2 \\ \text{Generalized Box-Cox:} & T_k(x;\alpha_k) = \left[(x+\alpha_{k,1})^{\alpha_{k,2}} - 1\right]/\alpha_{k,2} \\ \text{Spline:} & T_k(x;\alpha_k) = \sum_{j=1}^{x_k} \alpha_{k,[j/2]},\ \alpha_{k,0} = 1. \end{array} \tag{22}$$

When $k = 1$, x denotes years of schooling *exceeding* 7 years (which is the minimum value of X_1 in our data). When $k = 2$, x denotes potential experience, defined as age minus years of schooling minus seven years. The spline transformation of x has knots every even year $(2, 4, 6, 8, ...)$. Thus, because the maximum values of X_1 and X_2 in our sample is 18 years of schooling and 29 years of experience, respectively, we are able to identify five $\alpha_{1,[j/2]}$-parameters ($[(18-7)/2] = 5$) and 14 $\alpha_{2,[j/2]}$-parameters ($[29/2] = 14$). Note that the linear and quadratic transformations are special cases of the (generalized) Box-Cox transformation, obtained by setting $\alpha_{k,2} = 1$ and $\alpha_{k,2} = 2$, respectively.

The vector of explanatory variables in the earnings equation, Z_2, includes indicators about sector of occupation (public, private services, manufacturing), field of education (general, technical, humanistic, teaching, administrative, etc.) and indicators for each of 19 counties where the individual works. The vector of explanatory variables of the ordered probit model for schooling choice, Z_1, contains variables regarding the family background. These include dummy variables for birth cohort, indicators of whether the individual as a child lived with both parents or alone with either mother or father, the labor market status of the parents, indicators of household income (quintile and both the father's and mother's education level), and whether the person had a mother and/or father who was born abroad. In addition, the schooling choice equation contains indicator variables for the county where the individual grew up, for example, where the individual lived in 1970. The main exclu-

sion restriction in this application, which in addition to functional form assumptions identifies the parameters of the model, is that given all the other covariates in the model, the region where you grew up may affect your choice of schooling, but not your earnings. It is well documented that educational choices vary considerably across regions in Norway. This is true also when conditioning on, for example, family background variables. This exclusion restriction is in the spirit of Card (1995) who used college proximity as an instrument, but may be interpreted in a more general sense as variations in the opportunity cost of education.

3.2 Results with normally distributed error terms

Estimation results for some key combinations of transformations of earnings, schooling and experience are displayed in Table 1 in the case with normally distributed error terms. A full set of results is reported in Tables 2-4. When interpreting the results in Table 1, it is important to bear in mind that the parameter estimates of β_1 and β_2 are not comparable across different models, as they are coefficients of different transformations of schooling and experience. Moreover, whereas the models reported in the first three columns of Table 1 have log earnings as the dependent variable, the last column reports results from a specification with a general Box-Cox transformation of earnings.

From Table 1, we first note that the linear-quadratic specification with regard to schooling and experience, i.e., the traditional Mincer model, gives a substantially lower log-likelihood than the Box-Cox model (Model 2) and – particularly – the spline models (Models 3-4). On the other hand, when $\nu = 0$, the spline transformations of x_1 and x_2 give considerably higher likelihood than the Box-Cox transformations – but at the cost of 15 more parameters. Although the model with spline transformations of x_1 and x_2 is clearly the most flexible with respect to para-

meterization, it is not a special case of neither the Box-Cox nor the linear-quadratic specification. On the other hand, the linear-quadratic specification is a special case of Box-Cox, with three parameters less. Because the maximum likelihood estimates are $\widehat{\alpha}_{2,1} = 2.49$ and $\widehat{\alpha}_{2,2} \approx 0$, we see that the estimated Box-Cox transformation of experience amounts to $\ln(x_2 + 2.49)$.

With regard to the transformation of earnings, the general Box-Cox transformation leads to an estimate of ν equal to -.17, with a standard error of only .003. The results suggest that ν is significantly different from zero. However, from the point of view of economic significance $\nu = -.17$ is so close to zero that the Box-Cox and logarithmic transformation are equivalent for practical purposes. We illustrate this point below.

The estimated correlations between the stochastic terms have interesting economic interpretations and give information on the nature of self-selection. However, the pair-wise correlations reported in Table 1 show that many of these are not robust across different model specifications. For example, we find strong evidence of negative correlation between η_2 and ε_2 when $\nu = 0$, but not at the maximum likelihood estimate $\nu = -.17$. However, with regard to the correlations that have the clearest economic interpretation we get quite striking results. First of all, it is evident that self-selection does matter. Concentrating henceforth on the results from the Box-Cox and spline transformations of schooling and experience, which overall give the best fit to the data and the most plausible results, there are significant negative correlations between ε_1 and ε_2, i.e., the residual terms of the earnings and schooling equations. We also find strong positive correlations between η_1 and ε_1. Using spline transformations of x_1 and x_2, we obtain correlation coefficients of the same magnitude as for the Box-Cox transformations, regardless of whether $\nu = 0$ or $\nu = -.17$. The robust finding that $\text{Corr}(\eta_1, \varepsilon_1) > 0$ implies that individuals

who have a high preference for schooling (conditional on the exogenous variables) also have high marginal returns to schooling. On the other hand, the finding that $\text{Corr}(\varepsilon_1, \varepsilon_2) < 0$ means that if an individual with a high preference for schooling takes a short education, his earnings potential is lower than for an individual with the same education, but with a low preference for schooling. The correlations mentioned above have the interpretation of positive selection by comparative advantage and negative selection by absolute advantage, respectively. These patterns may also be interpreted as selection by different type of skills, with a high ε_2 reflecting high blue-collar skills, and a high ε_1 reflecting high white-collar skills. It should be kept in mind that the correlations reported in Table 1 depend on the respective specifications and cannot be interpreted independently of the chosen transformations of length of schooling and experience.

There is considerable heterogeneity in the returns to schooling and experience, as seen from the estimated standard deviations $\text{SD}(\eta_1)$ and $\text{SD}(\eta_2)$ of η_1 and η_2, respectively, which are of the same magnitude as the estimated fixed coefficients, $\widehat{\beta}_1$ and $\widehat{\beta}_2$. To evaluate the importance of individual heterogeneity in the returns to experience and schooling, it is natural to look at the variation coefficients $\text{SD}(\eta_1)/\widehat{\beta}_1$ and $\text{SD}(\eta_2)/\widehat{\beta}_2$. These ratios lie between 1/10 and 1 in all the model specifications and are smaller for schooling than for experience. Thus, it seems that relative to the fixed coefficient, β_1 and β_2, the unobserved heterogeneity in returns to experience is larger than in returns to schooling. This higher cross section dispersion in returns to experience is consistent with what is reported in Belzil and Hansen (2002b). As a further check of the importance of heterogeneity in the coefficients of schooling and experience, a model with only a fixed coefficient vector (i.e., no η-vector) has been estimated. This restriction reduces the number of parameters by nine. However, it is firmly rejected by a likelihood ratio test.

The differences in results across the four model specifications are illustrated in Figures 1–4, along with the results from a linear-quadratic specification without selection effects (equivalent to OLS estimation of a standard Mincer equation). Figure 1 shows expected log earnings as a function of years of schooling when all the other variables of the earnings equation are set equal to their sample mean. In particular, years of experience is fixed at 15 years. The intercepts of the different graphs in the figure are determined by the (identifying) condition that when all the variables are at their sample means, expected log earnings should be equal in all the four model specifications. We see that the two versions of the model with spline transformations of schooling depicted in Figure 1, i.e., with $\nu = 0$ and $\nu = -.17$ as the dependent variables, are almost identical, except for small discrepancies at low values of years of schooling.

In analyses of returns to schooling and experience, the marginal returns to schooling and the earnings-experience profiles are of key interest. In models allowing for heterogeneity in returns, there are several possible "marginal returns" or "treatment effects" that may be calculated, based on the estimation results. Which effects that are most relevant, depend on the purpose of the analysis. In models with no heterogeneity in the returns, all treatment effects coincide.

Figure 2 shows the expected marginal returns to schooling corresponding to the three specifications depicted in Figure 1 that have $\ln Y$ as the dependent variable. The natural interpretation of the estimates from the models with selection effects is as the "average treatment effect" of schooling (ATE). This means that the graphs show the marginal effect on earnings of the *last* year of schooling, for a randomly selected individual whose years of schooling is shown on the horizontal axis. In contrast, the interpretation of the OLS estimate shows the (conditional) earnings differentials between individuals with different levels of schooling. In the absence of

selection effects, OLS and the linear specification will coincide.

Comparing OLS with full information maximum likelihood estimation of the linear specification, we see from Figures 1 and 2 that allowing for selection effects does matter for the estimated returns to schooling. From Figure 2 we find a marginal returns to schooling which is around one percentage point higher when we allow for selection effects. When comparing the linear specification with the more flexible specifications, we see that there are considerable differences in the estimated marginal returns across different levels of schooling. In particular, there are high returns to completing upper secondary school (12 years) and to take one or two years of higher education, whereas the marginal return to the last year of schooling, if the current level of schooling is 15 years or more, is considerably smaller. This is consistent with the findings of several empirical studies of returns to schooling using Norwegian data, cf. e.g. Hægeland, Klette and Salvanes (1999). Thus the strongest non-linearity in returns to education in Norway appears to arise from a particular high return to taking *some* higher education. One may speculate that this partly reflects a positive signal of high productivity to employers.

Figure 3 shows expected log earnings as a function of years of experience, with all the other variables of the earnings equation fixed at their sample means. In contrast to the estimated returns to schooling, allowing for selection effects only has minor implications for the estimated returns to experience. We see from Figure 3 that the Box-Cox specification gives higher marginal returns for years of experience up to four to five years compared to the other specifications.

Concentrating on our preferred specification, with spline transformations of both schooling and experience and with log earnings as the dependent variable in the earnings equation, Figure 4 depicts the three different kinds of marginal returns to schooling defined in Section 2.3. We see that the average effect of the treatment

on the treated (TT) in general is higher than the average treatment effect (ATE). This reflects the positive correlation between ε_2 and η_1 that was reported in Table 1: Individuals with higher (idiosyncratic) returns to schooling also invest more in schooling. Hence the marginal returns at a specific level are higher for those who actually have completed this level of schooling than for the average individual. In other words, there is selection by comparative advantage. On the other hand, we also estimated a negative correlation between ε_1 and ε_2; conditional on idiosyncratic returns to schooling, those with higher earnings potential regardless of schooling – all else equal – tend to choose a lower level of schooling. This is clearly seen from the earnings-schooling profiles in Figure 1. The self-selection related to ε_2 gives a flatter profile, i.e., individuals with high ε_2 tend to have low levels of schooling and vice versa.

To evaluate the fit of our preferred specification, Figure 5 plots (i) the discrete probability density functions over a grid of 100 intervals, with equal width, for the estimated spline model with log earnings as the dependent variable, and (ii) histograms of the log earnings data. This is done conditional on the chosen level of schooling, i.e., for eight different levels. Note that the derived theoretical distributions are not normal. They are obtained from (8), by integrating out (Z_1, Z_2) using the empirical distribution function of these covariates (given the level of schooling). We see that the estimated model based on the normal distribution is unable to pick up the heavy tails that characterize the histograms in Figure 5.

3.3 Results for the case with mixture distributions

We confine our analyses here to the case with $\ln Y$ as the dependent variable and with spline transformations of both years of schooling and years of experience. The results in Table 5 refer to the mixture model with $Q = 2$ and $Q = 3$, i.e., two and three mixture distributions, respectively. The new results are comparable to Model

3 in Table 1, i.e., the (benchmark) model with normally distributed error terms and identical ε_2-distribution across levels of schooling. The benchmark model is a special case of a mixture model with $Q = 1$. Detailed results regarding the mixture parameters are given in Table 6 for the case with $Q = 3$. By comparing the next-to-last row in Table 1 and 5, we see a formidable increase in log-likelihood when we allow normal mixtures. When $Q = 3$ we obtain a log-likelihood which is 1600 points higher than the benchmark model with $Q = 1$ reported in Table 1 (Model 3). The increase in log-likelihood when going from $Q = 2$ to $Q = 3$ is also very large; about 230 points, at the cost of 17 additional parameters.

The estimated coefficients of skewness and kurtosis in Table 5 for the error term $\varepsilon_2^J(R)$ show clear evidence of non-normality.[4] The coefficient of kurtosis is significantly above 3 in both models. When $Q = 3$ we obtain the highest coefficient of kurtosis (about 6), thus indicating an aggregate earnings distribution (across schooling levels) with very heavy tails. Also the coefficient of skewness is significantly different from zero according to the latter model, and the estimate (0.15 in both models) indicates a modest skewness to the right. Regarding the correlation coefficients $\text{Corr}(\varepsilon_2^J(R), \varepsilon_1)$ and $\text{Corr}(\eta_1, \varepsilon_1)$, these have the same sign as in the benchmark model (Model 3) and they are both significantly different from zero.

The most interesting question is perhaps how the estimated returns to schooling are affected when we allow normal mixtures. The answer is evident from figures 6 and 7. The normal benchmark model (Model 3) and the two mixture models have estimated average returns to schooling (ATE) that are quite similar. The only notable difference is that the two latter models exhibit about 2 percentage points lower ATE when years of schooling is less than or equal to 11 years, and 1-2 percentage points higher for 15-16 years of schooling. We see that the estimated

[4] The results involving $\varepsilon_2^J(R)$ in Table 5 are obtained by simulations from the estimated distributions of $\varepsilon^j(r)$, J and R.

returns to schooling is modestly affected by whether we choose $Q = 2$ or $Q = 3$.

Concentrating henceforth on the mixture model with with $Q = 3$, Figure 7 depicts estimates of the three types of treatment effects regarding the returns to schooling defined in Section 2.3. The graphs are quite similar to the corresponding graphs in Figure 4, with normally distributed error terms. Again, we see that average effect of the treatment on the treated (TT) in general is higher than the average treatment effect (ATE). Note, however, that the differences between the graphs in Figure 7 are generally smaller than between the corresponding graphs in Figure 4. This is related to the fact that $\text{Corr}(\varepsilon_2^J(R), \varepsilon_1)$ and $\text{Corr}(\eta_1, \varepsilon_1)$ when $Q = 3$ (reported in Table 5) are smaller in magnitude than the corresponding correlations for Model 3 reported in Table 1.

Figure 8 reproduces Figure 5 in the case where the theoretical model depicted in Figure 5 is replaced by the normal mixture model with $Q = 3$. For all levels of schooling we see that the estimated conditional probability density functions fit the histograms of the log earnings data well. The improvement compared to Figure 5 is particularly striking for schooling levels 7 and 8, where the normal benchmark model fits the data quite poorly (cf. Figure 5, chart 7 and 8). A QQ- plot for the marginal distribution of log-earnings is presented in Figure 9. The plot compares the empirical distribution function (the straight line) with the mixture model ($Q = 3$) and the normal benchmark model ($Q = 1$). The overall impression from these graphs is that the depicted mixture model fits the data well, and that a substantial improvement compared to the normal benchmark model is achieved. Similar graphs for the mixture model with $Q = 2$ reveal a somewhat poorer fit than when $Q = 3$, especially for schooling level 7 and 8, but we still get a clear improvement compared to the benchmark model.

4 Conclusion

In this paper we have discussed maximum likelihood estimation of a joint model for earnings and the choice of level of schooling. The earnings relation is allowed to be very general with random coefficients and explanatory variables that are flexible transformations of schooling and experience. The choice of level of schooling is assumed to be an ordered probit model. Under the assumption that the random terms of the model have a mixed multinormal distribution, we have demonstrated that the joint distribution of the choice of level of schooling and earnings as well as explicit formulas for several types of treatment effects regarding the returns to schooling, can be expressed on closed form.

We have applied this framework and methodology to analyze the structure of the earnings relation on micro data for Norway. The estimation results show that if we constrain the transformation of the dependent variable to be of the Box-Cox type, the logarithm of earnings seems to be the best one in terms of fit. Within the class of Box-Cox transformations, or alternatively spline transformations of the independent variables "years of schooling" and "potential experience", the latter family turns out to give the best fit. Compared to a multinormal benchmark model, the mixed multinormal model offers a substantially improved fit to the (heavy-tailed) empirical distribution of the actual log-earnings data.

We believe that the econometric framework developed in this paper offers several advantages to the researcher compared to the two-stage control function approach. First, because it is a maximum likelihood approach based on the mixed Gaussian distribution, it offers considerable flexibility. Second, it allows for nonlinear transformations of the dependent variable that contain unknown parameters. Third, biases due to heteroscedasticity and imputed estimates from the first stage that typically plague the control function approach no longer exist. Fourth, the maximum likeli-

hood approach facilitates testing of alternative model specifications.

References

[1] Angrist JD, Imbens GW, Rubin DB. 1996. Identification of Causal Effects Using Instrumantal Variables. *Journal of the American Statistical Association* **91**: 444–472.

[2] Belzil C. 2007. The return to schooling in structural dynamic models: a survey. *European Economic Review* **51**: 1059–1105.

[3] Belzil C, Hansen J. 2002a. Unobserved Ability and the Return to Schooling. *Econometrica* **70**: 2075–2091.

[4] Belzil C, Hansen J. 2002b. A structural analysis of the correlated random coefficient wage regression model. IZA Discussion Paper 512.

[5] Belzil C, Hansen J. 2007. A structural analysis of the correlated random coefficient wage regression model. *Journal of Econometrics* **140**: 827–848.

[6] Cameron SV, Heckman JJ. 1998. Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males. *Journal of Political Economy* **106**: 262–333.

[7] Card D. 1995. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp,* L.N. Christofides, E.K. Grant and R. Swidinsky (eds.). University of Toronto Press: Toronto.

[8] Card D. 2001. Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica* **69**: 1127–1160.

[9] Carneiro P, Hansen K, Heckman JJ. 2003. Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement

of the Effects of Uncertainty on College Choice. *International Economic Review* **44**: 361–422.

[10] Garen J. 1984. The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable. *Econometrica* **52**: 1199–1218.

[11] Griliches Z. 1977. Estimating the Returns to Schooling: Some Econometric Problems. *Econometrica* **45**: 1–22.

[12] Gronau R. 1974. Wage Comparisons – A Selectivity Bias. *Journal of Political Economy* **82**: 1119–1143.

[13] Heckman JJ. 1974. Shadow Prices, Market Wage, and Labor Supply. *Econometrica* **42**: 679–694.

[14] Heckman, JJ. 1979. Sample Selection Bias as a Specification Error. *Econometrica* **47**: 153–162.

[15] Heckman, JJ, Lochner LJ, Todd PE. 2008. Earnings Functions and Rates of Return. NBER Working paper 13780.

[16] Heckman JJ, Polachek S. 1974. Empirical Evidence on the Functional Form of the Earnings-Schooling Relationship. *Journal of the American Statistical Association* **69**: 350–354.

[17] Heckman JJ, Vytlacil E. 2005. Structural Equations, Treatment Effects and Econometric Policy Evaluation. *Econometrica* **73**: 669–738.

[18] Hægeland T, Klette TJ, Salvanes KG. 1999. Declining returns to education in Norway? Comparing estimates across cohorts, sectors and over time. *Scandinavian Journal of Economics* **101**: 555–576.

[19] Keane MP. 2005. Structural vs. Atheoretic Approaches to Econometrics, Keynote Address at the Duke Conference on Structural Models in Labor, Aging and Health, September 17–19, 2005.

[20] Keane MP, Wolpin KI. 1997. The Career Decisions of Young Men. *Journal of Political Economy* **105**: 473–522.

[21] Mincer J. 1974. *Schooling, Experience and Earnings*. Columbia University Press: New York.

[22] Willis RJ, Rosen S. 1979. Education and Self-Selection. *Journal of Political Economy* **87**: 7–36.

[23] Wooldridge JM. 1997. On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model. *Economics Letters* **56**: 129–133.

[24] Wooldridge JM. 2002. Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model. *Economics Letters* **79**: 185–191.

Appendix A: Proofs

PROOF OF THEOREM 1.

Define $Y^{[\nu]} = (Y^\nu - 1)/\nu$. Inserting (5) into (3) we obtain

$$Y^{[\nu]} = T(X;\alpha)\beta + Z_2\gamma_2 + (T(X;\alpha)\rho + \theta)\varepsilon_1 + T(X;\alpha)\widetilde{\eta} + \widetilde{\varepsilon}_2.$$

Since J is independent of $\widetilde{\varepsilon}_2$ and $\widetilde{\eta}$, we have:

$$\begin{aligned} \mathrm{Var}(T(X;\alpha)\widetilde{\eta} + \widetilde{\varepsilon}_2 | J = j, Z) &= \mathrm{Var}(T(X^j;\alpha)\widetilde{\eta} + \widetilde{\varepsilon}_2 | Z) \\ &= g(T(X^j;\alpha))^2, \end{aligned}$$

using the definition of $g(\cdot)$ in (6). Given $(\varepsilon_1,\, J, Z)$ we therefore obtain

$$P(Y^{[\nu]} \in (z, z+dz) | \varepsilon_1, J = j, Z) = \frac{dz}{g(T(X^j;\alpha))} \phi\left(\frac{z - T(X^j;\alpha)\beta - Z_2\gamma_2 - (T(X^j;\alpha)\rho + \theta)\varepsilon_1}{g(T(X^j;\alpha))} \right). \tag{23}$$

Let $K_j = (\mu_{j-1} - Z_1\gamma_1, \mu_j - Z_1\gamma_1]$. Since $J = j \Leftrightarrow \varepsilon_1 \in K_j$, we obtain from (23) that

$$\begin{aligned} & P(Y^{[\nu]} \in (z, z+dz), J = j | Z) \\ = & \int P(Y^{[\nu]} \in (z, z+dz) | \varepsilon_1, J = j, Z) P(J = j | Z_1, \varepsilon_1) \phi(\varepsilon_1)\, d\varepsilon_1 \\ = & \int P(Y^{[\nu]} \in (z, z+dz) | \varepsilon_1, J = j, Z) 1(\varepsilon_1 \in K_j) \phi(\varepsilon_1)\, d\varepsilon_1 \\ = & \int_{\mu_{j-1} - Z_1\gamma_1}^{\mu_j - Z_1\gamma_1} \frac{dz}{g(T(X^j;\alpha))} \phi\left(\frac{z - T(X^j;\alpha)\beta - Z_2\gamma_2 - (T(X^j;\alpha)\rho + \theta)\varepsilon_1}{g(T(X^j;\alpha))} \right) \phi(\varepsilon_1)\, d\varepsilon_1, \end{aligned} \tag{24}$$

where $1(B)$ is the indicator function which is one if the event B is true and zero else. Since

$$\frac{1}{\sigma}\phi\left(\frac{a - b\varepsilon_1}{\sigma}\right)\phi(\varepsilon_1) = \frac{1}{(\sigma^2 + b^2)^{1/2}} \phi\left(\frac{a}{(\sigma^2 + b^2)^{1/2}}\right) \times \frac{d}{d\varepsilon_1} \Phi\left(\left[\varepsilon_1 - \frac{ab}{(\sigma^2 + b^2)} \right] \frac{(\sigma^2 + b^2)^{\frac{1}{2}}}{\sigma} \right), \tag{25}$$

we obtain from (24), using (25) with $a = z - T(X^j;\alpha)\beta - Z_2\gamma_2$, $b = T(X^j;\alpha)\rho + \theta$ and $\sigma = g(T(X^j;\alpha))$, that

$$\begin{aligned} P(Y^{[\nu]} \in (z, z+dz), J = j|Z) = \frac{dz}{\psi(T(X^j;\alpha))}\phi\left(\frac{z - T(X^j;\alpha)\beta - Z_2\gamma_2}{\psi(T(X^j;\alpha))}\right) \times \\ \left\{\Phi\left(\left[\mu_j - Z_1\gamma_1 - \frac{(y - T(X^j;\alpha)\beta - Z_2\gamma_2)\,(T(X^j;\alpha)\rho + \theta)}{\psi(T(X^j;\alpha))^2}\right]\frac{\psi(T(X^j;\alpha))}{g(T(X^j;\alpha))}\right)\right. \\ \left.-\Phi\left(\left[\mu_{j-1} - Z_1\gamma_1 - \frac{(y - T(X^j;\alpha)\beta - Z_2\gamma_2)\,(T(X^j;\alpha)\rho + \theta)}{\psi(T(X^j;\alpha))^2}\right]\frac{\psi(T(X^j;\alpha))}{g(T(X^j;\alpha))}\right)\right\}. \end{aligned} \tag{26}$$

Now, letting $z = (y^\nu - 1)/\nu$, we get $dz = y^{\nu-1}dy$ by the change-of-variables formula. Hence the density in terms of untransformed earnings, y, becomes equal to (8). This completes the proof. ■

Proof of Corollary 2.

Similarly to (5), we can write

$$\varepsilon_2(r) = \theta\varepsilon_1(r) + \widetilde{\varepsilon}_2(r),\ \eta(r) = \rho\varepsilon_1(r) + \widetilde{\eta}(r), \tag{27}$$

where $(\widetilde{\varepsilon}_2(r), \widetilde{\eta}(r)')$ has covariance matrix Σ and is independent of $\varepsilon_1(r)$. Using (9) and (27), we then obtain

$$\varepsilon_2^j(r) = \kappa_{jr}\theta\varepsilon_1(r) + \chi_{jr} + \widetilde{\varepsilon}_2^j(r), \tag{28}$$

where $\widetilde{\varepsilon}_2^j(r) = \kappa_{jr}\widetilde{\varepsilon}_2(r)$ is independent of $\varepsilon_1(r)$ and the covariance matrix of $(\widetilde{\varepsilon}_2^j(r), \widetilde{\eta}(r)')$ is $\Sigma_{jr}(r)$ – defined in (14). We realize that when $R = r$ and $J = j$, the proof of Corollary 2 is completely analogous to the proof of Theorem 1, with $(\varepsilon_1, \varepsilon_2, \eta')$ and $(\widetilde{\varepsilon}_2, \widetilde{\eta}')$ now being replaced by $(\varepsilon_1(r), \varepsilon_2^j(r), \eta(r)')$ and $(\widetilde{\varepsilon}_2^j(r), \widetilde{\eta}(r)')$, respectively. The modifications in (12) compared to (8) occur because the mean of $\varepsilon_2^j(r)$ is χ_{jr}, Σ in (6) is replaced by $\Sigma_{jr}(r)$ (yielding (13)); and θ in (7) is replaced by $\kappa_{jr}\theta$ (yielding (15)).

■

PROOF OF COROLLARY 3.

Recall that $J = j \Leftrightarrow \varepsilon_1 \in K_j$. Using the rule of double expectation and (27), we obtain

$$\begin{aligned} E(\eta(R)|J=j) &= \sum_{r=1}^{Q} q_r E(\eta(r)|\varepsilon_1(r) \in K_j, R=r) = \sum_{r=1}^{Q} q_r E(\rho\varepsilon_1(r) + \widetilde{\eta}(r)|\varepsilon_1(r) \in K_j, R=r) \\ &= \rho E(\varepsilon_1(r)|\varepsilon_1(r) \in K_j) = \rho \frac{E\left(\varepsilon_1(r) 1\{\varepsilon_1(r) \in K_j\}\right)}{P(J=j)} = \rho \frac{\int_{\mu_{j-1}-Z_1\gamma_1}^{\mu_j - Z_1\gamma_1} u\phi(u)\, du}{P(J=j)} \\ &= -\rho \frac{\phi(\mu_j - Z_1\gamma_1) - \phi(\mu_{j-1} - Z_1\gamma_1)}{\Phi(\mu_j - Z_1\gamma_1) - \Phi(\mu_{j-1} - Z_1\gamma_1)} = -\rho\lambda(j), \end{aligned} \tag{29}$$

where, in the third equation, we have used that $\widetilde{\eta}(r)$ is independent of $\varepsilon_1(r)$ and R is independent of $(\varepsilon_1(r), \widetilde{\eta}(r)')$ for every r. This proves (18). Consider next the proof of (19). Similarly to (29), we obtain

$$\begin{aligned} E(\varepsilon_2(R)|J=j, R=r) &= E(\theta\varepsilon_1(r) + \widetilde{\varepsilon}_2(r)|\varepsilon_1(r) \in K_j, R=r) = \theta E(\varepsilon_1(r)|\varepsilon_1(r) \in K_j) \\ &= \theta \frac{E\left(\varepsilon_1(r) 1\{\varepsilon_1(r) \in K_j\}\right)}{P(J=j)} = -\theta\lambda(j), \end{aligned} \tag{30}$$

where we used that $\widetilde{\varepsilon}_1(r)$ is independent of $\varepsilon_1(r)$ and R is independent of $(\varepsilon_1(r), \widetilde{\varepsilon}_2(r))$. From (9) and the rule of double expectation, we obtain

$$\begin{aligned} E(\varepsilon_2^k(R)|J=j) &= E\left(\kappa_{kR}\varepsilon_2(R) + \chi_{kR}|J=j\right) \\ &= \sum_{r=1}^{Q} q_r E\left(\kappa_{kR}\varepsilon_2(R) + \chi_{kR}|J=j, R=r\right) \\ &= \sum_{r=1}^{Q} q_r \left(\kappa_{kr} E(\varepsilon_2(R)|J=j, R=r) + \chi_{kr}\right) \\ &= -\left(\sum_{r=1}^{Q} q_r \kappa_{kr}\right) \theta\lambda(j), \end{aligned}$$

where, in the last equality, we used (30) and the summation restriction (16). This completes the proof of (19). ■

Figures and tables

Table 1: **Parameter estimates of earnings equation for different model specifications. Standard errors in parentheses**

	Model specification			
	Model 1	**Model 2**	**Model 3**	**Model 4**
Earnings:	$\ln y$	$\ln y$	$\ln y$	$(Y^\nu - 1)/\nu$
Schooling:	Linear	Box-Cox	Splines	Splines
Experience:	Quadratic	Box-Cox	Splines	Splines
ν	0 (-)	0 (-)	0 (-)	-.17 (.003)
β_1	.08 (.003)	.66 (.09)	.06 (.005)	.01 (.001)
β_2	-.0005 (.0001)	.21 (.01)	.01 (.001)	.05 (.005)
SD(ε_2)	.26 (.01)	.51 (.10)	.24 (.03)	.05 (.001)
SD(η_1)	.01 (.004)	.08 (.02)	.01 (.001)	.001 (.0002)
SD(η_2)	.0003 (.00004)	.13 (.02)	.01 (.001)	.03 (.004)
Corr($\varepsilon_2, \varepsilon_1$)	-.12 (.05)	-.34 (.06)	-.25 (.05)	-.25 (.04)
Corr(η_1, ε_1)	.03 (.24)	.39 (.06)	.35 (07)	.47 (.16)
Corr(η_1, ε_2)	.92 (.13)	-.82 (.06)	-.13 (.23)	.30 (.16)
Corr(η_2, ε_1)	-.04 (.07)	-.01 (.03)	-.03 (.03)	.06 (.05)
Corr(η_2, ε_2)	-.62 (.07)	-.93 (.02)	-.48 (.11)	-.15 (.08)
Corr(η_2, η_1)	-.87 (.17)	.77 (.06)	.64 (.11)	-.77 (.15)
$\alpha_{1,1}$		11.26 (1.39)		
$\alpha_{1,2}$	1 (-)	.28 (.03)		
$\alpha_{2,1}$	-29.5 (2.48)	2.49 (.62)		
$\alpha_{2,2}$	2 (-)	.002 (.002)		
log-likelihood	-27274	-27251	-27179	-27058
Sample size	29,332	29,332	29,332	29,332

Table 2: **Descriptive statistics for schooling, experience and earnings**

Variable	**Mean**	**St.dev.**	**Min**	**Max**
Years of schooling	12.2	2.3	7	18.0
Years of experience	15.2	5.9	0	29.0
Log of earnings	7.7	0.3	6.4	9.8

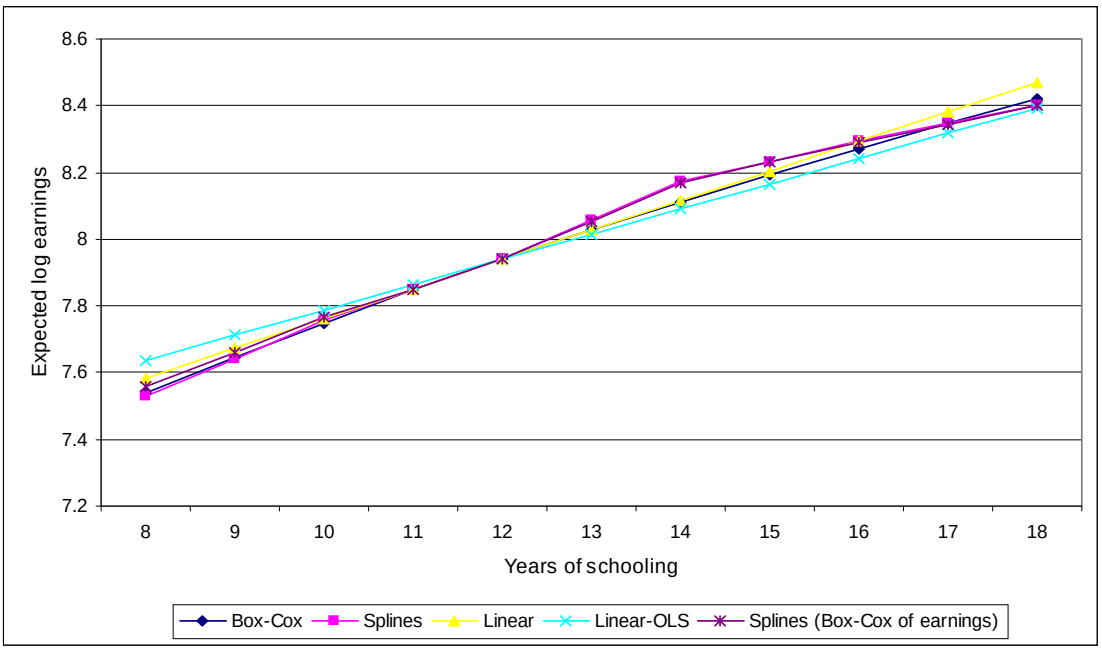

Figure 1: **Expected log earnings as a function of years of schooling**

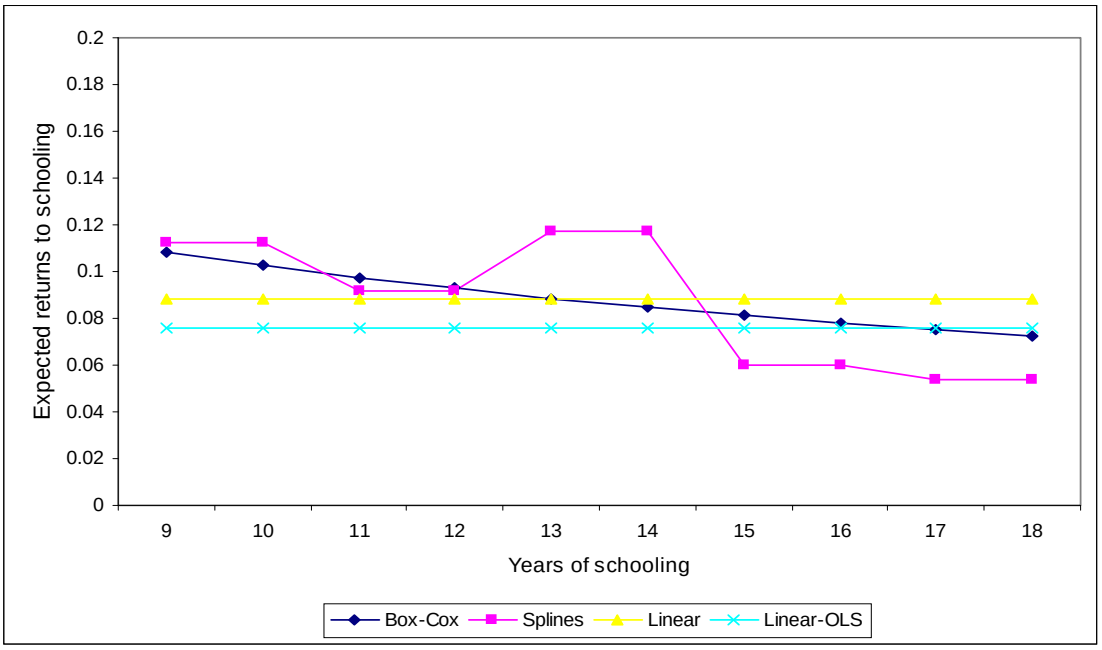

Figure 2: **Expected marginal returns to schooling (ATE)**

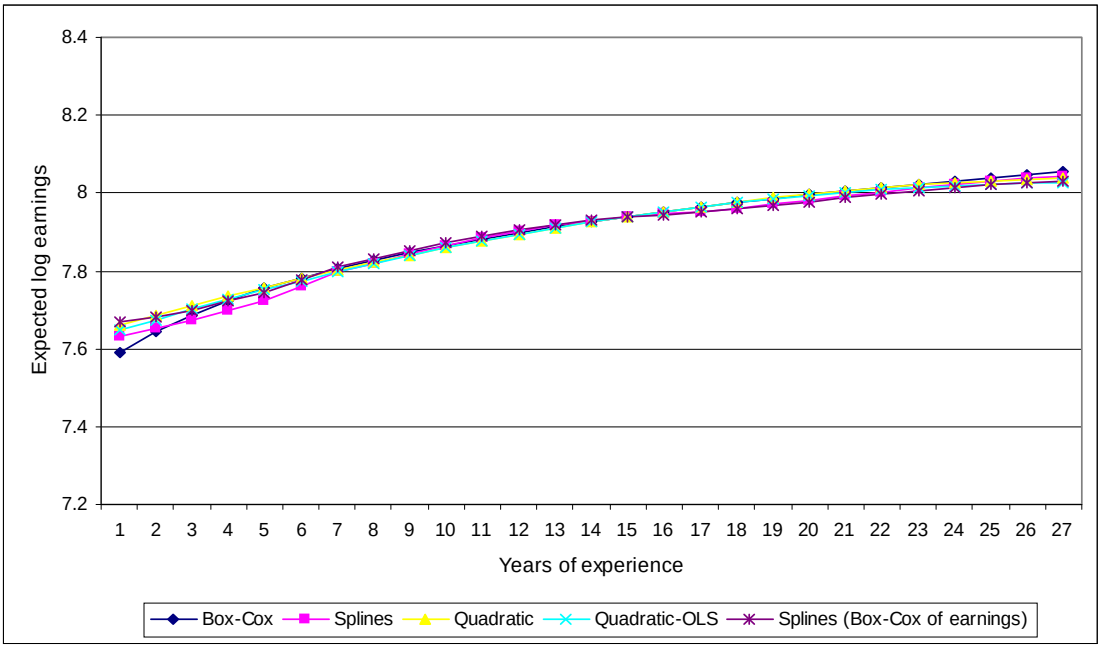

Figure 3: **Expected log earnings as a function of experience**

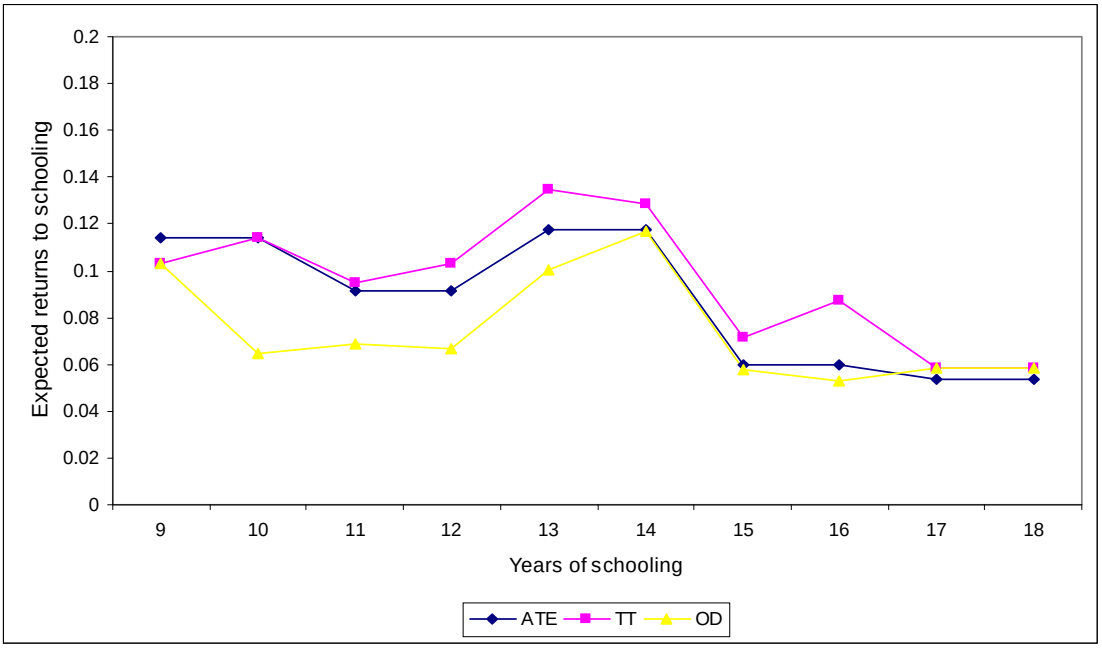

Figure 4: **Estimates of marginal treatment effects. Models with normally distributed error terms**

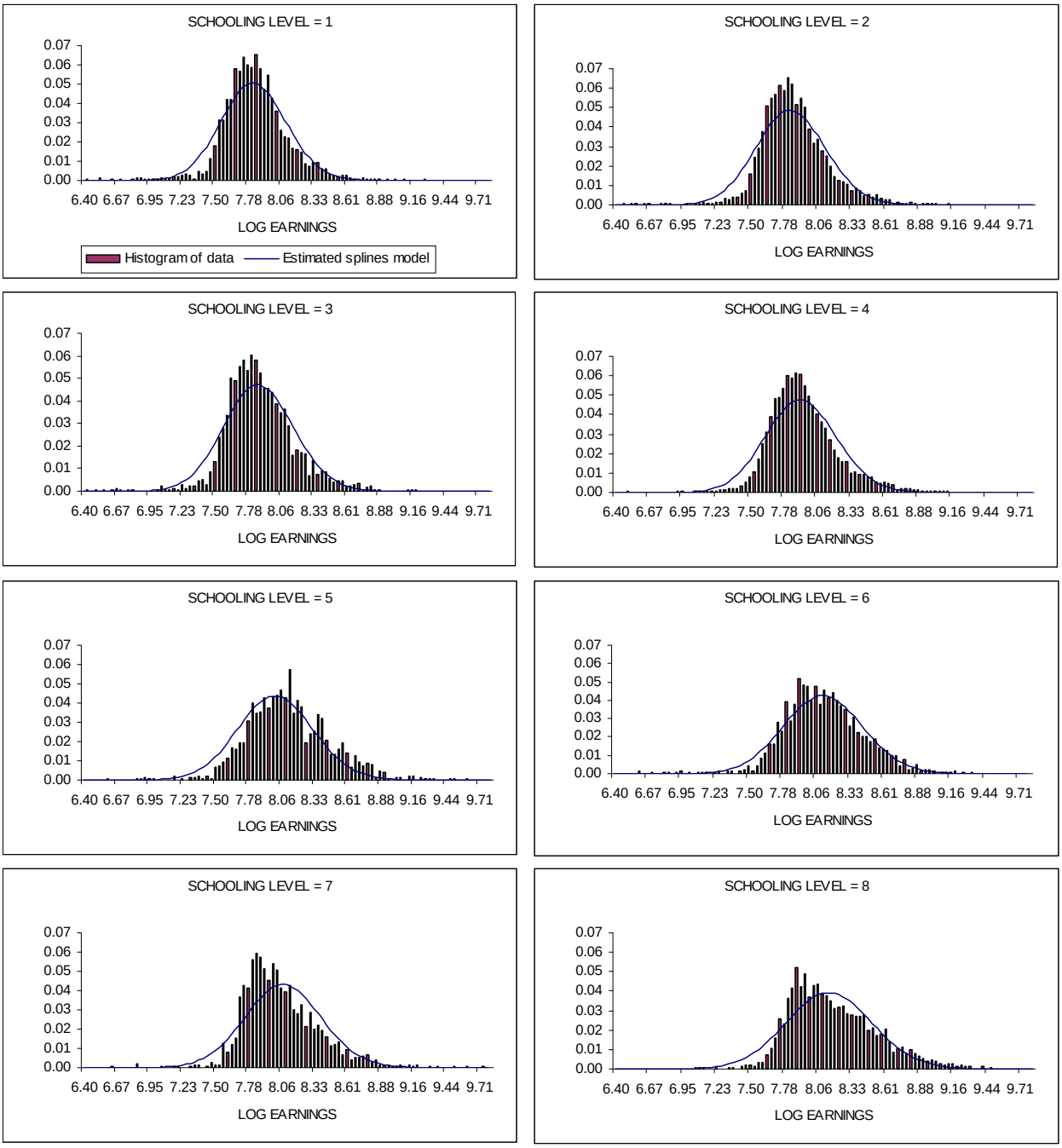

Figure 5: **Normally distributed error terms: The conditional distributions of log earnings given the level of schooling.** Estimated probability distribution functions and histograms of log earning data

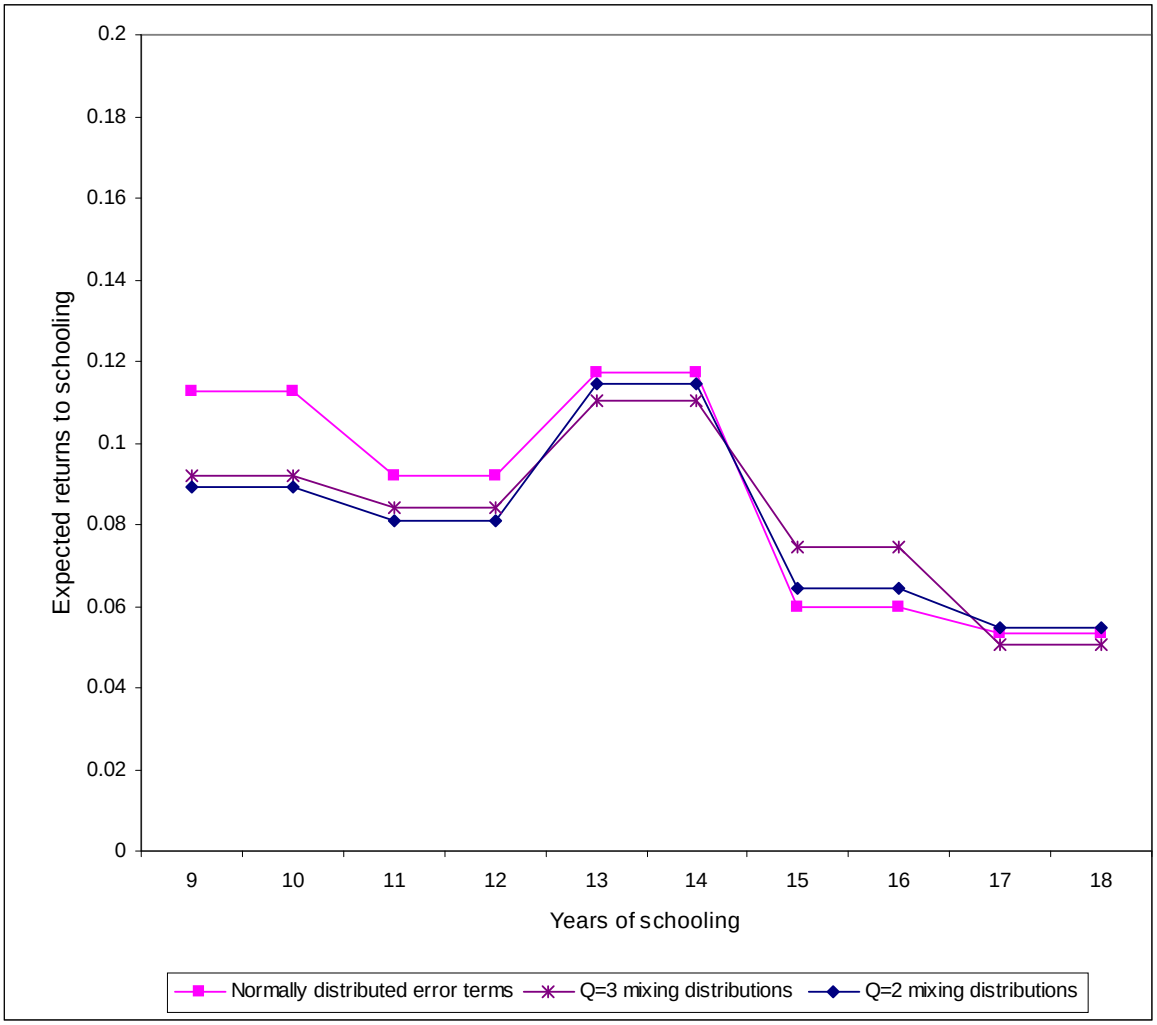

Figure 6: **Comparing estimates of the average treatment effect (ATE): Model with normally distributed error terms and two mixture models. Spline transformation of years of schooling**

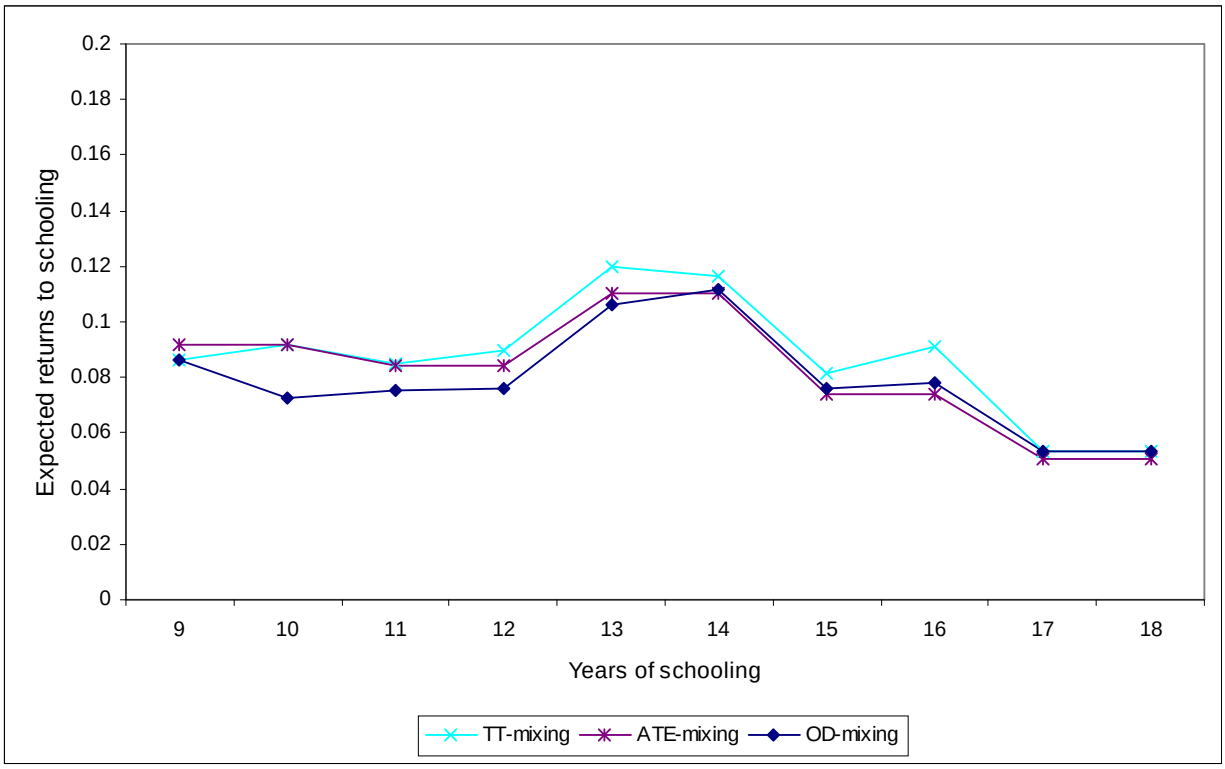

Figure 7: **Estimates of different treatment effects: Model with mixed multinormally distributed error terms (Q=3)**

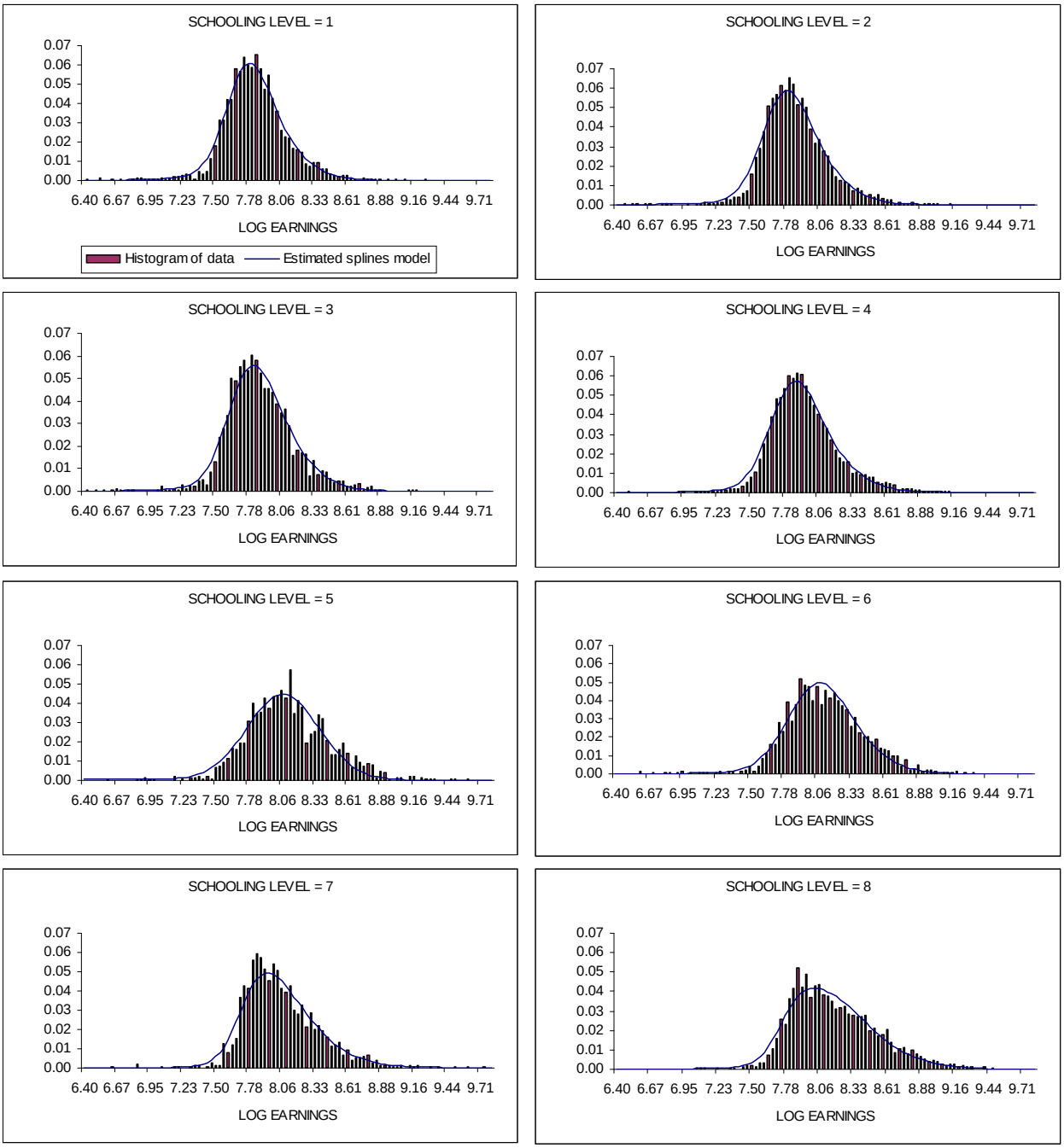

Figure 8: **Mixed multinormally distributed error terms (Q=3): The conditional distributions of log earnings given the level of schooling.** Estimated probability distribution functions and histograms of log earning data

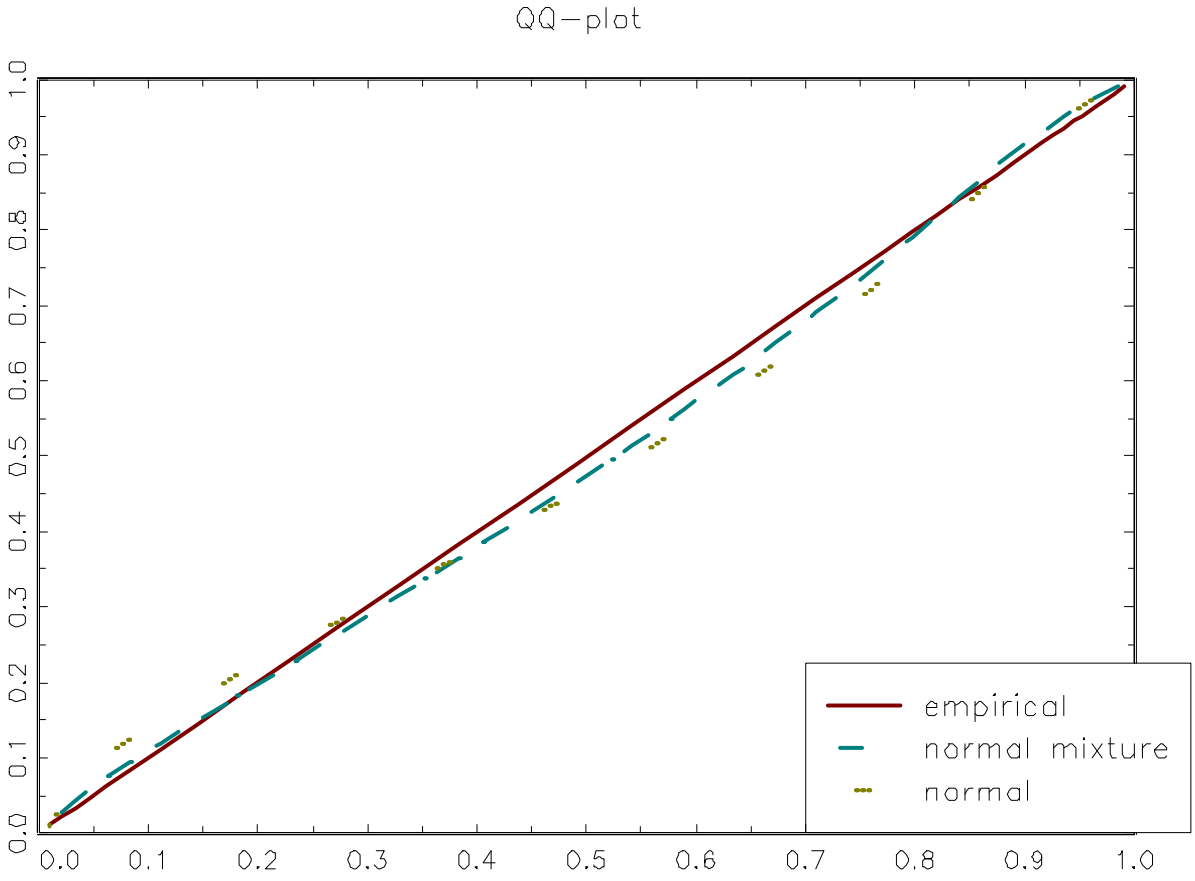

Figure 9: **Comparing estimated and empirical distribution functions: QQ-plot for the marginal distribution of log earnings**

Table 3: **Descriptive statistics for Z_1 with corresponding parameter estimates**

Z1-variables	**Mean**	**St.dev**	**Min**	**Max**	**Parameter estimate**	**S.E.**
Lone mother	0.05	0.21	0	1	0.27	0.07
Lone father	0.01	0.08	0	1	-0.05	0.08
No parents	0.01	0.09	0	1	-0.13	0.07
Mother working	0.33	0.47	0	1	0.02	0.01
Father working	0.93	0.25	0	1	0.07	0.04
Family income:						
quintile 2	0.20	0.40	0	1	0.09	0.02
quintile 3	0.21	0.41	0	1	0.15	0.02
quintile 4	0.21	0.41	0	1	0.23	0.02
quintile 5	0.20	0.40	0	1	0.32	0.03
Mother's schooling:						
lower secondary	0.15	0.35	0	1	0.35	0.02
upper secondary	0.07	0.25	0	1	0.45	0.03
lower tertiary	0.04	0.19	0	1	0.71	0.04
upper teritary	0.00	0.06	0	1	0.88	0.13
Father's schooling:						
lower secondary	0.14	0.35	0	1	0.31	0.02
upper secondary	0.13	0.34	0	1	0.39	0.02
lower tertiary	0.07	0.26	0	1	0.69	0.03
upper teritary	0.04	0.19	0	1	0.99	0.04
Born abroad	0.00	0.02	0	1	-0.31	0.27
Father born abroad	0.06	0.24	0	1	-0.08	0.05
Mother born abroad	0.03	0.16	0	1	-0.08	0.04
Østfold	0.06	0.24	0	1	0.03	0.03
Akershus	0.09	0.28	0	1	-0.04	0.03
Hedmark	0.04	0.20	0	1	0.07	0.04
Oppland	0.04	0.20	0	1	0.14	0.04
Buskerud	0.05	0.22	0	1	0.04	0.03
Vestfold	0.05	0.21	0	1	0.09	0.04
Telemark	0.04	0.19	0	1	0.14	0.04
A-Agder	0.02	0.14	0	1	0.24	0.05
V-Agder	0.04	0.19	0	1	0.22	0.04
Rogaland	0.08	0.28	0	1	0.11	0.03
Hordaland	0.11	0.31	0	1	0.20	0.03
Sogn Fj.	0.03	0.17	0	1	0.36	0.04
Møre Roms.	0.07	0.25	0	1	0.26	0.03
S-Tr.	0.06	0.24	0	1	0.21	0.03
N-Tr.	0.03	0.18	0	1	0.42	0.04
Nordland	0.06	0.24	0	1	0.34	0.03
Troms	0.03	0.18	0	1	0.24	0.04
Finnmark	0.02	0.13	0	1	0.27	0.05

Table 4: **Descriptive statistics for Z_2 with corresponding parameter estimates**

Z2-variables	**Mean**	**St.dev.**	**Min**	**Max**	**Parameter estimate**	**S.E.**
Intercept					7.33	0.04
Manufacturing					0	-
Public services	0.28	0.45	0	1	-0.20	0.03
Private services	0.40	0.49	0	1	-0.04	0.03
Unspecified	0.00	0.04	0	1	-0.23	0.04
General	0.19	0.40	0	1	0.05	0.03
Humanities	0.03	0.18	0	1	-0.10	0.04
Teaching	0.06	0.23	0	1	-0.14	0.05
Technical					0	-
Business/administrative	0.21	0.41	0	1	0.02	0.02
Transport	0.03	0.17	0	1	0.03	0.01
Health	0.05	0.23	0	1	0.07	0.00
Farming/fisheries	0.02	0.13	0	1	-0.06	0.05
Services/military	0.06	0.23	0	1	0.08	0.01
Østfold	0.05	0.21	0	1	-0.15	0.01
Akershus	0.09	0.28	0	1	-0.03	0.01
Oslo					0	-
Hedmark	0.03	0.18	0	1	-0.19	0.00
Oppland	0.03	0.18	0	1	-0.22	0.01
Buskerud	0.05	0.21	0	1	-0.11	0.01
Vestfold	0.04	0.19	0	1	-0.14	0.01
Telemark	0.03	0.17	0	1	-0.13	0.01
A-Agder	0.02	0.13	0	1	-0.18	0.01
V-Agder	0.03	0.17	0	1	-0.14	0.01
Rogaland	0.08	0.27	0	1	-0.02	0.01
Hordaland	0.09	0.29	0	1	-0.11	0.01
Sogn Fj.	0.02	0.15	0	1	-0.18	0.01
Møre Roms.	0.05	0.22	0	1	-0.16	0.01
S-Tr.	0.06	0.23	0	1	-0.18	0.01
N-Tr.	0.02	0.15	0	1	-0.23	0.01
Nordland	0.05	0.21	0	1	-0.20	0.01
Tromsø	0.03	0.18	0	1	-0.16	0.01
Finnmark	0.02	0.12	0	1	-0.21	0.01

Table 5: **Parameter estimates of earnings equation for two normal mixture models: Q=2 and Q=3. Standard errors in parentheses**

	Mixture distribution:	
	$Q = 2$	$Q = 3$
Earnings:	$\ln y$	$\ln y$
Schooling:	Splines	Splines
Experience:	Splines	Splines
β_1	0.003 (.00006)	0.002 (.00005)
β_2	1.61 (.21)	0.58 (.06)
SD ($\varepsilon_2^J(R)$)	0.31 (.029)	0.28 (.01)
Coeff. of skewness $\varepsilon_2^J(R)$	0.15 (.08)	0.15 (.07)
Coeff. of kurtosis $\varepsilon_2^J(R)$	4.99 (.86)	6.02 (.74)
SD(η_1)	0.0005 (.00005)	0.0003 (.00005)
SD(η_2)	0.63 (.05)	0.19 (.01)
Corr($\varepsilon_2^J(R), \varepsilon_1$)	−0.08 (.03)	−0.10 (.04)
Corr(η_1, ε_1)	0.41 (.13)	0.22 (.07)
Corr($\eta_1, \varepsilon_2^J(R)$)	0.84 (.08)	0.82 (.10)
Corr(η_2, ε_1)	−010 (.03)	−0.11 (.03)
Corr($\eta_2, \varepsilon_2^J(R)$)	−0.92 (.03)	−0.80 (.01)
Corr(η_2, η_1)	0.86 (.07)	−0.45 (.15)
log-likelihood	−25810	−25579
Sample size	29, 332	29, 332

Table 6: **Estimates of mixture parameters when Q=3. Standard errors in parentheses**

Parameter	**Mixture distribution (r):**		
	$r = 1$	$r = 2$	$r = 3$
q_r	0.80	0.19 (.013)	0.01 (.001)
κ_{1r}	1.24	0.00 (−)	0.15 (.17)
κ_{2r}	1.24	0.00 (−)	0.00 (−)
κ_{3r}	1.20	0.00 (−)	2.39 (.35)
κ_{4r}	0.71	2.13 (.09)	2.50 (.25)
κ_{5r}	0.13	2.95 (.20)	24.9 (4.82)
κ_{6r}	0.80	1.74 (.18)	2.32 (.48)
κ_{7r}	1.22	0.00 (−)	1.44 (.45)
κ_{8r}	0.75	1.92 (.12)	2.97 (.61)
χ_{1r}	0.05	0.27 (.03)	-0.73 (.05)
χ_{2r}	0.04	0.26 (.02)	-0.88 (.05)
χ_{3r}	0.05	0.27 (.03)	-0.77 (.08)
χ_{4r}	0.05	0.26 (.03)	-0.74 (.07)
χ_{5r}	-0.08	0.21 (.05)	-7.75 (1.05)
χ_{6r}	0.05	0.28 (.03)	-0.90 (.12)
χ_{7r}	-0.01	-0.11 (.04)	0.91 (.08)
χ_{8r}	0.06	0.30 (.03)	-0.69 (.17)

* All estimates for $r = 1$ determined by summation restrictions, cf. (16) and (17)