Estimation of a model for matched panel data with high-dimensional two-way unobserved heterogeneity[*]

Øivind A. Nilsen[†] Arvid Raknerud[‡] and Terje Skjerpen[§]

August 22, 2016

ABSTRACT: We consider a model for matched data with two types of unobserved effects: a random effect related to the main observational unit and a random or fixed effect related to a secondary unit to which the main unit is matched. In typical applications, e.g. on registry data, there is a curse of dimensionality which we propose to mitigate using an iterative feasible GLS approach on variables subjected to the Helmert transformation. Control functions allow for correlation between the explanatory variables and the random effects. This approach is illustrated by a wage equation with unobserved individual- and firm-specific effects and an endogenous years-of-schooling variable.

JEL classification: C23, C81, J31

Keywords: Matched employer–employee data, Helmert transformation, Random effects, Wage equation, Iterative feasible GLS

[*] We are grateful for valuable comments from an editor and the referee. The usual disclaimer applies.

[†] Corresponding author: Norwegian School of Economics. E-mail: oivind.nilsen@nhh.no.

[‡] Statistics Norway, Research Department. E-mail: arvid.raknerud@ssb.no.

[§] Statistics Norway, Research Department. E-mail: terje.skjerpen@ssb.no.

1 Introduction

Access to matched data sets enables consideration of unobserved heterogeneity corresponding to different types of units in regression analyses. Often the main focus is on one type of observational unit, while it is also necessary to account for unobserved heterogeneity caused by another type of observational unit that is matched to the main type. Wage modeling by means of matched employer-employee data may be the best known example. Here, the individual is considered the main observational unit, and the firm to which the individual is matched has the role of a secondary observational unit. The use of the two dimensional unobserved effects in panel data models is not limited to labour market applications. Other examples are bank-customers, student-teachers, and patients-general practitioners (see Ioannidou and Ongena, 2010; Rockoff, 2004; Biørn and Godager, 2010).

An important choice to make in panel data analysis with two types of observational units is how to specify unobserved time-invariant effects related to the primary and secondary type of units, i.e., whether they should be treated as fixed or random. Abowd et al. (1999), whose paper contributes seminally to wage modeling using employer-employee data, represent both unobserved individual- and firm-specific heterogeneity by fixed effects. Following Abowd et al. (1999), it is common in this literature to assume that both the unobserved effects are fixed.[1]

There are few examples in the literature of models for matched observation units where unobserved heterogeneity in both dimensions is represented by random effects. Notable exceptions are Woodcock (2008, 2015), who estimates a model with

[1] See for instance the two computer oriented articles by Cornelissen (2008), and Guimarães and Portugal (2010).

unobserved person, firm and match effects – all of which are assumed to be random – using what is labeled a 'hybrid mixed effect estimator'. Other contributions include Abowd and Kramarz (1999), Abowd et al. (2008), Dostie (2011) and Sørensen and Vejlin (2013). Dostie (2011), having access to data where each worker is observed in only one firm, did not have the option to choose a specification with fixed individual and firm effects, using instead a random effects specification. Thus, model specifications involving random individual and random firm effects are less data demanding than models involving fixed individual and fixed firm effects.

There are, however, potential problems related to the estimation of random effects models. One is related to dimensionality, and therefore computer-memory requirements. It is evident that matched registry data include several thousands of observational units, which again are matched with thousands of another type of observational units. Thus, when the model with two-way unobserved heterogeneity is estimated, one may end up with having to invert very large matrices, which may not be computationally feasible in terms of memory and reasonable computing time. Another problem is that the (pure) random effects specification imposes orthogonality between the unobserved time-invariant variable and the observed explanatory variables, which may lead to biased estimates of the slope parameters of the model.

A fundamental problem related to fixed effects models is that the coefficients corresponding to time-invariant individual specific explanatory variables are not identified. Within the framework of two-way fixed effects, e.g. Abowd et al. (1999), a two-step procedure is usually applied to identify the effects of such explanatory variables: First estimate a fixed effects model using only individual time-varying covariates. Then run an auxiliary regression of estimated fixed effects on individual-specific variables and the individual means of time-varying variables. This is called the fixed effects vector decomposition (FEVD) estimator and is applied in many empirical studies and advocated inter alia by Plümper and Troeger (2011). However, Greene (2011) and Greene (2012, pp. 364–370) make clear that the FEVD estimator is based on implicit exogeneity assumptions which are somewhat different from those employed by Hausman and Taylor (1981) in their instrumental variable approach. The implicit exogeneity assumption used in conjunction with the FEVD estimator is that the time-invariant observed variables are uncorrelated with the unobserved individual effects; only the time-varying variables are allowed to be correlated with the unobserved individual specific effect.[2]

Our paper entails two distinctive features that makes it different from earlier contributions using matched panel employer-employee data. The first is related to computational aspects. We transform our econometric relation using a backward orthogonal deviations operator, also known as the 'Helmert transformation', which sweeps out the unobserved effects corresponding to N main observation units (e.g., individuals).[3] Such a transformation does not distort the orthogonality property of the (transformed) genuine error terms. We show that the dimension reduction brought about by the Helmert transformation facilitates application of an iteratively feasible GLS (IFGLS) estimator. Hence, the transformation contributes to a simplification of the maximization problem that needs to be solved for obtaining parameter estimates. As far as we know, the Helmert transformation has not been utilized before when analyzing matched employer-employee panel data.

[2] Breusch et al. (2011) have also questioned the transparency and gain of the fixed effects vector decomposition. The articles by Breusch et al. (2011), Greene (2011) and Plümper and Troeger (2011) formed part of the Symposium on Fixed-Effect Vector Decomposition.

[3] As mentioned by Watson (2006), the Helmert transformation originates from geodesy. Balestra and Krishnakumar (2008) and Arellano and Bover (1995) comment on this transformation even though they do not use the label 'Helmert transformation'. Rather they refer to it as 'the backward and forward orthogonal deviations operator'. See also Keane and Runkle (1992) for the related concept of forward filtering.

The other distinctive feature is that we apply a control function approach to account for correlation between the time-invariant unobserved effects of the primary unit and the observed right-hand side variables. In our wage-equation application, where most of the observed right hand side variables are individual-specific, the Hausman-Taylor framework is not helpful. To remedy a potential endogeneity problem related to the main explanatory variable – education length – we use a control function approach based on the assumption that the choice of education length follows an ordered probit model, with some of the explanatory variables excluded from the wage equation. The control function captures the correlation between educational length and the unobserved individual-specific effect and enables us to relax the orthogonality assumption of the classical random effects model. This approach has previously not been applied in a setting with matched employer-employee panel data. With respect to unobserved time-invariant firm effects we consider both a fixed and a random effects specification.

The rest of the paper is organized as follows. In Section 2, we outline the general modeling framework and introduce the Helmert transformation. This transformation enables dimensionality reduction and facilitates the application of an IFGLS routine for estimation of the unknown parameters. We furthermore demonstrate how to control for correlation between individual time-invariant explanatory variables and random effects using a control function approach. In Section 3 we illustrate how the econometric framework can be applied in a wage equation setting. Section 4 provides some concluding remarks.

2 The general model

Let $i \in \{1, ..., N\}$ denote the main observation unit and $j \in \{1, ..., M\}$ denote the secondary unit. The unit, j, that is linked to i at t is conceptualized through a link function: $j = J(i, t)$. Adopting the notation of Abowd et al. (2008, p. 733) for a general linked linear model, the starting point of our analysis is the following regression equation:

$$y_{it} = x_{it}\beta + z_i\gamma + q_{J(i,t),t}\rho + \mu_i + \nu_{J(i,t)} + \eta_{it}, \tag{1}$$

where y_{it} is the dependent variable. Then x_{it} is a $1 \times p$ vector of time-varying covariates of the main unit, i, z_i is a $1 \times q$ vector of time-invariant covariates and $q_{J(i,t),t}$ is a $1 \times r$ vector of time-varying covariates of the secondary unit linked to i at t, i.e. $J(i, t)$. In matched employer–employee data, $J(i, t)$ will typically denote the firm where individual i is employed in period t.[4] For simplicity, we will henceforth refer to the main unit as an "individual" and the secondary unit as a "firm".

There are three types of unobserved components in (1): (i) The individual effect, μ_i, (ii) the firm effect, $\nu_{J(i,t)}$ (corresponding to the firm matched to i at t) and (iii) η_{it} – the genuine error term. The unobserved component attached to the individual, μ_i, is involved irrespective of the firm where the individual is working and covers inter alia intelligence of the inividual. The unobserved component attached to a given firm, j, equals ν_j and is shared by all the individual working in a specific firm.

[4] The adopted standard in the matched employer-employee data literature measures sorting as the extent to which high wage workers are found in high wage firms, conditional on observable characteristics. That means that sorting in these analyses is taken as given and not modelled explicitly. More recent empirical literature, often based on the theoretical models by Shimer and Smith (2000) or Shimer (2005), has started to develop matching models in which the sorting of workers into firms is modeled more explicitly (see for instance Postel-Vinay and Robin, 2002; Lopes de Melo, 2009; Le Maire and Scheuer, 2013; Abowd et al., 2014; and Bagger and Lentz, 2014). Our focus in this paper, however, is more on the econometric methodology, so we follow the adopted standard and assume the employer-employee matching is outside the model.

Note the important distinction between ν_j and $\nu_{J(i,t)}$: ν_j is the effect corresponding to a *given* firm, whereas $\nu_{J(i,t)}$ is the effect corresponding to the firm *matched* to i at t. Thus, whereas the underlying firm-effect ν_j is time-invariant, $\nu_{J(i,t)}$ will change when the match of individual i changes.[5,6]

We consider different types of specifications for μ_i and ν_j. First, ν_j is allowed to be either a random or a fixed effect. Second, μ_i is allowed to be either a standard random effect, or a random effect correlated with z_i. Of course, if the unobserved individual effect, μ_i, is correlated with z_i, treating it as a standard random effect yields biased estimates of γ. We therefore propose an IV/control function approach in Section 2.4.

The starting point of our analysis is the following standard assumptions: For all i and t: $E(\eta_{it}) = 0$, $E(\eta_{it}\eta_{is}) = 0$ for $t \neq s$, and $E(\eta_{it}^2) = \sigma_{\eta\eta}$. Let $\nu = (\nu_1, ..., \nu_M)'$ denote the vector of all the M firm effects and G_{it} the $1 \times M$ design matrix indicating which firm is matched to individual i at t:

$$G_{it}\nu = \nu_{J(i,t)}.$$

That is

$$G_{it} = \underbrace{\begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}}_{\text{position } J(i,t)}. \tag{2}$$

Then we can re-write equation (1) as;

$$y_{it} = x_{it}\beta + z_i\gamma + q_{J(i,t),t}\rho + \mu_i + G_{it}\nu + \eta_{it}. \tag{3}$$

[5] To make this clearer, assume individual i works in two different firms: $j = 2$ in years $t = 1,...,4$, and firm $j = 7$ in years $t = 5,...,9$. As $J(i,t)$ denotes the firm matched with individual i at time t, $J(i,t) = 2$ in years $t = 1,...,4$, and $J(i,t)$ =7 in years $t = 5,...,9$. Furthermore, $v_{J(i,t)} = v_2$ for $t = 1,...,4$, and $v_{J(i,t)} = v_7$ for $t = 5,...,9$, and $q_{J(i,t)t}$ is the vector of time-varying covariates collected for the relevant firm j in year t.

[6] As mentioned earlier, Woodcock (2008, 2015) also includes unobserved match effects, picking up the value of match quality. He finds that the conclusions are rather different when using models including match effects with models without this type of effect. We return to this in Section 3.

To reduce the number of latent variables in the model, we apply the Helmert transformation (see Lütkepohl, 1996, p. 249). Formally, the Helmert transformation of any $1 \times m$ *row vector* H_{it}, $t = 1, ..., T_i$, is given by $\overrightarrow{H}_{i,1}, ..., \overrightarrow{H}_{i,T_i}$, where

$$\overrightarrow{H}_{i,t} = \sqrt{t/(t+1)} \left(H_{i,t+1} - (1/t) \sum_{s=1}^{t} H_{is} \right), \; t = 1, ..., T_i - 1,$$

with the last observation on unit i being at $t = T_i$,[7] and

$$\overrightarrow{H}_{i,T_i} = \overline{H}_i \equiv (1/T_i) \sum_{s=1}^{T_i} H_{is}.$$

For example, for the firm-variables, $q_{J(i,t),t}$, the Helmert transformation is:

$$\begin{aligned} \overrightarrow{q}_{i,t} &= \sqrt{t/(t+1)} \left(q_{J(i,t+1),t+1} - (1/t) \sum_{s=1}^{t} q_{J(i,s),s} \right), \; t = 1, ..., T_i - 1 \\ \overrightarrow{q}_{i,T_i} &= \overline{q}_i \equiv (1/T_i) \sum_{s=1}^{T_i} q_{J(i,s),s}. \end{aligned}$$

Applying the Helmert transformation to each term in (3), it is easy to check that the Helmert-transformed error terms, $\overrightarrow{\eta}_{i,t}$ (corresponding to $\overrightarrow{y}_{i,t}$) are uncorrelated over t, given that η_{it} are uncorrelated and homoscedastic (i.e., have constant variance over time). Moreover, $Var(\overrightarrow{\eta}_{i,t}) = \sigma_{\eta\eta}$ for $t < T_i$ and $Var(\overrightarrow{\eta}_{i,T_i}) = \sigma_{\eta\eta}/T_i$.

2.1 Independent random individual and random firm effects

Assume now that the vector of the random firm effects,

$$\nu = (\nu_1, ..., \nu_M)',$$

and the vector of individual effects,

$$\mu = (\mu_1, ..., \mu_N)',$$

[7] For notational simplicity, we assume that all individuals enter the sample at $t = 1$. This convention entails no loss of generality since t can be reinterpreted as the t'th observation of individual i.

are mutually independent and distributed as

$$\begin{aligned} \nu &\sim IID(0, \sigma_{\nu\nu} I_M) \\ \mu &\sim IID(0, \sigma_{\mu\mu} I_N), \end{aligned}$$

where I_p is the identity matrix of dimension p. Then we have the following relation:

$$Y = XB + u, \tag{4}$$

where $B = (\beta', \gamma', \rho')'$ and

$$u = G\nu + e,$$

with

$$\begin{aligned} Y &= \begin{bmatrix} \overline{y}_1 \\ \vdots \\ \overline{y}_N \\ \overrightarrow{y} \end{bmatrix}_{\sum_{i=1}^N T_i \times 1} \qquad X = \begin{bmatrix} \overline{x}_1 & \overline{q}_1 & z_1 \\ \vdots & & \vdots \\ \overline{x}_N & \overline{q}_N & z_N \\ \overrightarrow{X} & \overrightarrow{q} & 0 \end{bmatrix}_{\sum_{i=1}^N T_i \times (p+r+q)} \\ G &= \begin{bmatrix} \overline{G}_1 \\ \vdots \\ \overline{G}_N \\ \overrightarrow{G} \end{bmatrix}_{\sum_{i=1}^N T_i \times M} \qquad e = \begin{bmatrix} \bar{\eta}_1 + \mu_1 \\ \vdots \\ \bar{\eta}_N + \mu_N \\ \overrightarrow{\eta} \end{bmatrix}_{\sum_{i=1}^N T_i \times 1.} \end{aligned} \tag{5}$$

The submatrices $\overrightarrow{y}$, $\overrightarrow{X}$,$\overrightarrow{q}$, $\overrightarrow{G}$, $\overrightarrow{\eta}$ in (5) are defined as follows:

$$\overrightarrow{y}= \begin{bmatrix} \overrightarrow{y}_{1,1} \\ \vdots \\ \overrightarrow{y}_{1,T_1-1} \\ \vdots \\ \overrightarrow{y}_{N,1} \\ \vdots \\ \overrightarrow{y}_{N,T_N-1} \end{bmatrix} \quad \overrightarrow{X}= \begin{bmatrix} \overrightarrow{x}_{1,1} \\ \vdots \\ \overrightarrow{x}_{1,T_1-1} \\ \vdots \\ \overrightarrow{x}_{N,1} \\ \vdots \\ \overrightarrow{x}_{N,T_N-1} \end{bmatrix} \quad \overrightarrow{q}= \begin{bmatrix} \overrightarrow{q}_{1,1} \\ \vdots \\ \overrightarrow{q}_{1,T_1-1} \\ \vdots \\ \overrightarrow{q}_{N,1} \\ \vdots \\ \overrightarrow{q}_{N,T_N-1} \end{bmatrix} \quad \overrightarrow{G}= \begin{bmatrix} \overrightarrow{G}_{1,1} \\ \vdots \\ \overrightarrow{G}_{1,T_1-1} \\ \vdots \\ \overrightarrow{G}_{N,1} \\ \vdots \\ \overrightarrow{G}_{N,T_N-1} \end{bmatrix} \quad \overrightarrow{\eta}= \begin{bmatrix} \overrightarrow{\eta}_{1,1} \\ \vdots \\ \overrightarrow{\eta}_{1,T_1-1} \\ \vdots \\ \overrightarrow{\eta}_{N,1} \\ \vdots \\ \overrightarrow{\eta}_{N,T_N-1} \end{bmatrix}.$$

Define $\overline{T} = \sum_{i=1}^{N} T_i/N$ and $\lambda_\mu = \sigma_{\mu\mu}/\sigma_{\eta\eta}$. Then the covariance matrix of the error term in (4) is:

$$Var(u) = \Sigma,$$

where

$$\Sigma = \sigma_{\nu\nu} GG' + \sigma_{\eta\eta} D, \tag{6}$$

with

$$D = \begin{bmatrix} \Omega & 0 \\ 0 & I_{(\overline{T}-1)N} \end{bmatrix} \tag{7}$$

and

$$\Omega = diag(T_1^{-1} + \lambda_\mu, ..., T_N^{-1} + \lambda_\mu).$$

The GLS estimator of B, for a given weighting matrix W, is:

$$\widehat{B} = (X'WX)^{-1}X'WY. \tag{8}$$

Moreover,

$$Var(\widehat{B}) = (X'WX)^{-1}X'W\Sigma WX(X'WX)^{-1}. \tag{9}$$

The optimal weighting matrix in (8) is therefore $W = \Sigma^{-1}$.

In matched employer-employee panel data models, the unobserved individual and firm effects are both often specified as fixed effects.[8] Then identification is caused by variation in the combination of individuals and firms over time. For instance the identification of the fixed firm effects are driven only by the individuals moving from one firm to another over time. In our approach, where none of the unobserved effects are necessarily assumed to be fixed, identification is based on the longitudinal and linked aspects of the data, in the combination with the parametric assumptions

[8] As emphasized by inter alia Hsiao (2003, p. 43), fixed and random effects have a common point of departure. Whereas fixed effects are related to conditional inference, random effects are related to unconditional inference.

embedded in the mixed model (see also Dostie, 2011), i.e. that both each individual and each firm are observed several years and that the individuals' characteristics change over time.

IFGLS To denote Σ as function of $\theta = (\sigma_{\eta\eta}, \lambda_\mu, \sigma_{\nu\nu})$, we use the notation $\Sigma(\theta)$. Let $B^{(n)}$ denote the GLS estimator obtained when using the weighting matrix $W^{(n)}$ in (8). IFGLS consists in generating a sequence $(W^{(n)}, B^{(n)})$, where the superscript (n) denotes iteration number, as follows:

$$W^{(n+1)} = \Sigma^{-1}(\theta^{(n)}),$$

where

$$\theta^{(n)} = \arg\max_{\theta} \; L(\theta, B^{(n)})$$

and $L(\theta, B)$ is the log-likelihood function under the assumption of normality of ν and e:

$$L(\theta, B) = -\frac{1}{2}\ln|\Sigma(\theta)| - \frac{1}{2}(Y - XB)'\Sigma(\theta)^{-1}(Y - XB). \tag{10}$$

Convergence of the iterative algorithm to a stationary point on the likelihood function $L(\theta, B)$ follows from Oberhofer and Kmenta (1974), cf. also Breusch (1987). If the model is misspecified, the IFGLS estimator of B is still consistent provided $E(y_{it}|x_{it}, z_i)$ is correctly specified (see Gourieroux and Monfort, 1995, Ch. 8.4). An estimator of the covariance matrix $Var(\widehat{B})$ that is robust to both autocorrelation, heteroscedasticity and non-normality can be calculated from the residuals, $\widehat{e}$, of the estimated model (using the general formula (9)).[9] The computational aspects of the IFGLS algorithm is considered below.

[9] Lack of normality implies that the p-values of different test statistics cannot be trusted. The normality assumption of the error components may be tested, cf. for instance Blanchard and Mátyás (1996) and Gilbert (2002), who address normality in the one-way panel data model.

Computational issues To obtain the optimal weighting matrix $W = \Sigma^{-1}$ in (8) (for given θ), we first define $\lambda_v = \sigma_{vv}/\sigma_{\eta\eta}$. Then – using the matrix inversion lemma (see Anderson and Moore, 1979, p. 138) –

$$\begin{aligned}\Sigma^{-1} &= \sigma_{\eta\eta}^{-1}\left[D^{-1} - D^{-1}G(\lambda_\nu^{-1}I_M + G'D^{-1}G)^{-1}G'D^{-1}\right] \qquad (11)\\ &= \sigma_{\eta\eta}^{-1}\left[D^{-1} - D^{-1}GP\right],\end{aligned}$$

with

$$P = \sigma_{\eta\eta}^{-1}VG'D^{-1} \qquad (12)$$

and

$$V = \sigma_{\eta\eta}\left(\lambda_\nu^{-1}I_M + G'D^{-1}G\right)^{-1}. \qquad (13)$$

Note that G has dimension $\overline{T}N \times M$ and Σ dimension $\overline{T}N \times \overline{T}N$, whereas D is a diagonal matrix of order $\overline{T}N$.

The main achievement of the Helmert-transformation is to reduce the problem of inverting the $N\overline{T} \times N\overline{T}$ covariance matrix Σ to a manageable problem of calculating – in opposite order – (11)-(13). First, the matrix to be inverted to obtain V in (13) consists of the $M \times M$ matrix $\lambda_\nu^{-1}I + G'D^{-1}G$. This is a highly sparse matrix due to the diagonality of D (a direct consequence of the Helmert transformation) and the fact that G is a sparse matrix.[10] Once V has been obtained, the calculation of P in (12), and then Σ^{-1} in (11) are computationally straightforward, as seen from these two equations.

To denote D, P and V (see Eqs. (7), (12) and (13)) as functions of θ, we use the

[10] The non-zero elements of $G'D^{-1}G$ only consist of terms $g_i'D^{-1}g_j$ $(i \neq j)$, where g_i is the i'th column of G. This corresponds to pairs of firms i and j with overlapping employees. In practice, only a very small fraction of the $M(M-1)/2$ pairs satisfies this condition, and the number of non-zero terms will be of order $O(M)$ rather than $O(M^2)$. As a consequence, the number of operations required to obtain V will typically be of order $O(M^2)$ rather than $O(M^3)$.

notation $D(\theta)$, $P(\theta)$ and $V(\theta)$. Then the IFGLS algorithm works as follows: Let (n) refer to iteration n and $B^{(1)}$ be given. For $n = 1, 2, ...;$

(i) Maximize $L(\theta, B^{(n)})$ with respect to θ using a quasi-Newton algorithm in combination with Proposition 2 in Appendix A to obtain the maximizer, $\theta^{(n)}$

(ii) Calculate $\Sigma(\theta^{(n)})$ and then $B^{(n+1)}$ from (8), using $W = \Sigma(\theta^{(n)})^{-1}$

(iii) Set $n = n+1$, and go to *(i)* unless $|B^{(n+1)} - B^{(n)}| < c$ for some tolerance level $c > 0$ and norm $|\cdot|$. In that case, set $\widehat{B} = B^{(n+1)}$.

The above algorithm gives IFGLS estimators of $B = (\beta', \gamma', \rho')'$ together with estimates of the variance parameters $\sigma_{\nu\nu}$, $\sigma_{\mu\mu}$ and $\sigma_{\eta\eta}$.

Another estimation method that is of relevance in our case is Restricted Maximum Likelihood (REML). By transforming the original data employing different contrasts one may formulate a log-likelihood in the transformed variables which only depends on second order parameters, that is in our case the variance of the random individual component, the variance of the random firm component and the variance of the genuine error term. The maximization of the log-likelihood in transformed variables yields REML estimates of these parameters. The first order parameters may be estimated by utilizing a GLS estimator.

A property of the Helmert-transformation is that it retains the distributional properties of the genuine error terms in the original model specification. This is not the case with REML. Besides IFGLS estimation utilizing the Helmert-transformation seems to be a better tool when it comes to handling computational issues related to large matrices.[11]

[11] Asymptotically, maximum likelihood estimation, in which one maximizes over all the unknown parameters simultaneously and REML will give the same estimates, cf. for instance Demidenko (2004, Ch. 3.6.3). It has been put forward that it may be advantageous to use REML rather than ML when one is faced with small sample issues, cf. for instance Fitzmaurice et al. (2004, Ch. 4.5).

2.2 Random individual effects and fixed firm effects

Assume now that only the individual effects are random, but that the firm effects are fixed. The model with fixed firm effects is a limiting case of the random effects model when λ_ν^{-1} approaches zero, which is equivalent to assuming a "diffuse" prior for the random firm effects.[12] When ν is a vector with fixed effects[13] in (4), the GLS estimator of ν, $\widehat{\nu}$, must be found simultaneously with $\widehat{B}$. The GLS estimator is the solution to:

$$\begin{bmatrix} X'WY \\ G'WY \end{bmatrix} = \begin{bmatrix} X'WX & X'WG \\ G'WX & G'WG \end{bmatrix} \begin{bmatrix} \widehat{B} \\ \widehat{\nu} \end{bmatrix} \tag{14}$$

The optimal weighting matrix is now $W = D^{-1}$, which is a diagonal matrix. This is in contrast to $W = \Sigma^{-1}$ in the model with both random individual and random firm effects. IFGLS then reduces to the problem of minimizing the log-likelihood function

$$L(\theta, \widehat{B}, \widehat{\nu}) = -\frac{TN}{2}\ln\sigma_{\eta\eta} - \frac{1}{2}\sum_{i=1}^{N}\ln(T_i^{-1}+\lambda_\mu) - \frac{1}{2\sigma_{\eta\eta}}(Y - X\widehat{B} - G\widehat{\nu})' D(\theta)^{-1}(Y - X\widehat{B} - G\widehat{\nu})$$

with respect to θ. Thus the numerical complexity is confined to solving (14). This is a sparse linear system of equations, for the reasons explained earlier.

2.3 Correlated individual effects (μ_i) and explanatory variables (z_i)

In the above model specifications, the unobserved individual-specific effect μ_i is a standard random effect (and hence uncorrelated with the explanatory variables x_{it} and z_i). We now consider the case where the row vector z_i can be partitioned as $z_i = (\chi_i, S_i)$, where χ_i and S_i are row vectors of exogenous and endogenous variables,

However, since we in our application have rather comprehensive data, small sample issues are not a great concern.

[12] See Francke et al. (2010) for more details about the relation between the fixed and random effects estimators.

[13] This can be interpreted as conditioning on the realized values of the unobserved firm effects.

respectively, the latter being correlated with μ_i. Similarly, let $\gamma' = (\gamma'_{\varkappa}, \gamma'_s)$ such that we can write our former equation (1) as

$$y_{it} = x_{it}\beta + \chi_i\gamma_{\varkappa} + S_i\gamma_s + q_{J(i,t),t}\rho + \mu_i + \nu_{J(i,t)} + \eta_{it}. \tag{15}$$

Two types of methods to deal with the endogeneity of S_i are feasible within our setup: First, the classic instrumental variables method, and second, a control function approach in the case where S_i only consists of a single binary or ordinal variable (e.g. level of schooling). The latter approach is in the tradition of Heckman (1979) and Garen (1984).

The IV approach First, consider the case where S_i is a vector of observed continuous variables determined by

$$S_i = \delta U_i + \varepsilon_i, \tag{16}$$

where ε_i a random vector with zero mean, δ is a fixed, unknown coefficient matrix and U_i is a column-vector of variables including some or all components of χ_i in addition to at least as many instrumental variables as there are components of S_i. As usual, the instrumental variables are variables excluded from χ_i and uncorrelated with the composite error term, $\mu_i + \nu_{J(i,t)} + \eta_{it}$, of (15). In general, we can write

$$\mu_i = \pi\varepsilon_i + \widetilde{\varepsilon}_i \tag{17}$$

where

$$\pi = Var(\varepsilon_i)^{-1}E(\varepsilon_i\mu_i) \tag{18}$$

and $\widetilde{\varepsilon}_i$ is independent of ε_i. Thus, the individual effect μ_i is correlated with the error term in (16), making S_i endogenous. We can write

$$E(\mu_i|S_i, U_i) = \pi\kappa(S_i, U_i), \tag{19}$$

with

$$\kappa(S_i, U_i) = S_i - \delta U_i. \tag{20}$$

Note that δ can be estimated directly from (16) and that we can re-express (15) as

$$y_{it} = x_{it}\beta + \chi_i\gamma_\varkappa + S_i\gamma_s + q_{J(i,t),t}\rho \; + \pi\kappa(S_i, U_i) + \varepsilon_i^* + \nu_{J(i,t)} + \eta_{it}, \tag{21}$$

where

$$\begin{aligned} \varepsilon_i^* &= \mu_i - E(\mu_i|S_i, U_i) \\ &= \pi\left(\varepsilon_i - \kappa(S_i, U_i)\right) + \widetilde{\varepsilon}_i. \end{aligned}$$

The term ε_i^* has the property that $E(\varepsilon_i^*|S_i, U_i) = 0$ and hence is a genuine random effect (uncorrelated with S_i).

Equation (21), which is a version of (1) with random individual effects uncorrelated with the explanatory variables, may be estimated using the techniques described above. It is a classic exercise to show that identification is achieved by imposing at least as many exclusion restrictions (variables included in U_i but not in χ_i) as the number of endogenous explanatory variables (the dimension of S_i).

The control function approach Next, assume that S_i is a (scalar) categorical variable with K possible categories; $S_i \in \{1, 2, ..., K\}$. We will consider an ordered probit model for the endogenous explanatory variable S_i. Thus S_i is related to a continuous latent variable S_i^* through the relation

$$S_i = s \text{ iff } \zeta_{s-1} < S_i^* < \zeta_s \;,\; s = 1, ..., K, \tag{22}$$

where $\{\zeta_s\}$ are unknown threshold parameters, except for $\zeta_0 = -\infty$ and $\zeta_K = \infty$. Furthermore, we assume that

$$S_i^* = \delta U_i + \varepsilon_i, \tag{23}$$

where the vector (ε_i, μ_i) is assumed to have a bivariate normal distribution with zero mean and a general covariance matrix, apart from the conventional identifying restriction that ε_i has unit variance. Equation (17) is still valid, with the additional assumption that $\widetilde{\varepsilon}_i$ is normally distributed. We then have the following result, which is analogous to (19)-(20) and similar to Heckman (1979):

Proposition 1 *$E(\mu_i|S_i = s, U_i) = \pi\kappa(s, U_i)$, where*

$$\kappa(s, U_i) = -\frac{\left(\phi(\zeta_s - \delta U_i) - \phi(\zeta_{s-1} - \delta U_i)\right)}{\Phi(\zeta_s - \delta U_i) - \Phi(\zeta_{s-1} - \delta U_i)}, \; s = 1, ..., K, \tag{24}$$

with $\phi(\cdot)$ and $\Phi(\cdot)$ denoting the density and cumulative distribution function, respectively, of an $\mathcal{N}(0, 1)$ variable.

Proof

From (17) and the independence of ε_i and $\widetilde{\varepsilon}_i$ it follows that

$$\begin{aligned} E(\mu_i|S_i &= s, U_i) = E(\pi\varepsilon_i + \widetilde{\varepsilon}_i|S_i = s, U_i) = E(\pi\varepsilon_i|S_i = s, U_i) \\ &= \pi E(\varepsilon_i|\zeta_{s-1} - \delta U_i < \varepsilon_i \leq \zeta_s - \delta U_i) = \pi \frac{\int_{\zeta_{s-1}-\delta U_i}^{\zeta_s - \delta U_i} \omega\phi(\omega)d\omega}{P(S_i = s)} \\ &= \pi\kappa(s, U_i). \end{aligned}$$

Equation (21) is still valid. Specifically, a conventional ordered probit analysis based on (22)-(23) yields estimates of the parameters $\zeta_1, ..., \zeta_{K-1}$ and the parameter vector δ.

3 Application: Wage equation estimation

We illustrate our modelling approach by estimating a wage equation, where we allow for correlation between the level of schooling and the individual effect, μ_i. The estimated equation is a version of (21) (see the previous section). The dependent variable, y_{it}, is given as the log of annual wage earnings for (full-time employee) i employed in firm $J(i,t)$ in year t. The endogenous explanatory variable level of schooling, is denoted S_i, with $S_i \in \{1, 2, ..., 9\}$. Level 1 corresponds to 10 years of schooling, which is the mandatory level in Norway, whereas the three last categories comprise longer tertiary education. The exogenous time-invariant variables, χ_i, are dummies for type of education and gender. The time-varying individual-specific exogenous variables, x_{it}, are powers of labour market experience (represented by potential experience) up to the third order, labour market area dummies and year dummies. Finally, the vector of time-varying firm-specific exogenous variables, q_{jt}, includes i) log of number of employees and ii) return on total assets.

The initial sample used in the application of our method includes 241,904 observations for 53,665 individuals. The sample covers the period 1995–2006 and is collected for individuals and firms in the Norwegian machinery industry (NACE 29). In total, there are 2,593 firms in the initial sample. We include only individuals whose annual earnings are between 50,000 and 3,500,000 NOK (fixed prices), that is, we exclude the one per cent highest and lowest annual earnings.[14] Potential experience is defined as age minus years of schooling minus seven years (school starting age). For those individuals whose length of education changed over the sample period, we retain only the observations with maximum length of education. The labour market area dummies are constructed utilizing information on characteristics such

[14] 1 Euro $\approx$ 8 NOK in the sample period.

as size and centrality.[15] Mainly workers with the following three types of education are represented in the chosen industry: education in “General Programs”, “Business and Administration” and “Natural Sciences, Vocational and Technical subjects (Sci & Tech)”. Only these categories are therefore represented by education-type dummies in the model. The earnings measure used is total annual taxable (full-time) labor income. Because the earnings measure reflects annual earnings, observations where employment relationships begin or are terminated within the actual year are excluded. Holders of multiple jobs and individuals who received unemployment benefits or participated in active labour market programs are also excluded. It is also required that each individual has at least two observations after the above-mentioned exclusion criteria are applied. For the given individuals we also collect information about the educational level of their parents and where the parents are born. After the data are cleaned, the sample includes 178,381 observations, 37,562 individuals and 2,162 firms. Descriptive statistics of key variables is presented in Table 1.

[Table 1 about here]

Because we focus on models with both individual- and firm-specific unobserved effects (which may be either random or fixed), identification is facilitated by a substantial proportion of the individuals being observed in at least two different firms over the period they occur in the sample. Table 2 provides some information about worker mobility for the workers in our data set.

[Table 2 about here]

[15] See http://www.ssb.no/a/publikasjoner/pdf/sos110/sos110.pdf.

We consider three main specifications for the firm effects in (21): No firm effects (NO), random firm effects (RE) and fixed firm effects (FE).[16] Henceforth, we use the notation RENO for the combination of random individual effects (RE) and no firm effects (NO), and analogously for REFE and RERE.

The unobserved individual-specific effect, μ_i, is treated as a random variable that is (possibly) correlated with level of schooling, S_i. The level of schooling is determined by the ordered probit model (22)-(23). For the vector of explanatory variables, U_i, of the ordered probit model, we include father's and mother's education level and world region of origin as identifying instruments – in addition to the exogenous variables from the wage equation (see Table B1). This is in line with a long tradition of using family background variables as instruments (see Card, 1999). The identifying instruments may affect the choice of schooling, but are assumed not to influence the wage. In addition to functional form assumptions, these exclusion restrictions identify the parameters of the model.

A full set of estimation results for the ordered probit model is presented in Table B1. Without going into details, we see that most of the family background variables are statistically significant. As seen from Table B1, a test of the relevance of the eight proposed instruments yields an F-statistic of 440 (with 8 degrees of freedom in the nominator), so that we clearly do not have a problem with weak instruments. To calculate the F-statistic of the test, we utilize that an F-statistic with d degrees of freedom in the nominator is asymptotically equivalent to W$/d$, where W is the Wald statistic involved when testing d zero restrictions on the parameters of the ordered probit model. The estimates reported in Table B1 were used to estimate the control function $\kappa(S_i, U_i)$ occurring in the "augmented" wage equation (21) to

[16] The importance of accounting for firm effects when estimating wage equations using employer–employee data has been emphasized among others by Lallemand et al. (2005), Plasman et al. (2007), Heyman (2007) and Grütter and Lalive (2009).

control for the endogeneity of schooling.

[Table 3 about here]

Table 3 contains estimation results of the wage equation under different assumptions with respect to the treatment of unobserved individual and firm-specific heterogeneity.[17] In the specification corresponding to columns (1)–(2), no firm effects are included, the results reported in columns (3)–(4) correspond to a model with random firm effects, and the last two columns to a model with fixed firm effects. For issues related to software and computing time, see Appendix A.

There is a positive selection into education, as seen from the fact that the estimate of the coefficient π of the control function is significantly positive in all three firm effects specifications (NO, RE or FE).[18] The test of overidentification reported in Table 3, shows that we do not reject the overidentification restrictions, except in the RENO model (i.e., the model without unobserved firm effects). In line with this, the estimated coefficient of years of schooling is higher in columns (1), (3) and (5), where the control function is not included, compared to the corresponding specifications that include the control function, i.e., (2), (4) and (6).

The estimated returns to an additional year of education is 0.068 in the model with no firm effects, when we control for self-selection. The estimated returns to education clearly become smaller when firm effects are included: 0.063 and 0.062 in the RERE and REFE specification (see columns (4) and (6), respectively). As long as we correct for the correlation between the individual effect and education,

[17] All the estimation results are robust to initiating the estimation algorithm from different sets of starting values. Thus the parameter estimates reported in Table 3 seem to correspond to global maxima.

[18] We have also estimated the three models controlling for selection using a continuous education variable instead of the category-based one which the results reported in Table 3 are based on. These results – not shown here but available from the authors upon request – also show positive self-selection.

it makes no difference whether one uses the RERE or REFE model. However, if we consider the model without unobserved firm effects (RENO) on the one hand and the models with firm effects (RERE and REFE) on the other, we find that the estimated returns to education for the former is 0.5–0.6 percentage points higher. Thus the differences are quite substantial and also significant since the standard error of the parameter estimate is less than 0.002 in the models with unobserved firm effects. If we also exclude the *observed* firm variables the difference becomes wider (about one percentage point).[19]

The parameter estimates for the experience coefficients do not vary greatly between the models. The maximum returns to experience are found to be at 25–30 years of experience, and the returns are more or less flat thereafter. The estimate of the male dummy is about 0.25, showing that the estimated gender wage gap is significant. The estimates of the education-type parameters are significant in all the models and do not seem to be influenced by the inclusion of a control function. Comparing the estimates for the three different specifications RENO, RERE and REFE, the estimates are somewhat higher in the former compared to the two latter specifications. Thus, to include unobserved firm-effects is more important than the particular choice of a random vs a fixed effects specification in the firm effects.

Using a Hausman test, we have tested the RERE model against the REFE model (i.e., fixed firm effects), in which the null hypothesis is that the RERE model is correct. The p-value was practically equal to zero. Because Hausman tests routinely reject the random effect specification in large samples, this test may not be very informative. However, as emphasized above, as long as we control for selection into education the parameter estimates of the two models, RERE and REFE, are very

[19] For the RERE and REFE models, the inclusion of firm-effects is of minor importance for the other parameter estimates. These results are not reported, but available from the authors upon request.

similar. The high estimated values of $\sigma_{\mu\mu}$ compared with $\sigma_{\nu\nu}$ reported in Table 3 (about four times as high), show that the individual effects have a much more dispersed distribution than the firm effects.

We have also estimated the FEVD model using the *felsdvreg* routine for STATA (see Cornelissen, 2008) followed by a vector decomposition to identify the effects of the time-invariant explanatory variables and individual means of the time-varying variables. The estimated returns to an additional year of education then becomes 0.072.[20] This is substantially higher than our estimates of both the REFE and RERE model with the control function included. This higher estimate is in accordance with the general criticism of the FEVD estimator, which – in our application – fails to address the problem of correlation between years of schooling and the individual effects.

Our model does not include match effects. In our notation, such effects can be described by the error structure $\eta_{it} = \phi_{i,J(i,t)} + \widetilde{\eta}_{it}$, where the match effect $\phi_{i,J(i,t)}$ depends on the matched pair $(i, J(i,t))$. Note that if the match effects are uncorrelated with the explanatory variables, our IFGLS estimator is still consistent with regard to the slope coefficients. The presence of match effects is often associated with assortative sorting, implying that an individual will move to a new job to obtain a better match, represented by a higher $\phi_{i,J(i,t)}$. This hypothesis implies that the conditional expectation $E\left[\phi_{i,J(i,t)} - \phi_{i,J(i,t')}|J(i,t) \neq J(i,t'), t > t'\right]$ should be positive. That is, job changes are, on average, associated with increasing match effects. We tested this assumption using the residuals from our RERE and REFE models (see Table 3, columns (4) and (6), respectively). The residuals were used as the dependent variable in an auxiliary regression where each new job of a worker is

[20] The full set of results for the FEVD estimator is not reported, but available from the authors upon request.

assigned a separate dummy (an indicator of the order of the job). Then we tested the null hypothesis that the coefficients of these dummies were jointly equal to zero. The p-value of the test was 0.24. This clear non-rejection, which contradicts other findings in the literature (see especially Woodcock 2008, 2015), is likely to be due to the fact that wages in Norwegian manufacturing to a large extent are determined in negotiations between employer and labour unions (the labour union coverage is close to 80 percent in NACE 29). Thus, there might be little to gain in terms of wage increase associated with a job change. Sørensen and Vejlin (2013), using Danish data, also found that the importance of the match effect was less than what was found by Woodcock (2008) on US data. Denmark resembles Norway with respect to union density and coverage.

4 Concluding remarks

More and more panel datasets are constructed by merging information from several registers. Merged employer-employee datasets give researchers the ability to control for a wide variety of observable characteristics as well as unobserved heterogeneity related to the two types of observation units: The main unit (in our application, an individual), and the secondary unit with whom the main unit is matched (in our case a firm). In this paper, we consider a general regression model with unobserved random effects corresponding to the main observational unit, and unobserved random or fixed effects corresponding to the unit with whom the main unit is matched.

To assume that the effects corresponding to the main-unit are random (in our case an individual), makes it possible to identify the effect of time-invariant individual-specific variables directly. This contrasts the approach in more traditional models for analyzing linked data models where the unobserved effects for the main units and the secondary units both are assumed to be fixed. In such approaches it is common to rely on the fixed effects vector decomposition (FEVD) estimator where one, after having estimated individual specific fixed effects in a first stage, run an auxiliary regression to estimate the effects of time-invariant individual-specific variables. However, this approach does not solve any endogeneity problem – contrary to a common belief – so one might instead use a random effects estimator, which is generally more efficient. In the case of endogenous regressors, we propose a control function approach based on instrumental variables, where the estimated control function is included as a regressor in the original regression equation to control for the endogeneity of explanatory variables.

A computation advantage of our approach is that it is mitigating the curse of dimensionality in high-dimensional two-way random effects models. This is done by using an IFGLS estimation procedure on variables subjected to the Helmert transformation. Compared to for instance the mixed model approach implemented in STATA this is a huge advantage in terms on computing time and memory requirements when it comes to handling large matrices.

Another advantage of our approach is that it utilizes more of the total variation in the data than fixed effects approaches. For instance, in the matched employer-employee data models identification of the fixed effects is driven only by the individuals moving from one firm to another over time. Thus with short panels, where typically only a small share of the individuals is observed in more than one firm, identification might be hard. In our approach, all the individuals contribute to the identification of the unobserved effects. Thus, there are likely to be substantial efficiency gains from our approach compared to models where the unobserved effects for both the main units and the secondary units are assumed to be fixed.

In our empirical application, we find that if the endogeneity of the time-invariant education variable is ignored – as done in matched two-way fixed effects employer employee models – the returns to education is biased upwards. Controlling for unobserved firm heterogeneity is only partly able to reduce the bias.

There are a set of issues we have not addressed and that need to be explored in future work. It would be useful to apply our approach also to applications outside the labour market area – as used for illustration in this paper. Furthermore, it would be useful to extend our model also to include match effects, to control for the value of match quality. A related issue, at least in employer-employee models, is sorting of workers with different levels of skill into particular firms, and therefore endogenous mobility. Still, the ideas and empirical evidence provided in this paper show the importance and potential fruitfulness of departing from traditional models where the unobserved heterogeneity of both the main units and the secondary units are assumed to be fixed.

References

[1] Abowd JM, Kramarz F (1999) Econometric analyses of linked employer–employee data. Lab Econ 6: 53–74

[2] Abowd JM, Kramarz F, Margolis DN (1999) High wage workers and high wage firms. Econometrica 67: 251–333

[3] Abowd JM, Kramarz F, Woodcock SD (2008) Econometric analyses of linked employer-employee data. In: Mátyás L, Sevestre P (eds) The econometrics of panel data: fundamentals and recent developments in theory and practice. Springer, Berlin, 727–760

[4] Abowd JM, Kramarz F, Perez-Duarte S, Schmutte IM (2014) Testable Models of Assortative Matching in the Labor Market. Mimeo, CREST(ENSAE)

[5] Anderson BDO, Moore JB (1979) Optimal Filtering. Prentice-Hall, Englewood Cliffs, NJ

[6] Arellano M, Bover O (1995) Another look at the instrumental variables estimation of error component models. J Econometrics 68: 29–51

[7] Bagger J, Lentz R (2014) An Empirical Model of Wage Dispersion with Sorting. NBER working paper 20031, Cambridge, MA.

[8] Balestra P, Krishnakumar J (2008) Fixed Effects Models and Fixed Coefficient Models. In: Mátyás L, Sevestre P (eds) The econometrics of panel data: fundamentals and recent developments in theory and practice. Springer, Berlin, pp 23–48

[9] Biørn E, Godager G (2010) Does quality influence choice of general practioners? An analysis of matched doctor-patient panel data. Econ Modelling 27: 842–853

[10] Blanchard P, Mátyás L (1996) Robustness of tests for error components models to non-normality. Econ Letters 51: 161–167

[11] Breusch TS (1987) Maximum likelihood estimation of random effects models. J Econometrics 36: 383–389

[12] Breusch TS, Ward MB, Nguyen HTM, Kompas T (2011) On the fixed-effects vector decomposition. Pol Anal 19: 123–134

[13] Card D (1999) The causal effect of education on earnings. In: Ashenfelter O, Card D (eds) Handbook of Labor Economics, Vol. III, part A. North-Holland, Elsevier, Amsterdam, pp 1801–1863

[14] Cornelissen T (2008) The Stata command felsdvreg to fit a linear model with two high-dimensional fixed effects. Stata J 8: 170–189

[15] Demidenko E (2004) Mixed Models: Theory and Application. Wiley, Hoboken New Jersey

[16] Dempster AP, Laird NM, Rubin, DB (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Royal Statistical Society B 39:1–38

[17] Dostie B (2011) Wages, Productivity and Aging. De Economist 159: 139–158

[18] Fahrmeir L, Tutz G (1994) Multivariate statistical modelling based on generalized linear models. Springer, New York

[19] Fitzmaurice GM, Laird NM, Ware JH (2004): Applied Longitudinal Analysis. Wiley, Hoboken New Jersey

[20] Francke MK, Koopman SJ, De Vos AF (2010). Likelihood functions for state space models with diffuse initial conditions. J Time Ser Anal 31: 407–414

[21] Garen J (1984) The returns to schooling: a selectivity bias approach with a continuous choice variable. Econometrica 52: 1199–1218

[22] Gilbert S (2002). Testing the distribution of error components in panel data models. Econ Letters 77*,* 47–53

[23] Gourieroux C, Monfort A (1995) Statistics and econometric models. Volume 1. Cambridge University Press, Cambridge

[24] Greene WH (2011) Fixed effects vector decomposition: a magical solution to the problem of time-invariant variables in fixed effects models. Pol Anal 19: 135–146

[25] Greene WH (2012) Econometric analysis. Seventh edition. Prentice Hall, London

[26] Grütter M, Lalive R (2009) The importance of firms in wage determination. Lab Econ 16: 149–160

[27] Guimarães P, Portugal P (2010) A simple feasible procedure to fit models with high-dimensional fixed effects. Stata J 10: 628–649

[28] Hausman JA, Taylor WE (1981) Panel data and unobservable individual effects. Econometrica 49: 1377–1398

[29] Heckman JJ (1979) Sample selection bias as a specification error. Econometrica 47: 153–162

[30] Heyman F (2007) Firm size or firm age? The effects on wages using matched employer–employee data. Labour 21: 237–263

[31] Hsiao C (2003) Analysis of panel data. Second edition. Cambridge University Press, Cambridge

[32] Ioannidou V, Ongena S (2010) 'Time for a change': loan conditions and bank behavior when firms switch banks. J Finance 65: 1847–1877

[33] Keane MP, Runkle DE (1992) On the estimation of panel-data models when instruments are not strictly exogenous. J Bus Econ Statist 10: 1–9

[34] Lallemand T, Plasman R, Rycx F (2005) Why do large firms pay higher wages? Evidence from matched worker–firm data. Int J Manpower 26: 705–723

[35] Le Maire D, Scheuer C (2013) Job Sampling and Sorting. Mimeo, University of Copenhagen

[36] Lopes de Melo R (2009) Sorting in the Labor Market: Theory and Measurement. Mimeo, University of Chicago

[37] Lütkepohl H (1996) Handbook of Matrices. Wiley, Chichester

[38] Oberhofer W, Kmenta J (1974) A general procedure for obtaining maximum likelihood estimates in generalized regression models. Econometrica 42: 579–590

[39] Plasman R, Rycx F, Tojerow I (2007) Wage differentials in Belgium: the role of worker and employer characteristics. Cah Econ Bruxelles 50: 11–40

[40] Plümper T, Troeger VE (2011) Fixed-effects vector decomposition: properties, reliability, and instruments. Pol Anal 19: 147–164

[41] Postel-Vinay F, Robin J-M (2002) Equilibrium Wage Dispersion with Worker and Employer Heterogeneity. Econometrica 70: 2295–2350

[42] Rockoff JE (2004) The impact of individual teachers on student achievement: Evidence from panel data. Amer Econ Rev 94: 247–252

[43] Shimer R (2005) The Assignment of Workers to Jobs in an Economy with Coordination Frictions. J Polit Econ 113: 996–1025

[44] Shimer R, Smith L (2000) Assortative Matching and Search. Econometrica 68: 343–369

[45] Sørensen T, Vejlin R (2013) The Importance of Worker, Firm and Match Effects in the Formation of Wages. Empirical Econ 45: 435–464

[46] Watson GA (2006). Computing Helmert transformations. J Comput Appl Math 197: 387–394

[47] Woodcock SD (2008) Wage differentials in the presence of unobserved worker, firm and match heterogeneity. Lab Econ 15: 772–794

[48] Woodcock SD (2015) Match Effects. Res Econ 69: 100–121

Table 1: Descriptive statistics of key variables

Variable	Mean	Standard deviation
Log-earnings (in 1000 NOK)	12.416	0.421
Years of schooling	12.102	2.283
Experience	14.882	10.097
Male	0.883	0.321
Education type:		
general programs	0.148	0.355
business and administration	0.088	0.283
sci & tech	0.702	0.457
World region of origin:		
Nordic countries except Norway	0.014	0.118
Western Europe except Turkey	0.007	0.084
East-Europe	0.002	0.041
North America	0.004	0.062
Rest of the world	0.003	0.055
Length of father's education	10.901	2.628
Length of mother's education	10.326	2.133
Firm variables		
number of employees[1)]	38	229
return on total assets[2)]	0.074	0.178

[1)] We do not apply the exclusion criteria involved when constructing the sample of individuals when deriving the firm-sample. Neither, we exclude individuals with missing relevant variables. Thus, when calculating the summary statistics for the "number of employees" more individuals are recorded compared to the sample of individuals.

[2)] Results before extra ordinary items and taxes plus interest payments divided by total assets

Table 2. Overview of number of firms in workers' employment history

Number of firms	Number of individuals having worked in the indicated number of firms
1	28,649
2	6,376
3	1,806
4	593
5	127
6	11
Total	37,562

Table 3: Empirical results for wage equations. Dependent variable: log-earnings

Specification:	RENO		RERE		REFE	
Control function included:	No	Yes	No	Yes	No	Yes
	(1)	(2)	(3)	(4)	(5)	(6)
Control function, $\widehat{\kappa}(S_i, U_i)$	–	0.018	–	0.015	–	0.014
	–	(0.003)	–	(0.003)	–	(0.003)
Worker characteristics						
years of schooling	0.075	0.068	0.069	0.063	0.067	0.062
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
experience	0.055	0.054	0.055	0.054	0.054	0.054
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
experience^2/100	-0.172	-0.171	-0.172	-0.170	-0.170	-0.169
	(0.005)	(0.005)	(0.010)	(0.010)	(0.010)	(0.010)
experience^3/1000	0.017	0.017	0.017	0.016	0.016	0.016
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
male	0.248	0.247	0.256	0.255	0.259	0.259
	(0.005)	(0.005)	(0.008)	(0.008)	(0.008)	(0.008)
Education type:						
general programs	0.096	0.096	0.085	0.084	0.077	0.077
	(0.008)	(0.008)	(0.031)	(0.011)	(0.011)	(0.031)
business and administration	0.069	0.068	0.060	0.059	0.053	0.052
	(0.008)	(0.008)	(0.034)	(0.033)	(0.012)	(0.034)
sci & tech	0.049	0.048	0.038	0.037	0.033	0.032
	(0.007)	(0.007)	(0.023)	(0.010)	(0.010)	(0.022)
Firm variables						
log-number of employees	0.068	0.068	0.039	0.039	0.028	0.028
	(0.006)	(0.006)	(0.001)	(0.001)	(0.001)	(0.001)
returns to total assets	0.033	0.033	0.065	0.065	0.066	0.066
	(0.001)	(0.001)	(0.004)	(0.004)	(0.004)	(0.004)
Variance components:						
$\sigma_{\eta\eta}$ (idiosyncratic noise)	0.035	0.035	0.027	0.027	0.027	0.027
$\sigma_{\mu\mu}$ (individual effect)	0.072	0.072	0.040	0.040	0.039	0.039
$\sigma_{\nu\nu}$ (firm effect)	–	–	0.009	0.009	–	–
Sargan test of overidentification (7 d.f.)						
J-statistics		19.67		10.61		8.21
p-value		0.006		0.15		0.31
Sample: No. of individuals 37,562, no. of firms 2,162, no. of observations 178,381						

Notes: Standard errors in parentheses. Year and labor market area effects are accounted for in all the models. NO, FE and RE denote, respectively, the model with no firm effects, the model with fixed firm effects and the model with random firm effects.

Appendix A. Supplementary materials

Obtaining derivatives of $L(\theta, B)$ Direct differentiation of $L(\theta, B)$ (see (10)) w.r.t. θ is intractable, because the number of computations involved is of order $O\left(\left(\overline{T}N\right)^2\right)$. To see this, it follows from Lütkepohl (1996, p. 198) that

$$\frac{\partial((Y-XB)'\Sigma(\theta)^{-1}(Y-XB))}{\partial\theta} = -\frac{\partial vec(\Sigma(\theta))'}{\partial\theta}\Sigma(\theta)^{-1}(Y-XB)\otimes\Sigma(\theta)^{-1}(Y-XB)$$

($\otimes$ denotes the Kronecker product), where $\Sigma(\theta)^{-1}(Y-XB)\otimes\Sigma(\theta)^{-1}(Y-XB)$ is a $\left(\overline{T}N\right)^2\times 1$ vector. In Proposition 2, we obtain analytical derivatives of $L(\theta, B)$ in an indirect way by performing operations that typically will be only of order $O(M^2)$, which is quite feasible even for large M.

Proposition 2

$$\left.\frac{\partial L(\theta,B)}{\partial\theta}\right|_{\theta=\theta_0} = \left.\frac{\partial M(\theta|\theta_0;B)}{\partial\theta}\right|_{\theta=\theta_0} \tag{25}$$

where

$$\begin{aligned} M(\theta|\theta_0;B) \;=\; & -\frac{1}{2}\sum_{i=1}^{N}T_i\ln\sigma_{\eta\eta}-\frac{1}{2}\sum_{i=1}^{N}\ln(\frac{1}{T_i}+\lambda_\mu) \\ & -\frac{1}{2}\sigma_{\eta\eta}^{-1}\sum_{i=1}^{N}(\frac{1}{T_i}+\lambda_\mu)^{-1}\left\{(\overline{y}_i-\overline{x_i}\beta-z_i\gamma-\overline{G}_i\widehat{\nu}(\theta_0;B))^2+\overline{G}_iV(\theta_0)\overline{G}_i'\right\} \\ & -\frac{1}{2}\sigma_{\eta\eta}^{-1}\left((\overrightarrow{y}-\overrightarrow{X}\beta-\overrightarrow{G}\widehat{\nu}(\theta_0))'(\overrightarrow{y}-\overrightarrow{X}\beta-\overrightarrow{G}\widehat{\nu}(\theta_0))+tr(\overrightarrow{G}V(\theta_0)\,\overrightarrow{G}')\right) \\ & -\frac{M}{2}\ln\sigma_{\nu\nu}-\frac{1}{2}\sigma_{\nu\nu}^{-1}\left((\widehat{\nu}(\theta_0)'\widehat{\nu}(\theta_0)+tr(V(\theta_0))\right)). \end{aligned} \tag{26}$$

with

$$\widehat{\nu}(\theta_0;B)\equiv\mathsf{E}\{\nu\,|\,Y;(\theta_0,B)\}=P(\theta_0)(Y-XB) \tag{27}$$

and

$$\mathsf{Var}\{\nu\,|\,Y;\theta_0\}=V(\theta_0), \tag{28}$$

where $V(\ \theta_0)$ and $P(\ \theta_0)$ are calculated from (13) and (12), respectively.

Proof:

We first show that

$$\begin{aligned} M(\theta|\theta_0;B) &= \int \ln f(Y,\nu;(\theta,B))\, f(\nu|\,Y;(\theta_0,B))\, d\nu \qquad (29) \\ &\equiv \mathsf{E}\left\{\ln f(Y,\nu;(\theta,B))\,|Y;(\theta_0,B)\right\}, \end{aligned}$$

where $f(\cdot;\psi)$ and $f(\cdot|\cdot;\psi)$ is generic notation for joint and conditional probability densities, respectively, that belong to a parametric family, with parameter value ψ. The expectation in (29) is with respect to the latent variables (firm effects) ν conditional on the data Y, given B and with θ evaluated at θ_0. The function defined on the right-hand side of (29) is well-known from the EM algorithm and is usually referred to as the "complete data" log-likelihood. Given (29), the result (25) is well-known from the literature. See for example Dempster et al. (1977), with discussions, and Fahrmeir and Tutz (1994).

By definition

$$\begin{aligned} &\mathsf{E}\left\{\ln f(Y,\nu;(\theta,B))\,|Y;(\theta_0,B)\right\} \\ &= \mathsf{E}\left\{(\ln f(Y|\nu;(\theta,B)) + \ln f(\nu;(\theta,B)))\ |Y;(\theta_0,B)\right\} \qquad (30) \\ &= -\frac{1}{2}\sum_{i=1}^{N} T_i \ln\sigma_{\eta\eta} - \frac{1}{2}\sum_{i=1}^{N}\ln(\frac{1}{T_i}+\lambda_\mu) \qquad (31) \\ &-\frac{1}{2}\sigma_{\eta\eta}^{-1}\mathsf{E}\left\{\sum_{i=1}^{N}(\frac{1}{T_i}+\lambda_\mu)^{-1}(\overline{y}_i-\overline{x_i}\beta-z_i\gamma-\overline{G}_i\nu)^2\,|Y;\theta_0\right\} \\ &-\frac{1}{2}\sigma_{\eta\eta}^{-1}\mathsf{E}\left\{\left(\overrightarrow{y}-\overrightarrow{X}\beta-\overrightarrow{G}\nu\right)'\left(\overrightarrow{y}-\overrightarrow{X}\beta-\overrightarrow{G}\nu\right)\,|Y;\theta_0\right\} \\ &-\frac{M}{2}\ln\sigma_{\nu\nu}-\frac{1}{2}\sigma_{\nu\nu}^{-1}\mathsf{E}\left\{\nu'\nu|\,Y;\theta_0\right\}. \qquad (32) \end{aligned}$$

To evaluate the expectations in (32), we only need to calculate the conditional expectations

$$\widehat{\nu}(\theta_0) \equiv \mathsf{E}\left\{\nu\,|\,Y;\theta_0\right\})$$

and the conditional covariance matrix $\mathsf{Var}\,\{\nu\,|\,Y;\theta_0\}$. By applying the general formulae in Francke et al. (2010), we verify that

$$\mathsf{Var}\,\{\nu\,|\,Y;\theta_0\} = V(\theta_0),$$

and

$$\mathsf{E}\,\{\nu\,|\,Y;\theta_0\} = P(\theta_0)(Y - X\widehat{B}),$$

where $V(\ \theta_0)$ and $P(\ \theta_0)$ are already defined (and calculated) in (13) and (12), respectively. Furthermore,

$$\mathsf{E}\,\{\nu'\nu|\,Y;\theta_0\} = \widehat{\nu}(\theta_0)'\widehat{\nu}(\theta_0) + tr(V(\theta_0)).$$

Thus we have established that both (29) and (26) hold and hence Proposition 2.

In contrast to $L(\theta, B)$, $M(\theta|\ \theta_0; B)$ is trivial to differentiate with respect to θ (for given θ_0), since no matrix-inversions are required in (26) (the number of operations needed to calculate $V(\ \theta_0)$ and $\widehat{\nu}(\ \theta_0, B)$ are of order $O(M^2)$, as discussed above).

Computational issues The whole estimation procedure is programmed in GAUSS. For the sample used for illustration, we get convergence after approximately 8 minutes for the RERE and REFE models on a 64 cores Linux server with a maximum clock rate of 2.5 GHz (HP BL685c G7). For comparison, estimating the RERE model (on the same server) using the STATA command mixed convergence was not obtained within 24 hours even for a 10% subsample (it is acknowledged in the documentation of the STATA command *mixed*, that the approach is feasible only when the dimensionality is small to moderate).[21]

[21] The syntax is

$$mixed\ \ depvar\ [indepvars]...||\_all : R.individual||\_all : R.firm$$

(see the user manual at http://www.stata.com/manuals13/me.pdf)

We also estimate a real wage equation with random individual effects and fixed firm effects. Convergence is achieved after approximately 8 minutes. For comparison, the STATA command xtreg (with random individual effects and fixed firm effects entered as dummy variables) takes 2 hours and 13 minutes to converge on the same server using the full sample. We take this as evidence that our approach provides a substantial improvement relative to popular approaches for analyzing models with two-way unobserved heterogeneity.

Appendix B: Estimation results of ordered probit model

Table B1: Ordered probit parameter estimates. Dependent variable: level of schooling

Variable/Parameter	Estimate	Standard error
Male	.024	.016
Length of father's education	.088	.009
Length of mother's education	.063	.010
Length of father's education x		
Length of mother's education	.001	.001
World region of origin:		
Nordic countries except Norway	.015	.046
Western Europe except Turkey	–.092	.072
East-Europe	–.189	.141
North America	.222	.093
Rest of the world	.094	.112
Threshold parameters:		
ζ_1	.821	.109
ζ_2	1.285	.109
ζ_3	2.319	.110
ζ_4	2.556	.109
ζ_5	3.375	.109
ζ_6	4.790	.114
ζ_7	5.982	.284
Test of weak instruments:		
F-statistic (p-value)[1)]	440.7 (0.000)	
Number of observations (individuals)	37,562	

Notes: Robust standard errors. Region of residence is accounted for.

[1)] F=W(8)/8, where W(8) is the Wald test-statistic with 8 d.f.