Econometric Analyses of Linked Employer-Employee Data

John M. Abowd, Francis Kramarz, and Simon Woodcock*

May 2006

1 Introduction

There has been a recent explosion in the use of linked employer-employee data to study the labor market. This was documented, in part, in our *Handbook of Labor Economics* chapter (Abowd and Kramarz, 1999a).[1] Various new econometric methods have been developed to address the problems raised by integrating longitudinal employer and employee data. We first described these methods in Abowd and Kramarz (1999b). In this chapter, we present a survey of these new econometric methods, with a particular emphasis on new developments since our earlier articles.

Linked employer-employee data bring together information from both sides of the labor market. They therefore permit, for the first time, equilibrium analyses of labor market outcomes. They also allow researchers to investigate the joint role of worker and firm heterogeneity, both observed and unobserved, on labor market outcomes. Labor economists have taken full advantage of these data to revisit classic questions and to formulate new ones, and much has been learned as a result. For example, Abowd, Kramarz, Lengermann, and Roux (2005) have revisited the classic question of inter-industry wage differentials to determine whether they are attributable to workers or firms. Abowd, Kramarz, Lengermann, and Perez-Duarte (2003) use linked employer-employee data to examine whether "good" workers are employed by "good" firms. Dostie (2005) presents new evidence on the returns to seniority and its relation to turnover; and Woodcock (2003) examines the role of heterogeneity and worker-firm learning on employment and wage dynamics. These applied endeavors have demonstrated the value of linked employer-employee data. They have also spurred

*Abowd is Professor of Labor Economics, Cornell University, Distinguished Senior Research Fellow at the US Census Bureau and affiliated with the NBER and CREST. Kramarz is head of the Department of Research, Centre de Recherche en Economie et Statistique (CREST) at INSEE, Professor at Ecole Polytechnique and is a fellow at CEPR and IZA. Woodcock is Assistant Professor at Simon Fraser University. Abowd acknowledges financial support from the NSF (SBER 96-18111).

[1] See also Lane, Burgess and Theeuwes (1997) for a review of uses of longitudinal linked employer-employee data.

the development of new econometric methods to analyze these data. These new methods, rather than specific applications, are the primary focus of this chapter.

A distinguishing feature of longitudinal linked employer-employee data is that individuals and their employers are identified and followed over time. Furthermore, the relation between employer and employee, called a job relation, is continuously monitored. From a statistical perspective, there are three populations under simultaneous study. Individuals are sampled from the population of households, workplaces are sampled from the population of businesses, and jobs are sampled from the population of employment histories. Because of the multiple sampling frames involved, it is necessary to be precise about the statistical structure of the variables under study, since they may come from the individual, employer, or job frame. Measured characteristics of the individual, employer, and job are collected at multiple points in time, which may or may not be synchronous. To make clear the importance of careful elaboration of the sample structure for the variables under study, we will consider a prototypical integrated employer-employee database before turning to specific statistical models. The specific statistical models that we consider are generalizations of the specifications we first used in Abowd, Kramarz and Margolis (1999, AKM hereafter) as well as in more recent research.

We have noted a general misunderstanding of some recent, and some not-so-recent, empirical methods used by statisticians. We therefore make an effort to relate these methods to those used by panel data econometricians. We show the relation between various fixed-effects estimators and estimators popular in the variance components literature – in particular, mixed-effects estimators (see Searle, Casella and McCulloch, 1992). As we will see, statisticians and econometricians have different parameters of interest, the former relying more on the variance components and the design of the data, the latter being more concerned with endogeneity in its various guises. These generate a variety of distinct computational issues. Consequently econometricians and statisticians have independently developed a variety of tools to estimate the effects of interest. However, the realized effects have the same interpretation under all methods that we consider.

We begin, in section 2, by describing a prototypical longitudinal linked data set and discussing the related problems of missing data and sampling from integrated data. In section 3, we present two specifications for linear statistical models that relate linked employer and employee data to outcomes measured at the individual level. In the first and more general specification, person effects and firm effects can reflect interaction between observable person or firm characteristics and unobserved person and firm effects. For instance, match effects are potentially representable in this setting. In the second and simpler specification, a typical individual has a zero mean for the measured outcomes. Person effects measure deviations over time from this zero mean that do not vary as the employee moves from firm to firm. Firm effects measure deviations from this zero mean that do not vary as the firm employs different individuals. We continue, in section 4, by defining a variety of effects that are functions of the basic person and firm effects. Section 5 considers the estimation of the person and firm effects by fixed-effects methods. Section 6 discusses the use of mixed-effects estimators, the question of orthogonal design, and their relation with various correlated random-effects specifications. In section 7 we discuss the important heterogeneity biases that arise when either the person or firm effects are missing or incompletely specified. We discuss the consequences of endogenous mobility in section 8, and conclude in section 9.

2 A Prototypical Longitudinal Linked Data Set[2]

To summarize the complete likelihood function for linked longitudinal employer-employee data, we adopt the formalization in Abowd and Woodcock (2001). They considered statistical models for imputing missing data in linked databases using the full-information techniques developed by Rubin (1987). Their prototypical longitudinal linked data set contains observations about individuals and their employers linked by means of a work history that contains information about the jobs each individual held with each employer. The data are longitudinal because complete work history records exist for each individual during the sample period and because longitudinal data exist for the employer over the same period.

Suppose we have linked data on N workers and J firms with the following file structure. There are three data files. The first file contains data on workers, U, with elements denoted u_i, $i = 1, ..., N$. In the discussion below these data are time-invariant but in other applications they need not be. Call U the individual characteristics file. The second data file contains longitudinal data on firms, Z, with elements z_{jt}, $j = 1, ..., J$ and $t = 1, ..., T_j$. Call Z the employer characteristics file. The third data file contains work histories, W, with elements w_{it}, $i = 1, ..., N$ and $t = 1, ..., T_i$. Call W the work history file. It contains data elements for each employer who employed individual i during period t. The data U and W are linked by a person identifier. The data Z and W are linked by a firm identifier; we conceptualize this by the link function $j = J(i, t)$ which indicates the firm j at which worker i was employed at date t. For clarity of exposition, we assume throughout that all work histories in W can can be linked to individuals in U and firms in Z and that the employer link $J(i, t)$ is unique for each (i, t).[3]

2.1 Missing Data

Abowd and Woodcock consider the problem of imputing missing data in a longitudinal linked database. Their approach is based on the Sequential Regression Multivariate Imputation (SRMI; see Ragunathan et al. 1998). When imputing missing data in each of the three files, they condition the imputation on as much

[2] This section is based on Abowd and Woodcock (2001).

[3] The notation to indicate a one-to-one relation between work histories and indviduals when there are multiple employers is cumbersome. See Abowd and Stinson (2003) for a complete development of the likelihood function allowing for multiple employers during the period.

available information as possible. For example, when imputing missing data in the individual characteristics file U they condition not only on the non-missing data in U (observed characteristics of the individual) but also on characteristics of the jobs held by the individual (data in W) and the firms at which the individual was employed (data in Z). Similarly, when conditioning the imputation of missing data in W and Z, they condition on non-missing data from all three files. In this manner, their imputation is based on the complete likelihood function for the linked longitudinal data.

The Abowd and Woodcock technique necessitates some data reduction. To understand the data reduction, consider imputing missing data in the individual characteristics file U. Since individuals have work histories with different dynamic configurations of employers, explicitly conditioning the missing data imputation of individual characteristics on every variable corresponding to each job held by each worker is impractical – there are a different number of such variables for each observation to be imputed. A sensible alternative is to condition on some function of the available data that is well defined for each observation. For example, to impute missing data in U, one could condition on the person-specific means of time-varying work history and firm variables. Similar data reductions are required to impute missing data in the other files. In what follows, we use the functions g, h, m and n to represent data reductions that span sampling frames.

Abowd and Woodcock note the importance of conditioning the imputation of time-varying variables on contemporaneous data and leads and lags of available data. Because the dynamic configuration of work histories varies from worker to worker and the pattern of firm "births" and "deaths" varies from firm to firm, not every observation with missing data has the same number of leads and lags available to condition the imputation. In some cases, there are no leads and lags available at all. They suggest grouping observations by the availability of dynamic conditioning data (*i.e.*, the number of leads and lags available to condition missing data imputations) and separately imputing missing data for each group. This maximizes the set of conditioning variables used to impute each missing value. Again, some data reduction is generally necessary to keep the number of groups reasonable. For example, one might only condition on a maximum of s leads and lags, with $s = 1$ or $s = 2$. They parameterize the set of dynamic conditioning data available for a particular observation by κ_{it} in the work history file, and γ_{jt} in the firm file. It may also be desirable to split the observations into separate groups on the basis of observable characteristics, for example sex, full-time/part-time employment status, or industry. They parameterize these groups by λ_i in the individual file, μ_{it} in the work history file, and ν_{jt} in the firm file.

The key aspects of the SRMI algorithm are as follows. One proceeds sequentially and iteratively through variables with missing data from all three files, at each stage imputing missing data conditional on all non-missing data and the most recently imputed values of missing data. The optimal imputation sequence is in increasing degree of missingness. As each variable in the sequence comes up for imputation, observations are split into groups based on the value of κ_{it},

γ_{jt}, λ_i, μ_{it}, and/or ν_{jt}. The imputed values are sampled from the posterior predictive distribution of a parametric Bayesian imputation model that is specific to each group. After the imputes are drawn, the source file for the variable under imputation is reassembled from each of the group files. Before proceeding to the next variable, all three files must be updated with the most recent imputations, since the next variable to be imputed may reside in another file (U, W, or Z). At the same time, the functions of conditioning data (including leads and lags) described above generally need to be re-computed. The procedure continues for a pre-specified number of rounds or until the imputed values are stable.

Explicitly specifying the posterior predictive densities from which the imputations are drawn is notationally cumbersome. For completeness, we reproduce these directly from Abowd and Woodcock in (1), (2), and (3). For a particular variable under imputation, subscripted by k, they denote by $U_{<k}$ the set of variables in U with less missing data than variable k; $W_{<k}$ and $Z_{<k}$ are defined analogously. They denote by $U_{>k}$ the set of variables in U with more missing data than variable k, and define $W_{>k}$ and $Z_{>k}$ similarly. They use the subscript *obs* to denote variables with no missing data. They also subscript conditioning variables by i, j, and t as appropriate to make clear the relationships between variables in the three data files. The predictive densities from which the round $\ell+1$ imputations are drawn are

$$\int f_{u_k}\left(u_k \left| \begin{array}{c} U_{<k,i}^{(\ell+1)}, U_{>k,i}^{(\ell)}, U_{obs,i}, \\ g_k\left(\left\{Z_{<k,J(i,t)}^{(\ell+1)}, Z_{>k,J(i,t)}^{(\ell)}, Z_{obs,J(i,t)}\right\}_{t=1}^{t=T_i}\right), \\ h_k\left(\left\{W_{<k,it}^{(\ell+1)}, W_{>k,it}^{(\ell)}, W_{obs,it}\right\}_{t=1}^{t=T_i}\right), \lambda_i, \theta_k \end{array}\right.\right) p_k\left(\theta_k|.\right) d\theta_k \quad (1)$$

$$\int f_{w_k}\left(w_k \left| \begin{array}{c} U_{<k,i}^{(\ell+1)}, U_{>k,i}^{(\ell)}, U_{obs,i}, \\ \left\{Z_{<k,J(i,\tau)}^{(\ell+1)}, Z_{>k,J(i,\tau)}^{(\ell)}, Z_{obs,J(i,\tau)}\right\}_{\tau=t-s}^{\tau=t+s}, \\ \left\{w_{k,i\tau}^{(\ell)}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \\ \left\{W_{<k,i\tau}^{(\ell+1)}, W_{>k,i\tau}^{(\ell)}, W_{obs,i\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \kappa_{it}, \mu_{it}, \theta_k \end{array}\right.\right) p_k\left(\theta_k|.\right) d\theta_k \quad (2)$$

$$\int f_{z_k}\left(z_k \left| \begin{array}{c} m_k\left(U_{<k,J^{-1}(i,t)}^{(\ell+1)}, U_{>k,J^{-1}(i,t)}^{(\ell)}, U_{obs,J^{-1}(i,t)}\right), \\ \left\{z_{k,j\tau}^{(\ell)}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{Z_{<k,j\tau}^{(\ell+1)}, Z_{>k,j\tau}^{(\ell)}, Z_{obs,j\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \\ n_k\left(\left\{\begin{array}{c} W_{<k,J^{-1}(i,\tau)\tau}^{(\ell+1)}, W_{>k,J^{-1}(i,\tau)\tau}^{(\ell)}, \\ W_{obs,J^{-1}(i,\tau)\tau} \end{array}\right\}_{\tau=t-s}^{\tau=t+s}\right), \\ \gamma_{jt}, \nu_{jt}, \theta_k \end{array}\right.\right) p_k\left(\theta_k|.\right) d\theta_k, \quad (3)$$

where $f_{\cdot k}$ is the likelihood defined by an appropriate generalized linear model for variable k, θ_k are unknown parameters, and the posterior densities $p_k\left(\theta_k|.\right)$

are conditioned on the same information as $f_{\cdot k}$. Repeating the missing data imputation method M times yields M sets of completed data files (U^m, W^m, Z^m) which they call the completed data implicates $m = 1, ..., M$.

Equations (1-3) describe the complete set of conditional distributions of each variable in the linked longitudinal employer-employee data, given all other variables. Hence, they form the basis for sampling from this complete distribution. One can use these equations in a Gibbs sampler or other Monte Carlo Markov Chain algorithm to draw a complete sample of linked longitudinal data that has the same likelihood function as the original analysis sample. Abowd and Woodcock use this property to draw partially synthetic data from the joint posterior predictive distribution.

2.2 Sampling from Linked Data

Many of the estimators discussed below are computationally intensive. Because many longitudinal linked databases are constructed from administrative records they are very large.[4] Thus researchers are sometimes faced with the prospect of sampling from the linked data to facilitate estimation. In principle, sampling from any one of the frames (workers, firms, or jobs) that comprise the linked data is straightforward. However, the estimators discussed below rely on links between sampling frames (i.e., observed worker mobility between firms) for identification. Small simple random samples of individuals may not retain sufficient "connectedness" between sampling frames for identification.[5]

Woodcock (2003) considers the problem of sampling from linked data while preserving a minimum degree of connectedness between sampling frames. He presents a "dense" sampling algorithm that guarantees each sampled worker is connected to at least n others by a common employer. The sample is otherwise representative of the population of individuals employed in a reference period. The dense sampling algorithm is straightforward. It operates on the population of jobs at firms with at least n employees in the reference period t. In the first stage, sample firms with probabilities proportional to their employment in period t. In the second stage, sample a minimum of n employees from each sampled firm, with probabilities inversely proportional to the firm's employment in period t. A simple application of Bayes' rule demonstrates that all jobs active in period t have an equal probability of being sampled. The sample is thus equivalent to a simple random sample of jobs active in period t, but guarantees that each sampled worker is connected to at least n others.

[4] See Abowd and Kramarz (1999a) for a typology.

[5] See section 5.1.2 below for a discussion of connectedness and its role in identifying person and firm effects.

3 Linear Statistical Models with Person and Firm Effects

3.1 A General Specification

We consider general linear statistical model:

$$y_{it} = x_{it}\beta + q_{it,\mathrm{J}(i,t)}\theta_i + r_{it,\mathrm{J}(i,t)}\psi_{\mathrm{J}(i,t)} + \varepsilon_{it} \tag{4}$$

where y_{it} is an observation for individual $i = 1, \ldots, N$, $t = n_{i1}, ..., n_{iT_i}$, T_i is the total number of periods of data available for individual i, and the indices $n_{i1}, ..., n_{iT_i}$ indicate the period corresponding to the first observation on individual i through the last observation on that individual, respectively. The vectors x_{it} contain P time-varying, exogenous characteristics of individual i; the vectors $q_{it,\mathrm{J}(i,t)}$, and $r_{it,\mathrm{J}(i,t)}$ contain respectively Q and R exogenous characteristics of individual i and (or) firm $\mathrm{J}(i,t)$. Both vectors include indicators that associate an observation and a person (for Q) or a firm (for R). We denote the design matrices of these indicators by D and F, respectively. The vector θ_i is a size Q vector of person effects; $\psi_{\mathrm{J}(i,t)}$ is a size R vector of firm effects; and ε_{it} is the statistical residual. The first period available for any individual is arbitrarily dated 1 and the maximum number of periods of data available for any individual is T. Assemble the data for each person i into conformable vectors and matrices

$$\begin{aligned} y_i &= \begin{bmatrix} y_{i,n_{i1}} \\ \cdots \\ y_{i,n_{iT_i}} \end{bmatrix}, \\ X_i &= \begin{bmatrix} x_{i,n_{i1},1} \cdots \ x_{i,n_{i1},P} \\ \cdots \\ x_{i,n_{iT_i},1} \cdots \ x_{i,n_{iT_i},P} \end{bmatrix}, \\ \varepsilon_i &= \begin{bmatrix} \varepsilon_{i,n_{i1}} \\ \cdots \\ \varepsilon_{i,n_{iT_i}} \end{bmatrix} \end{aligned}$$

where y_i and ε_i are $T_i \times 1$ and X_i is $T_i \times P$ with similar definitions for $Q_{i,\mathrm{J}(i,.)}$ and $R_{i,\mathrm{J}(i,.)}$.

We assume that a simple random sample of N individuals is observed for a maximum of T periods. Assume further that ε_i has the following properties:

$$\mathrm{E}\left[\varepsilon_i | X_i, Q_{i,\mathrm{J}(i,.)}, R_{i,\mathrm{J}(i,.)}\right] = 0$$

and

$$\mathrm{Cov}\left[\varepsilon_i, \varepsilon_m | X_i, Q_{i,\mathrm{J}(i,.)}, R_{i,\mathrm{J}(i,.)}, X_m, Q_{m,\mathrm{J}(m,.)}, R_{m,\mathrm{J}(m,.)}\right] = \begin{cases} \left\{\Sigma_{T_i}\right\}_i, \ i = m \\ 0, \text{ otherwise} \end{cases}$$

where $\left\{\Sigma_{T_i}\right\}_i$ means the selection of rows and columns from a $T \times T$ positive definite symmetric matrix Σ such that the resulting $T_i \times T_i$ positive definite symmetric matrix corresponds to the periods $\{n_{i1}, n_{i2}, \ldots, n_{iT_i}\}$.[6] In full matrix notation we have

$$y = X\beta + \left[D, \widetilde{Q}\right]\theta + \left[F, \widetilde{R}\right]\psi + \varepsilon \tag{5}$$

where: X is the $N^* \times P$ matrix of observable, time-varying characteristics (in deviations from the grand means); D is the $N^* \times N$ design matrix of indicator variables for the individual; $\widetilde{Q}$ is the $N^* \times (Q-1)N$ matrix of the observable characteristics in q with person-specific effects; F is the $N^* \times J$ design matrix of indicator variables for the firm; $\widetilde{R}$ is the $N^* \times (R-1)N$ matrix of observable characteristics in r with firm-specific effects; y is the $N^* \times 1$ vector of dependent data (also in deviations from the grand mean); ε is the conformable vector of residuals; and $N^* = \sum_{i=1}^{N} T_i$. The vector y is ordered according to individuals as

$$y = \begin{bmatrix} y_1 \\ \cdots \\ y_N \end{bmatrix} \tag{6}$$

and X, Q, R and ε are ordered conformably. A typical element of y is y_{it} and a typical element of X, or any similarly organized matrix, as $x_{(i,t)p}$ where the pair (i, t) denotes the row index and p denotes the column index. The effects in equations (4) and (5) are: β, the $P \times 1$ vector of coefficients on the time-varying personal characteristics; θ, the $N \times Q$ vector of individual effects; and ψ, the $J \times R$ vector of firm effects. When estimating the model by fixed effects methods, identification of the effects is accomplished by imposing a zero sample mean for θ_i and $\psi_{\mathrm{J}(i,t)}$ taken over all (i, t).[7] In the mixed effects case, identification is achieved by assuming the random effects have zero conditional mean and finite conditional variance.

3.2 The Pure Person and Firm Effects Specification

A simpler linear statistical model is specified as:

$$y_{it} = x_{it}\beta + \theta_i + \psi_{\mathrm{J}(i,t)} + \varepsilon_{it} \tag{7}$$

with variables defined as above except that θ_i is the *pure person effect* and $\psi_{\mathrm{J}(i,t)}$ is the *pure firm effect*. We now assume that ε_i has the following properties:[8]

$$\mathrm{E}\left[\varepsilon_i | D_i, F_i, X_i\right] = 0 \tag{8}$$

and

$$\mathrm{Cov}\left[\varepsilon_i, \varepsilon_m | D_i, D_m, F_i, F_m, X_i, X_m\right] = \begin{cases} \{\Sigma_{T_i}\}_i, & i = m \\ 0, & \text{otherwise} \end{cases}$$

[6] See section 6 for a specific example of $\{\Sigma_{T_i}\}_i$.

[7] Further details of identification requirements are discussed in section 5.1.

[8] The zero conditional mean assumption (8) has been interpreted as an assumption of "exogenous mobility," since it precludes any relationship between an individuals employment location (measured by F_i) and the errors ε_i. See AKM for further discussion, and section 8 below for recent work that accomodates endogenous mobility.

where D_i and F_i are those elements of D and F, respectively, corresponding to person i. In full matrix notation we have

$$y = X\beta + D\theta + F\psi + \varepsilon \tag{9}$$

where: X is the $N^* \times P$ matrix of observable, time-varying characteristics (in deviations from the grand means); D is the $N^* \times N$ design matrix of indicator variables for the individual; F is the $N^* \times J$ design matrix of indicator variables for the employer at which i works at date t (J firms total); y is the $N^* \times 1$ vector of dependent data (also in deviations from the grand mean); ε is the conformable vector of residuals; and $N^* = \sum_{i=1}^{N} T_i$.

The effects in equations (7) and (9) are: β, the $P \times 1$ vector of coefficients on the time-varying personal characteristics; θ, the $N \times 1$ vector of individual effects; and ψ, the $J \times 1$ vector of firm effects. As above, identification of the effects is accomplished by imposing a zero sample mean for θ_i and $\psi_{\mathrm{J}(i,t)}$ taken over all (i,t) for fixed-effects estimators, and by assuming of zero conditional mean and finite conditional variance for random-effects estimators.

4 Definition of Effects of Interest

Many familiar models are special cases of the linear model in equations (4) and (5) or the simpler version in equations (7) and (9). In this section we define a variety of effects of interest that are functions of the person and firm effects specified in the preceding section. These definitions allow us to consider these familiar models using common notation and internally coherent definitions. We use the example of estimating inter-industry wage differentials, frequently called industry effects, to illustrate some important issues.

4.1 Person Effects and Unobservable Personal Heterogeneity

The person effect in equation (7) combines the effects of observable time-invariant personal characteristics and unobserved personal heterogeneity. We decompose these two parts of the pure person effect as

$$\theta_i = \alpha_i + u_i\eta \tag{10}$$

where α_i is the unobservable personal heterogeneity, u_i is a vector of time-invariant personal characteristics, and η is a vector of effects associated with the time-invariant personal characteristics. An important feature of the decomposition in equation (10) is that estimation can proceed for the person effects, θ_i, whether random or fixed, without direct estimation of η. Since many linked employer-employee data sets contain limited, or missing, information on the time-invariant characteristics u_i, we describe the estimation algorithms in terms of θ_i; however, when data on u_i are available, equivalent techniques can be used for estimation in the presence of α_i (see AKM for the fixed effects case, Woodcock (2003) for the mixed effects case). The design matrix D in equation (9) can be augmented by columns associated with the observables u_i so that the statistical methods discussed below are applicable to the estimation of the effect specified in equation (10).

This specification can be further generalized by incorporating time-varying observable characteristics of the worker, q_{it}, or of the firm, q_{jt}, that may well be interacted as in (4) and (5) to give:

$$\theta_{jit} = \alpha_i + u_i \eta + q_{it}\mu_i + q_{jt}\delta_i \tag{11}$$

where μ_i and δ_i are vectors of effects associated with the time-varying person and firm observable characteristics. Statistical analysis of the effects defined by equation (11) is accomplished by augmenting the columns of D to reflect the data in q_{jt} and q_{it}. The formulae shown in the estimation sections below can then be applied to the augmented design matrix.

4.2 Firm Effects and Unobservable Firm Heterogeneity

The firm effect in equation (7) combines the effects of observable and unobserved time-invariant characteristics of the firm. It can also be generalized to contain the effects of time-varying characteristics of the firm and time-varying characteristics of the employee-employer match as in equations (4) and (5). We illustrate each of these possibilities in this subsection.

We can decompose the pure firm effect of equation (7) into observable and unobservable components as

$$\psi_j = \phi_j + v_j \rho \tag{12}$$

where ϕ_j is unobservable firm heterogeneity, v_j is a vector of time-invariant firm characteristics, and ρ is a vector of associated effects.

Time-varying firm and employer-employee match characteristics require a redefinition of the simple firm effect as ψ_{jit}. The addition of the i and t subscripts allows the firm effect to vary over time and across employer-employee matches. Now let the firm observable characteristics be time-varying, v_{jt}, and denote the observable match characteristics by r_{jit}. Then we can write the firm effect as

$$\psi_{jit} = \phi_j + v_{jt}\rho + r_{jit}\gamma_j \tag{13}$$

where γ_j is a vector of effects associated with the match characteristics. Statistical analysis of the effects defined by equation (13) is accomplished by augmenting the columns of F to reflect the data in v_{jt} and r_{jit}. The formulas shown in the estimation sections below can then be applied to the augmented design matrix.

4.3 Firm-Average Person Effect

For each firm j we define a firm-average person effect

$$\bar{\theta}_j \equiv \bar{\alpha}_j + \bar{u}_j \eta = \frac{\sum_{\{(i,t)|\,\mathrm{J}(i,t)=j\}} \theta_i}{N_j} \tag{14}$$

where

$$N_j \equiv \sum_{\forall(i,t)} 1(\mathrm{J}(i,t)=j)$$

and the function $1(A)$ takes the value 1 if A is true and 0 otherwise. The importance of the effect defined in equation (14) may not be apparent at first glance. Consider, the difference between ψ_j and $\bar{\theta}_j$. The former effect measures the extent to which firm j deviates from the average firm (averaged over individuals and weighted by employment duration) whereas the latter effect measures the extent to which the average employee of firm j deviates from the population of potential employees. In their analysis of wage rate determination, AKM refer to the firm-average person effect, $\bar{\theta}_j$, as capturing the idea of high (or low) wage workers while the pure firm effect, ψ_j, captures the idea of a high (or low) wage firm. Both effects must be specified and estimable for the distinction to carry empirical import.

4.4 Person-Average Firm Effect

For each individual i consider the person-average firm effect defined as

$$\bar{\psi}_i \equiv \bar{\phi}_i + \bar{v}_i \rho = \frac{\sum_t \psi_{\mathrm{J}(i,t)it}}{T_i}. \tag{15}$$

This effect is the individual counterpart to the firm-average person effect. Limited sample sizes for individuals make estimates of this effect less useful in their own right; however, they form the basis for conceptualizing the difference between the effect of heterogeneous individuals on the composition of a firm's workforce, as measured by the effect defined in equation (14), and the effect of heterogeneous firms on an individual's career employment outcomes, as measured by the effect in equation (15).

4.5 Industry Effects[9]

Industry is a characteristic of the employer. As such, the analysis of industry effects in the presence of person and firm effects can be accomplished by appropriate definition of the industry effect with respect to the firm effects. We call the properly defined industry effect a "pure" industry effect. Denote the pure industry effect, conditional on the same information as in equations (7) and (9), as κ_k for some industry classification $k = 1, \ldots, K$. Our definition of the pure industry effect is simply the correct aggregation of the pure firm effects within the industry. We define the pure industry effect as the one that corresponds to putting industry indicator variables in equation (9) and, then, defining what is left of the pure firm effect as a deviation from the industry effects. Hence, κ_k can be represented as an employment-duration weighted average of the firm

[9] This section is based upon the analysis in Abowd, Kramarz and Margolis (1999).

effects within the industry classification k:

$$\kappa_k \equiv \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\frac{1(\mathrm{K}(\mathrm{J}(i,t)) = k)\psi_{\mathrm{J}(i,t)}}{N_k} \right]$$

where

$$N_k \equiv \sum_{j=1}^{J} 1(\mathrm{K}(j) = k) N_j$$

and the function $\mathrm{K}(j)$ denotes the industry classification of firm j. If we insert this pure industry effect, the appropriate aggregate of the firm effects, into equation (7), then

$$y_{it} = x_{it}\beta + \theta_i + \kappa_{\mathrm{K}(\mathrm{J}(i,t))} + (\psi_{\mathrm{J}(i,t)} - \kappa_{\mathrm{K}(\mathrm{J}(i,t))}) + \varepsilon_{it}$$

or, in matrix notation as in equation (9),

$$y = X\beta + D\theta + FA\kappa + (F\psi - FA\kappa) + \varepsilon \tag{16}$$

where the matrix A, $J \times K$, classifies each of the J firms into one of the K industries; that is, $a_{jk} = 1$ if, and only if, $\mathrm{K}(j) = k$. Algebraic manipulation of equation (16) reveals that the vector κ, $K \times 1$, may be interpreted as the following weighted average of the pure firm effects:

$$\kappa \equiv (A'F'FA)^{-1}A'F'F\psi. \tag{17}$$

and the effect $(F\psi - FA\kappa)$ may be re-expressed as $M_{FA}F\psi$, where $M_Z \equiv I - Z\,(Z'Z)^{-}\,Z'$ denotes the column null space of an arbitrary matrix Z, and $()^{-}$ is a computable generalized inverse. Thus, the aggregation of J firm effects into K industry effects, weighted so as to be representative of individuals, can be accomplished directly by the specification of equation (16). Only rank$(F'M_{FA}F)$ firm effects can be separately identified using unrestricted fixed-effects methods; however, there is neither an omitted variable nor an aggregation bias in the estimates of (16), using either of class of estimators discussed below. Equation (16) simply decomposes $F\psi$ into two orthogonal components: the industry effects $FA\kappa$, and what is left of the firm effects after removing the industry effect, $M_{FA}F\psi$. While the decomposition is orthogonal, the presence of X and D in equation (16) greatly complicates the estimation by either fixed-effects or mixed-effects techniques.

4.6 Other Firm Characteristic Effects

Through careful specification of the firm effect in equation (13), we can estimate the average effect associated with any firm characteristic, v_{jt}, or any interaction of firm and personal characteristics, r_{jit}, while allowing for unobservable firm and personal heterogeneity.

4.7 Occupation Effects and Other Person × Firm Interactions

If occupation effects are interpreted as characteristics of the person, then they are covered by the analysis above and can be computed as functions of θ as described in equation (11). Occupation effects are often interpreted as an interaction between person and firm effects (Groshen 1991a, 1991b, 1996, implicitly). Mixed effects specifications are most appropriate in this case, and are discussed in section 6.

5 Estimation by Fixed-effects Methods

In this section we present methods for estimating the pure person and firm effects specification (7) by direct least squares, and consistent methods for estimating generalizations of this specification.

5.1 Estimation of the Fixed-Effects Model by Direct Least Squares

This subsection directly draws from Abowd, Creecy and Kramarz (2002) (ACK, hereafter). The normal equations for least squares estimation of fixed person, firm, and characteristic effects are of very high dimension. Thus estimating the full model by fixed-effect methods requires special algorithms. In our earlier work, e.g., Abowd, Finer and Kramarz (1999) (AFK, hereafter) and AKM, we relied on statistical approximations to render the estimation problem tractable. More recently, ACK developed new algorithms that permit the exact least squares estimation of all the effects in equation (7). These algorithms are based on the iterative conjugate gradient method and rely on computational simplifications admitted by the sparse structure of the least squares normal equations. They have some similarity to methods used in the animal and plant breeding literature.[10] ACK also developed new methods for computing estimable functions of the parameters of (7).

5.1.1 Least Squares Normal Equations

The full least squares solution to the estimation problem for equation (7) solves the normal equations for all estimable effects:

$$\begin{bmatrix} X'X & X'D & X'F \\ D'X & D'D & D'F \\ F'X & F'D & F'F \end{bmatrix} \begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} X'y \\ D'y \\ F'y \end{bmatrix} \tag{18}$$

[10] See Abowd and Kramarz (1999a) for a longer dicussion of the relation of these models to those found in the breeding literature. The techniques are summarized in Robinson (1991) and the random-effects methods are thoroughly discussed in Neumaier and Groenveld (1998). The programs developed for breeding applications cannot be used directly for the linked employer-employee data application because of the way the breeding effects are parameterized.

In typical applications, the cross-product matrix on the left-hand side of the equation is too high-dimensional to solve using conventional algorithms (*e.g.*, those implemented in SAS, Stata, and other general purpose linear modeling software based on variations of the sweep algorithm for solving (18)). AKM present a set of approximate solutions based on the use of different conditioning effects, Z. AFK applies the best of these approximations with a much higher-dimension Z.

5.1.2 Identification of Individual and Firm Effects

Many interesting economic applications of equation (7) make use of the estimated person and firm effects. Estimation requires a method for determining the identified effects[11]. The usual technique of sweeping out singular row/column combinations from the normal equations (18) is not applicable to the ACK method because they solve the normal equations without inverting the cross-product matrix. Hence, identification requires finding conditions under which the normal equations (18) can be solved exactly for some estimable functions of the person and firm effects. In this sub-section we ignore the problem of identifying the coefficients β because in practice this is rarely difficult.

The identification problem for the person and firm effects can be solved by applying graph-theoretic methods to determine groups of connected individuals and firms. Within a connected group of persons/firms, identification can be determined using conventional methods from the analysis of covariance. Connecting persons and firms requires that some of the individuals in the sample be employed at multiple employers. When a group of persons and firms is connected, the group contains all the workers who ever worked for any of the firms in the group and all the firms at which any of the workers were ever employed. In contrast, when a group of persons and firms is not connected to a second group, no firm in the first group has ever employed a person in the second group, nor has any person in the first group ever been employed by a firm in the second group. From an economic perspective, connected groups of workers and firms show the realized mobility network in the economy. From a statistical perspective, connected groups of workers and firms block-diagonalize the normal equations and permit the precise statement of identification restrictions on the person and firm effects.

The following algorithm constructs G mutually-exclusive groups of connected observations from the N workers in J firms observed over the sample period.[12]

`For` $g = 1, ...,$ `repeat until no firms remain:`

  `The first firm not assigned to a group is in group` g`.`

[11] Standard statistical references, for example Searle et *al.* (1992), provide general methods for finding the estimable functions of the parameters of equation (7). These methods also require the solution of a very high dimension linear system and are, therefore, impractical for our purposes.

[12] This algorithm finds all of the maximally connected sub-graphs of a graph. The relevant graph has a set of vertices that is the union of the set of persons and the set of firms and edges that are pairs of persons and firms. An edge (i, j) is in the graph if person i has worked for firm j.

```
  Repeat until no more firms or persons are added to group g:
   Add all persons employed by a firm in group g to group g.
   Add all firms that have employed a person in group g to group g.
  End repeat.
End for.
```

At the conclusion of the algorithm, the persons and firms in the sample have been divided into G groups. Denote the number of individuals in group g by N_g, and the number of employers in the group by J_g. Some groups contain a single employer and, possibly, only one individual. For groups that contain more than one employer, every employer in the group is connected (in the graph-theoretic sense) to at least one other employer in the group. Within each group g, the group mean of y and $N_g - 1 + J_g - 1$ person and firm effects are identified. After the construction of the G groups, exactly $N + J - G$ effects are estimable. See the proof in Appendix 1 of ACK[13].

5.1.3 Normal Equations after Group Blocking

The identification argument can be clarified by considering the normal equations after reordering the persons and firms by group. For simplicity, let the arbitrary equation determining the unidentified effect set it equal to zero, *i.e.*, set one person or firm effect equal to zero in each group. Then the column associated with this effect can be removed from the reorganized design matrix and we can suppress the column associated with the group mean. The resulting normal equations are:

$$\left[\begin{array}{c|cc|cc|c|cc}
X'X & X'D_1 & X'F_1 & X'D_2 & X'F_2 & \cdots & X'D_G & X'F_G \\
\hline
D_1'X & D_1' & D_1 & 0 & 0 & \cdots & 0 & 0 \\
F_1'X & F_1' & D_1 & 0 & 0 & \cdots & 0 & 0 \\
\hline
D_2'X & 0 & 0 & D_2'D_2 & D_2'F_2 & \cdots & 0 & 0 \\
F_2'X & 0 & 0 & F_2'D_2 & F_2'F_2 & \cdots & 0 & 0 \\
\hline
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\hline
D_G'X & 0 & 0 & 0 & 0 & \cdots & D_G'D_G & D_G'F_G \\
F_G'X & 0 & 0 & 0 & 0 & \cdots & F_G'D_F & F_G'F_G
\end{array}\right]
\left[\begin{array}{c}
\beta \\ \hline \theta_1 \\ \psi_1 \\ \hline \theta_2 \\ \psi_2 \\ \hline \cdots \\ \hline \theta_G \\ \psi_G
\end{array}\right]
=
\left[\begin{array}{c}
X'y \\ \hline D_1'y \\ F_1'y \\ \hline D_2'y \\ F_2'y \\ \hline \cdots \\ \hline D_G'y \\ F_G'y
\end{array}\right] \tag{19}$$

After reordering by group, the cross-products matrix is block diagonal. This matrix has full column rank and the solution for the parameter vector is unique.

[13] The grouping algorithm constructs groups within which "main effect" contrasts due to persons and firms are identified. In the linear models literature the "groups" are called "connected data". See Searle (1987), pp. 139-149 for a discussion of connected data. See Weeks and Williams (1964) for the general algorithm in analysis of variance models.

ACK do not solve equation (19) directly. Rather, they apply the technique discussed below to estimate the identifiable effects.

5.1.4 Estimation by Direct Solution of the Least Squares Problem

Appendix 2 in ACK shows the exact algorithm used to solve equation (18). It is a variant of the conjugate gradient algorithm, customized to exploit the sparse representation of equation (18) and to accommodate very large problems with many X variables. In practice, ACK apply this algorithm to the full set of persons, firms and characteristics shown in the design matrices of equations (7) and (18). Unlike equation (19), the cross-product matrix in equation (18) is not of full rank. Although the algorithm ACK use converges to a least squares solution, the parameter estimates are not unique. They subsequently apply the following identification procedure to the estimated effects. In each group, they eliminate one person effect by normalizing the group mean person effect to zero. ACK also normalize the overall mean person and firm effects to zero. This procedure identifies the grand mean of the dependent variable (or the overall regression constant if X and y have not been standardized to mean zero) and a set of $N + J - G - 1$ person and firm effects measured as deviations from the grand mean of the dependent variable[14].

5.2 Consistent Methods for β and γ (the firm-specific returns to seniority)

The preceding discussion focused on estimation of the pure person and firm effects model (7). In this subsection, we discuss methods presented in AKM for consistent estimation of more general representations of the person and firm effects. In particular, we discuss consistent estimation of β and γ_j in the general representation of the firm effect (13). The method relies on within-individual-firm differences of the data. It is robust in the sense that it requires no additional statistical assumptions beyond those specified in equation (4) and the general definition of the firm effect (13).[15] We should note, however, that this estimation technique relies heavily on the assumption of no interaction between X and F. Consider the first differences:

$$y_{i,n_{it}} - y_{in_{it-1}} = (x_{in_{it}} - x_{in_{it-1}})\beta + \gamma_{\mathrm{J}(i,n_{it})}(s_{in_{it}} - s_{in_{it-1}}) + \varepsilon_{in_{it}} - \varepsilon_{in_{it-1}} \quad (20)$$

[14] The computer software is available from the authors for both the direct least squares estimation of the two-factor analysis of covariance and the grouping algorithm. Computer software that implements both the random and fixed effects versions of these models used in breeding applications can be found in Groeneveld (1998). The specific algorithm we use can be found in Dongarra et *al.* (1991) p. 146.

[15] We have excluded $v_{jt}\rho$ from the firm effect (13), and assume a pure person effect θ_i.

for all observations for which $\text{J}(i, n_{it}) = \text{J}(i, n_{it-1})$, and where $s_{in_{it}}$ represents worker i's seniority at firm $J(i, n_{it})$ in period n_{it}.[16] In matrix form:

$$\triangle y = \triangle X\beta + \widetilde{F}\gamma + \triangle\varepsilon \tag{21}$$

where $\triangle y$ is $\widetilde{N^*} \times 1$, $\triangle X$ is $\widetilde{N^*} \times P$, $\widetilde{F}$ is $\widetilde{N^*} \times J$, $\triangle\varepsilon$ is $\widetilde{N^*} \times 1$, and $\widetilde{N^*}$ is equal to the number of (i, t) combinations in the sample that satisfy the condition $\text{J}(i, n_{it}) = \text{J}(i, n_{it-1})$. The matrix $\widetilde{F}$ is the rows of the design of γ that correspond to the person-years (i, t) for which the condition $\text{J}(i, n_{it}) = \text{J}(i, n_{it-1})$ is satisfied. The least squares estimates of β and γ are,

$$\widetilde{\beta} = (\triangle X' M_{\widetilde{F}} \triangle X)^{-1} \triangle X' M_{\widetilde{F}} \triangle y \tag{22}$$

$$\widetilde{\gamma} = (\widetilde{F}'\widetilde{F})^{-1}\widetilde{F}'(\triangle y - \triangle X\widetilde{\beta}). \tag{23}$$

A consistent estimate of $\text{V}[\widetilde{\beta}]$ is given by

$$\widetilde{\text{V}[\widetilde{\beta}]} = (\triangle X' M_{\widetilde{F}} \triangle X)^{-1}(\triangle X' M_{\widetilde{F}}\widetilde{\Omega}M_{\widetilde{F}} \triangle X)(\triangle X' M_{\widetilde{F}} \triangle X)^{-1}$$

where

$$\widetilde{\Omega} \equiv \begin{bmatrix} \widetilde{\Omega}[\triangle\varepsilon_1] & 0 & \cdots & 0 \\ 0 & \widetilde{\Omega}[\triangle\varepsilon_2] & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \widetilde{\Omega}[\triangle\varepsilon_{N^*}] \end{bmatrix}$$

and

$$\widetilde{\Omega}[\triangle\varepsilon_i] \equiv \begin{bmatrix} \widetilde{\triangle\varepsilon}^2_{in_2} & \widetilde{\triangle\varepsilon}_{in_2}\widetilde{\triangle\varepsilon}_{in_3} & \cdots & \widetilde{\triangle\varepsilon}_{in_2}\widetilde{\triangle\varepsilon}_{in_{T_i}} \\ \widetilde{\triangle\varepsilon}_{in_3}\widetilde{\triangle\varepsilon}_{in_2} & \widetilde{\triangle\varepsilon}^2_{in_3} & \cdots & \widetilde{\triangle\varepsilon}_{in_3}\widetilde{\triangle\varepsilon}_{in_{T_i}} \\ \cdots & \cdots & \cdots & \cdots \\ \widetilde{\triangle\varepsilon}_{in_{T_1}}\widetilde{\triangle\varepsilon}_{in_2} & \widetilde{\triangle\varepsilon}_{in_{T_1}}\widetilde{\triangle\varepsilon}_{in_3} & \cdots & \widetilde{\triangle\varepsilon}_{in_{T_i}}\widetilde{\triangle\varepsilon}_{in_{T_i}} \end{bmatrix}.$$

It is understood that only the rows of $\triangle\varepsilon$ that satisfy the condition $\text{J}(i, n_{it}) = \text{J}(i, n_{it-1})$ are used in the calculation of $\widetilde{\Omega}$, which is therefore $\widetilde{N^*} \times \widetilde{N^*}$. Notice that this estimator does not impose all of the statistical structure of the basic linear model (7).

6 The Mixed Model

In this Section, we focus on a mixed model specification of the pure person and firm effects model. The mixed model arises when some, or all, of the observable characteristics, person effects, and firm effects in equation (9) are treated as random, rather than fixed, effects. There is considerable confusion in the literature about the comparison of fixed and mixed effects specifications,

[16] In our preceding notation for the general firm effect (13), seniority is an element of observable match-specific characteristics r_{ijt}.

and so we take pains in this section to define terms in a manner consistent with the enormous statistical literature on this subject.

Consider the matrix formulation of the pure person and firm effects model, given in (9). We focus on the cases treated by Woodcock (2003) and Abowd and Stinson (2003), where the parameters β on observable characteristics are treated as fixed, and where the pure person and firm effects θ and ψ are random.[17] This specification corresponds closely to the hierarchical models that are common in some other applied settings, for instance in the education literature.[18]

The mixed model is completely specified by (9) and the stochastic assumptions[19]

$$E[\theta|X] = E[\psi|X] = E[\varepsilon|D,F,X] = 0 \quad (24)$$

$$Cov\left[\begin{array}{c}\theta\\ \psi\\ \varepsilon\end{array}\middle| X\right] = \left[\begin{array}{ccc}\sigma_\theta^2 I_N & 0 & 0\\ 0 & \sigma_\psi^2 I_J & 0\\ 0 & 0 & R\end{array}\right]. \quad (25)$$

It is worth noting that unlike some random effects specifications encountered elsewhere in the econometric literature, the mixed model we have specified does not assume that the design of the random effects (D and F) is orthogonal to the design (X) of the fixed effects (β). Such an assumption is almost always violated in economic data.

A variety of parameterizations of the residual covariance R are computationally feasible. Woodcock (2003) considers several in detail. Abowd and Stinson (2003) consider two more in the context of specifications that allow for multiple jobs in the same (i,t) pair and multiple measures of the dependent variable. The simplest parameterization is $R = \sigma_\varepsilon^2 I_{N^*}$. This specification is useful for making comparisons with the fixed-effect estimation procedure.

The most general parameterization estimated by Woodcock (2003) allows for a completely unstructured residual covariance within a worker-firm match. Let M denote the number of worker-firm matches (jobs) in the data, and let $\bar{\tau}$ denote the maximum observed duration of a worker-firm match. Suppose the data are ordered by t within j within i. In the balanced data case, where there are $\bar{\tau}$ observations on each worker-firm match, we can write

$$R = I_M \otimes W \quad (26)$$

[17] In fact, Woodcock (2003) decomposes the pure person effect θ_i into observable ($u_i\eta$) and unobserved components (α_i) as in equation (10). He treats η as fixed and α_i as random. For clarity of exposition we focus here on the simpler case where θ_i is random.

[18] In the education literature, schools are analogous to firms and students are analogous to workers. Because education data typically exhibit far less mobility (of students between schools) than we observe in labor market data, the usual specification nests student effects within school effects. The analogous hierarchical specification is therefore $y_{it} = x_{it}\beta + \theta_{ij} + \psi_j + \varepsilon_{it}$, where θ_{ij} is the person effect (nested within firm), and where ψ_j and θ_{ij} are specified as random effects. Dostie (2005) and Lillard (1999) estimate related mixed effects specifications for wages where the firm effect is nested within individuals, e.g., $y_{it} = x_{it}\beta + \theta_i + \psi_{ij} + \varepsilon_{it}$.

[19] In general, statisticians do not explicitly condition these expectations on X because they are primarily concerned with experimental data, where X constitutes part of the experimental design. Econometricians, however, are most often confronted with observational data. In this setting, X can rarely be considered a fixed component of the experimental design.

where W is the $\bar{\tau} \times \bar{\tau}$ within-match error covariance.[20] The extension to unbalanced data, where each match between worker i and firm j has duration $\tau_{ij} \leq \bar{\tau}$, is fairly straightforward. Define a $\bar{\tau} \times \tau_{ij}$ selection matrix S_{ij} with elements on the principal diagonal equal to 1, and off-diagonal elements equal to zero.[21] S_{ij} selects those rows and columns of W that correspond to observed earnings outcomes in the match between worker i and firm j. Then in the unbalanced data case, we have

$$R = \begin{bmatrix} S'_{11}WS_{11} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \ddots & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & S'_{1J_1}WS_{1J_1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & S'_{N1}WS_{N1} & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & S'_{NJ_N}WS_{NJ_N} \end{bmatrix}. \tag{27}$$

6.1 REML Estimation of the Mixed Model

Mixed model estimation is discussed at length in Searle et al. (1992) and McCulloch and Searle (2001). There are three principal methods that can be applied to estimate the variance components $\left(\sigma_\alpha^2, \sigma_\psi^2\right)$ and R : ANOVA, Maximum Likelihood (ML), and Restricted Maximum Likelihood (REML). ANOVA and ML methods are familiar to most economists; REML less so.[22] Since REML is by far the most commonly used estimation method among statisticians, it is worth giving it a brief treatment.

REML is frequently described as maximizing that part of likelihood that is invariant to the fixed effects (e.g., β). More precisely, REML is maximum likelihood on linear combinations of the dependent variable y, chosen so that the linear combinations do not contain any of the fixed effects. As Searle *et al.* (1992, pp. 250-251) show, these linear combinations are equivalent to residuals obtained after fitting the fixed portion of the model (e.g., $X\beta$) via least

[20] Woodcock (2003) estimates this parameterization of R under the assumption that W is symmetric and positive semi-definite.

[21] For example, if $\bar{\tau} = 3$ and a match between worker i and firm j lasts for 2 periods,

$$S_{ij} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

[22] REML estimation of mixed models is commonplace in statistical genetics and in the plant and animal breeding literature. In recent years, REML has in fact become the mixed model estimation method of choice in these fields, superceding ML and ANOVA.

squares.[23] The linear combinations $k'y$ are chosen so that

$$k'X\beta = 0 \;\; \forall\beta \tag{28}$$

which implies

$$k'X = 0. \tag{29}$$

Thus k' projects onto column null space of X, and is therefore

$$\begin{aligned} k' &= c'\left[I_{N^*} - X\left(X'X\right)^{-} X'\right] & (30)\\ &\equiv c'M_X & (31)\end{aligned}$$

for arbitrary c', and where A^{-} denotes the generalized inverse of A. When X has rank $r \leq p$, there are only $N^* - r$ linearly independent vectors k' satisfying (28).

Define $K' = C'M_X$ with rows k' satisfying (28), and where K' and C' have full row rank $N^* - r$. REML estimation of the variance parameters is maximum likelihood on $K'y$ under normality. For $y \sim \mathcal{N}\left(X\beta, \mathbf{V}\right)$ it follows that

$$K'y \sim N\left(0, K'\mathbf{V}K\right) \tag{32}$$

where $\mathbf{V} = DD'\sigma_\alpha^2 + FF'\sigma_\psi^2 + R$ is the conditional covariance of y implied by (25). The REML log-likelihood (i.e., the log-likelihood of $K'y$) is

$$\log L_{REML} = -\frac{1}{2}\left(N^* - r\right)\log 2\pi - \frac{1}{2}\log\left|K'\mathbf{V}K\right| - \frac{1}{2}y'K\left(K'\mathbf{V}K\right)^{-1}K'y. \tag{33}$$

The REML estimator of the variance parameters has a number of attractive properties. First, REML estimates are invariant to the choice of K'.[24] Second, REML estimates are invariant to the value of the fixed effects (i.e., β). Third, in the balanced data case, REML is equivalent to ANOVA.[25] Under normality, it thus inherits the minimum variance unbiased property of the ANOVA estimator.[26] Finally, since REML is based on the maximum likelihood principle, it inherits the consistency, efficiency, asymptotic normality, and invariance properties of ML.

Inference based on REML estimates of the variance components parameters is straightforward. Since REML estimation is just maximum likelihood on (33), REML likelihood ratio tests (REMLRTs) can be used. In most cases, REMLRTs are equivalent to standard likelihood ratio tests. The exception is

---

[23] Note this exercise is heuristic and serves only to motivate the REML approach. Under the stochastic assumptions (24) and (25), the least squares estimator of β is not BLUE. The BLUE of β is obtained by solving the mixed model equations (35).

[24] Subject to rows k' satisfying (28).

[25] The usual statistical definition of balanced data can be found in Searle (1987). Under this definitions, longitudinal linked data on employers and employees are balanced if we observe each worker employed at every firm, and all job spells have the same duration. Clearly, this is not the usual case.

[26] In contrast, ML estimators of variance components are biased since they do not take into account degrees of freedom used for estimating the fixed effects.

testing for the presence of some random effect γ. The null is $\sigma_\gamma^2 = 0$. Denote the restricted REML log-likelihood by $\log L^*_{REML}$. The REMLRT statistic is $\Lambda = -2\left(\log L^*_{REML} - \log L_{REML}\right)$. Since the null puts σ_γ^2 on the boundary of the parameter space under the alternative hypothesis, Λ has a non-standard distribution. Stram and Lee (1994) show the asymptotic distribution of Λ is a 50:50 mixture of a χ_0^2 and χ_1^2. The approximate p-value of the test is thus $0.5\left(1 - \Pr\left(\chi_1^2 \leq \Lambda\right)\right)$.

6.2 Estimating the Fixed Effects and Realized Random Effects

A disadvantage of REML estimation is that it provides no direct means for estimating the fixed covariate effects β. Henderson, in Henderson et al. (1959) derived a system of equations that simultaneously yield the BLUE of β and Best Linear Unbiased Predictor (BLUP) of the random effects. These equations have become known as the mixed model equations or Henderson equations. Define the matrix of variance components

$$G = \begin{bmatrix} \sigma_\theta^2 I_N & 0 \\ 0 & \sigma_\psi^2 I_J \end{bmatrix}. \tag{34}$$

The mixed model equations are

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}\begin{bmatrix} D & F \end{bmatrix} \\ \begin{bmatrix} D' \\ F' \end{bmatrix} R^{-1}X & \begin{bmatrix} D' \\ F' \end{bmatrix} R^{-1}\begin{bmatrix} D & F \end{bmatrix} + G^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{\theta} \\ \tilde{\psi} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ \begin{bmatrix} D' \\ F' \end{bmatrix} R^{-1}y \end{bmatrix} \tag{35}$$

where $\tilde{\beta}$ denotes solutions for the fixed effects, and $\tilde{\theta}$ and $\tilde{\psi}$ denote solutions for the random effects. In practice, of course, solving (35) requires estimates of R and G. Common practice is to use REML estimates $\tilde{G}$ and $\tilde{R}$.

The BLUPs $\tilde{\theta}$ and $\tilde{\psi}$ have the following properties. They are *best* in the sense of minimizing the mean square error of prediction

$$E\left(\begin{bmatrix} \tilde{\theta} \\ \tilde{\psi} \end{bmatrix} - \begin{bmatrix} \theta \\ \psi \end{bmatrix}\right)' A \left(\begin{bmatrix} \tilde{\theta} \\ \tilde{\psi} \end{bmatrix} - \begin{bmatrix} \theta \\ \psi \end{bmatrix}\right) \tag{36}$$

where A is any positive definite symmetric matrix. They are *linear* in y, and *unbiased* in the sense $E(\tilde{\theta}) = E(\theta)$ and $E(\tilde{\psi}) = E(\psi)$.

The solutions to equation (35) also have a Bayesian interpretation. If we suppose that the prior distribution for β is $N(0, \Omega)$ and the prior distribution for (θ, ψ) is $N(0, G)$, then the posterior mean $E\left[(\beta, \theta, \psi)\,|y\right] \to \left(\tilde{\beta}, \tilde{\theta}, \tilde{\psi}\right)$, the solution of (35), as $|\Omega| \to \infty$. (See Goldberger (1962), Searle *et al.* (1992, pp. 331-3) and Robinson (1991)).

The mixed model equations make clear the relationship between the fixed and mixed model estimation. In particular, as $|G| \to \infty$ with $R = \sigma_\varepsilon^2 I_{N^*}$, the mixed model equations (35) converge to the normal equations (18). Thus the mixed model solutions $(\tilde{\beta}, \tilde{\theta}, \tilde{\psi})$ converge to the least squares solutions $(\hat{\beta}, \hat{\theta}, \hat{\psi})$. In this sense the least squares estimator is a special case of the mixed model estimator.

6.3 Mixed Models and Correlated Random-effects Models

Since Chamberlain (1984) introduced his extension of methods by Cramér (1946) and Mundlak (1978) for handling balanced panel data models with random effects that were correlated with the X variables, econometricians have generally referred to the Chamberlain class of models as "correlated random-effects models." Statisticians, on the other hand, usually mean the Henderson (1953) formulation of the mixed-effects model that gives rise to equation (35), with G nondiagonal, when they refer to a correlated random-effects model.

It is important to distinguish between correlated random-effects models based on the mixed model equations (G nondiagonal) and orthogonal design models, which can occur within either a fixed-effects or random-effects interpretation of the person and firm effects. Orthogonal design means that one or more of the following conditions hold:

$$\begin{aligned} X'D &= 0, \text{ orthogonal person-effect design and personal characteristics} \\ X'F &= 0, \text{ orthogonal firm-effect design and personal characteristics} \\ D'F &= 0, \text{ orthogonal person-effect and firm-effect designs} \end{aligned}$$

An economy with random assignment of persons to firms could satisfy these conditions. However, virtually all longitudinal linked employer-employee data, as well as most other observational data in economics, violate at least one of these orthogonal design assumptions. Recognition of the absence of orthogonality between the effects is the basis for the fixed-effects estimator approximations discussed in section 5 and the difficulty associated with solving the mixed-model equations, in general (see Robinson, 1991, Searle, et al. 1992, Neumaier and Groeneveld, 1996, and Groeneveld, 1998).

To relate the Chamberlain-style correlated random-effects model to the mixed model estimator, we consider a single time-varying X, which we give the components of variance structure:

$$x_{it} = \upsilon_i + \varsigma_{it} \tag{37}$$

where

$$\text{Corr}[\upsilon_i, \theta_i] \neq 0$$

$$\text{V}[\varsigma_{it}] = \Delta$$

and

$$\text{Corr}[\varsigma_{it}, \varepsilon_{ns}] = 0 \ \forall i, n, s, t$$

This specification implies that $\text{Corr}[\upsilon_i, \psi_{\text{J}(i,t)}] \neq 0$ as long as G is nondiagonal. Then, to derive the Chamberlain estimating system for a balanced panel data model, assume that $T_i = T$ for all i and compute the linear projection of y_i on x_i

$$y_i = x_i \Pi + \nu_i \tag{38}$$

where Π is the $T \times T$ matrix of coefficients from the projection and ν_i is the $T \times 1$ residual of the projection. Chamberlain provides an interpretation of the coefficients in Π that remains valid under our specification.

Because the firm effect is shared by multiple individuals in the sample, however, the techniques proposed by Chamberlain for estimating equation (38) require modification. The most direct way to accomplish the extension of Chamberlain's methods is to substitute equation (37) into equation (7), then restate the system of equations as a mixed model. For each individual i in period t we have

$$\begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix} = \begin{bmatrix} \tau_i + \psi_{\mathrm{J}(i,t)} + \xi_{it} \\ \upsilon_i + \varsigma_{it} \end{bmatrix}. \tag{39}$$

where $\tau_i = \theta_i + \upsilon_i \beta$ and $\xi_{it} = \varepsilon_{it} + \varsigma_{it}\beta$. Stacking y_i and x_i, define

$$m_i \equiv \begin{bmatrix} y_i \\ x_i \end{bmatrix}, \text{ and } m \equiv \begin{bmatrix} m_1 \\ \cdots \\ m_N \end{bmatrix}$$

All other vectors are stacked conformably. Then, the mixed-effects formulation of equation (39) can be written as

$$m = D_1 \tau + D_2 \upsilon + F_3 \psi + \nu \tag{40}$$

where D_1, D_2, and F_3 are appropriately specified design matrices, τ is the $N \times 1$ vector of person effects entering the y equation, υ is the $N \times 1$ vector of person effects entering the x equation, and

$$\nu = \begin{bmatrix} \xi_1 \\ \varsigma_1 \\ \cdots \\ \xi_N \\ \varsigma_N \end{bmatrix}$$

is the stacked joint error vector. Problems of this form, with τ, υ, and ψ correlated and D_1, D_2, and F_3 nonorthogonal look unusual to economists but are quite common in animal science and statistical genetics. Software to solve the mixed model equations and estimate the variance matrices for equation (40) has been developed by Groeneveld (1998) and Gilmour *et al.* (1995) and some applications, other than the one presented above, are discussed in Robinson (1991) and Tanner (1996). The methods exploit the sparse structure of D_1, D_2, and F_3 and use analytic derivatives to solve (35). Robert (2001) and Tanner (1996) provide algorithms based on simulated data techniques.

7 Models of Heterogeneity biases in incomplete models

The analyses in this section are based upon the exact fixed-effects estimator for model (9) given by the solution to (18).

7.1 Omission of the firm effects

When the estimated version of equation (9) excludes the firm effects, ψ, the estimated person effects, θ^*, are the sum of the underlying person effects, θ, and the employment-duration weighted average of the firm effects for the firms in which the worker was employed, conditional on the individual time-varying characteristics, X:

$$\theta^* = \theta + (D'M_X D)^{-1} D'M_X F\psi. \tag{41}$$

Hence, if X were orthogonal to D and F, so that $D'M_X D = D'D$ and $D'M_X F = D'F$, then the difference between θ^* and θ, which is just an omitted variable bias, would be an $N \times 1$ vector consisting, for each individual i, of the employment-duration weighted average of the firm effects ψ_j for $j \in \{\mathrm{J}(i, n_{i1}), ..., \mathrm{J}(i, n_{iT})\}$:

$$\theta_i^* - \theta_i = \sum_{t=1}^{T_i} \frac{\psi_{\mathrm{J}(i,n_{it})}}{T_i},$$

the person-average firm effect. Similarly, the estimated coefficients on the time-varying characteristics in the case of omitted firm effects, β^*, are the sum of the parameters of the full conditional expectation, β, and an omitted variable bias that depends upon the conditional covariance of X and F, given D:

$$\beta^* = \beta + (X'M_D X)^{-1} X'M_D F\psi.$$

7.2 Omission of the person effects

Omitting the pure person effects (θ) from the estimated version of equation (9) gives estimates of the firm effects, ψ^{**}, that can be interpreted as the sum of the pure firm effects, ψ, and the employment-duration weighted average of the person effects of all of the firm's employees in the sample, conditional on the time-varying individual characteristics:

$$\psi^{**} = \psi + (F'M_X F)^{-1} F'M_X D\theta. \tag{42}$$

Hence, if X were orthogonal to D and F, so that $F'M_X F = F'F$ and $F'M_X D = F'D$, the difference between ψ^{**} and ψ, again an omitted variable bias, would be a $J \times 1$ vector consisting of the employment-duration weighted average of person effects θ_i for $(i,t) \in \{\mathrm{J}(i,t) = j \text{ and } t \in \{n_{i1}, \dots, n_{iT_i}\}\}$ for each firm j. That is,

$$\psi_j^{**} - \psi_j = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \left[\frac{\theta_i \, 1(\mathrm{J}(i, n_{it}) = j)}{N_j} \right],$$

the firm-average person effect. The estimated coefficients on the time-varying characteristics in the case of omitted individual effects, β^{**}, are the sum of the effects of time-varying personal characteristics in equation (9), β, and an omitted variable bias that depends upon the covariance of X and D, given F:

$$\beta^{**} = \beta + (X'M_F X)^{-1} X'M_F D\theta. \tag{43}$$

This interpretation applies to studies like Groshen (1991a, 1991b, 1996).

7.3 Inter-industry wage differentials

We showed above that industry effects are an aggregation of firm effects that may be inconsistently estimated if either person or firm effects are excluded from the equation. We consider these issues now in the context of inter-industry wage differentials as in Dickens and Katz (1987), Krueger and Summers (1987, 1988), Murphy and Topel (1987), Gibbons and Katz (1992). The fixed or random effects estimation of the aggregation of J firm effects into K industry effects, weighted so as to be representative of individuals, can be accomplished directly by estimation of equation (16). Only rank$(F'M_{FA}F)$ fixed firm effects can be separately identified; however, the mixed-effects model can produce estimates of all realized industry and firm effects.

As shown in AKM, fixed-effects estimates of industry effects, κ^*, that are computed on the basis of an equation that excludes the remaining firm effects, $M_{FA}F\psi$, are equal to the pure industry effect, κ, plus an omitted variable bias that can be expressed as a function of the conditional variance of the industry effects, FA, given the time-varying characteristics, X, and the person effects, D:

$$\kappa^* = \kappa + \left(A'F'M_{\begin{bmatrix} D & X \end{bmatrix}} FA \right)^{-1} A'F'M_{\begin{bmatrix} D & X \end{bmatrix}} M_{FA}F\psi$$

which simplifies to $\kappa^* = \kappa$ if, and only if, the industry effects, FA, are orthogonal to the subspace $M_{FA}F$, given D and X, which is generally not true even though FA and $M_{FA}F$ are orthogonal by construction. Thus, consistent fixed-effects estimation of the pure inter-industry wage differentials, conditional on time-varying personal characteristics and unobservable non-time-varying personal characteristics requires identifying information on the underlying firms unless this conditional orthogonality condition holds. Mixed-effects estimation without identifying information on both persons and firms likewise produces realized inter-industry wage effects that confound personal and firm heterogeneity.

Similarly, AKM show that fixed-effects estimates of the coefficients of the time-varying personal characteristics, β^*, are equal to the true coefficients of the linear model (9), β, plus an omitted variable bias that depends upon the conditional covariance between these characteristics, X, and the residual subspace of the firm effects, $M_{FA}F$, given D:

$$\beta^* = \beta + \left(X'M_{\begin{bmatrix} D & FA \end{bmatrix}} X \right)^{-1} X'M_{\begin{bmatrix} D & FA \end{bmatrix}} M_{FA}F\psi$$

which, once again, simplifies to $\beta^* = \beta$ if, and only if, the time-varying personal characteristics, X, are orthogonal to the subspace $M_{FA}F$, given D and FA, which is also not generally true. Once again, both fixed-effects and mixed-effects estimation of the β coefficients produces estimates that confound personal and firm heterogeneity when both types of identifying information are not available.

To assess the seriousness of the heterogeneity biases in the estimation of industry effects, AKM propose a decomposition of the raw industry effect into the part due to individual heterogeneity and the part due to firm heterogeneity. Their formulas apply directly to the fixed-effects estimator of equation (9) and can be extended to the estimated realized effects in a mixed-effects model. When equation (16) excludes both person and firm effects, the resulting raw industry effect, κ_k^{**}, equals the pure industry effect, κ, plus the employment-duration weighted average residual firm effect inside the industry, given X, and the employment-duration weighted average person effect inside the industry, given the time-varying personal characteristics X:

$$\kappa^{**} = \kappa + (A'F'M_X FA)^{-1}A'F'M_X(M_{FA}F\psi + D\theta)$$

which can be restated as

$$\kappa^{**} = (A'F'M_X FA)^{-1}A'F'M_X F\psi + (A'F'M_X FA)^{-1}A'F'M_X D\theta, \qquad (44)$$

which is the sum of the employment-duration weighted average firm effect, given X and the employment-duration weighted average person effect, given X. If industry effects, FA, were orthogonal to time-varying personal characteristics, X, and to the design of the personal heterogeneity, D, so that $A'F'M_X FA = A'F'FA$, $A'F'M_X F = A'F'F$, and $A'F'M_X D = A'F'D$, then, the raw inter-industry wage differentials, κ^{**}, would simply equal the pure inter-industry wage differentials, κ, plus the employment-duration-weighted, industry-average pure person effect, $(A'F'FA)^{-1} A'F'D\theta$, or

$$\kappa_k^{**} = \kappa_k + \sum_{i=1}^{N}\sum_{t=1}^{T_i} \frac{1[\mathrm{K}(\mathrm{J}(i, n_{it})) = k]\theta_i}{N_k}$$

Thus, statistical analyses of inter-industry differentials that exclude either person or firm effects confound the pure inter-industry wage differential with an average of the person effects found in the industry, given the measured personal characteristics, X.

8 Endogenous Mobility

The problem of endogenous mobility occurs because of the possibility that individuals and employers are not matched in the labor market on the basis of observable characteristics and the person and firm effects. A complete treatment of this problem is beyond the scope of this article; however, it is worth noting that the interpretation of equations (7) and (9) as conditional expectations given the person and firm effects is not affected by some forms of endogenous mobility. If the mobility equation is also conditioned on X, D, and, F, then the effects in the referenced equations are also structural as long as mobility does not depend upon ε.

Matching models of the labor market, such as those proposed by Jovanovic (1979) and Woodcock (2003) imply the existence of a random effect that is the interaction of person and firm identities. Such models are amenable to the statistical structure laid out in section 6; however, to our knowledge the application of such techniques to this type of endogenous mobility model has only been attempted recently using linked employer-employee data. We present these attempts now.

8.1 A Generalized Linear Mixed Model

Mixed model theory and estimation techniques have been applied to nonlinear models with linear indices. These are usually called generalized linear mixed models, and include such familiar specifications as the probit, logit, and tobit models augmented to include random effects. See McCulloch and Searle (2001) for a general discussion.

Woodcock (2003) estimates a mixed probit model with random person and firm effects as the first step of a modified Heckman two-step estimator. The goal is to correct for truncation of the error distribution in a mixed model of earnings with random person and firm effects. This truncation arises from endogenous mobility in the context of an equilibrium matching model. Specifically, the Woodcock (2003) matching model predicts that earnings are observed only if the worker-firm match continues, and that the continuation decision depends on person-, firm-, and tenure-specific mobility effects that are correlated with the person and firm effects in the earnings equation. At tenure τ, the match continues only if $\varepsilon_{it} \geq \bar{\varepsilon}_{i\tau}$ where

$$\begin{aligned} \bar{\varepsilon}_{i\tau} &= -\mu_\tau - \zeta_{i\tau} - \xi_{j\tau} \qquad (45) \\ \begin{bmatrix} \zeta_{i\tau} \\ \xi_{j\tau} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\zeta_\tau} & 0 \\ 0 & \sigma^2_{\xi_\tau} \end{bmatrix} \right). \end{aligned}$$

When $\varepsilon_{it} \sim \mathcal{N}(0, V_\tau)$, the marginal probability of observing the earnings outcome y_{it} is

$$\begin{aligned} \Pr(\varepsilon_{it} \geq \bar{\varepsilon}_{i\tau}) &= 1 - \Phi\left(\frac{-\mu_\tau - \zeta_{i\tau} - \xi_{j\tau}}{V_\tau^{1/2}} \right) \\ &= \Phi\left(\frac{\mu_\tau + \zeta_{i\tau} + \xi_{j\tau}}{V_\tau^{1/2}} \right) \qquad (46) \end{aligned}$$

where Φ is the standard normal CDF. Then we have

$$\begin{aligned} E\left[y_{it}|\varepsilon_{it} \geq \bar{\varepsilon}_{i\tau}\right] &= \mu + x_{it}'\beta + \theta_i + \psi_j + V_\tau^{1/2}\frac{\phi\left(\frac{\mu_\tau+\zeta_{i\tau}+\xi_{j\tau}}{V_\tau^{1/2}}\right)}{\Phi\left(\frac{\mu_\tau+\zeta_{i\tau}+\xi_{j\tau}}{V_\tau^{1/2}}\right)} \\ &= \mu + x_{it}'\beta + \theta_i + \psi_j + V_\tau^{1/2}\lambda_{i\tau} \end{aligned} \quad (47)$$

where $\lambda_{i\tau}$ is the familiar Inverse Mills' Ratio.

The truncation correction based on (46) and (47) proceeds as follows. The first step is to estimate a continuation probit at each tenure level with random person- and firm-specific mobility effects $\zeta_{i\tau}$ and $\xi_{j\tau}$. Woodcock (2003) estimates probits using the Average Information REML algorithm of Gilmour, Thompson, and Cullis (1995), applied to the method of Schall (1991). The Schall (1991) method extends standard methods for estimating generalized linear models to the random effects case. The basic idea is to perform REML on a linearization of the link function Φ. The process requires an iterative reweighting of the design matrices of fixed and random effects in the linearized system, see Schall (1991) for details. With estimates of the realized random effects $\tilde{\zeta}_{it}$ and $\tilde{\xi}_{j\tau}$ in hand, Woodcock (2003) constructs an estimate $\tilde{\lambda}_{i\tau}$ of the Inverse Mills' Ratio term for each observation. Including $\tilde{\lambda}_{i\tau}$ as an additional time-varying covariate in the earnings equation corrects for truncation in the error distribution due to endogenous mobility.

8.2 A Model of Wages, Endogenous Mobility and Participation with Person and Firm Effects

Following Buchinsky, Fougère, Kramarz, and Tchernis (2003), and the structural interpretation they develop, Beffy, Kamionka, Kramarz, and Robert (2003, BKKR hereafter) jointly model wages with a participation equation and an inter-firm mobility equation that include state-dependence and unobserved heterogeneity. A firm-specific unobserved heterogeneity component is added to the person-specific term. Like the linear models discussed in detail above, the wage equation includes person and firm effects.

Inter-firm mobility at date t depends on the realized mobility at date $t-1$. Similarly, participation at date t depends on past participation and mobility. Hence, we include initial conditions, modeled following Heckman (1981). This yields the following system of equations:

Initial Conditions:

$$\begin{aligned} z_{i1} &\sim \mathcal{U}_{1,\ldots,J} \\ y_{i1} &= \mathbb{I}\left(X_{i1}^Y\delta_0^Y + \alpha_{z_{i1}}^{Y,E} + v_{i1} > 0\right) \\ w_{i1} &= y_{i1}\left(X_{i1}^W\delta^W + \theta_{z_{i1}}^{W,E} + \epsilon_{i1}\right) \\ m_{i1} &= y_{i1}\mathbb{I}\left(X_{i1}^M\delta_0^M + \alpha_{z_{i1}}^{M,E} + u_{i1} > 0\right). \end{aligned}$$

Main Equations: $\forall t > 1$,

$$
\begin{aligned}
z_{it} &= y_{it-1}\left((1-m_{it-1})z_{it-1} + m_{it-1}\widetilde{\eta}_{it}\right) + (1-y_{it-1})\eta \\
\eta &\sim \mathcal{U}_{1,\dots,J} \quad \widetilde{\eta}_{it} \sim \mathcal{U}_{(1,\dots,J)-(z_{it-1})} \\
y_{it} &= \mathbb{I}\left(\underbrace{\gamma^M m_{it-1} + \gamma^Y y_{it-1} + X_{it}^Y \delta^Y + \theta_{z_{it}}^{Y,E} + \theta_i^{Y,I} + v_{it}}_{y_{it}^*} > 0\right) \\
w_{it} &= y_{it}\left(X_{it}^W \delta^W + \theta_{z_{it}}^{W,E} + \theta_i^{W,I} + \epsilon_{it}\right) \\
m_{it} &= y_{it}\mathbb{I}\left(\underbrace{\gamma m_{it-1} + X_{it}^M \delta^M + \theta_{z_{it}}^{M,E} + \theta_i^{M,I} + u_{it}}_{m_{it}^*} > 0\right).
\end{aligned}
$$

The variable z_{it} denotes the latent identifier of the firm and $J(i,t)$ denotes the realized identifier of the firm at which worker i is employed at date t. Therefore, $J(i,t) = z_{it}$ if individual i participates at date t. y_{it} and m_{it} denote, respectively, participation and mobility, as previously defined. y_{it} is an indicator function, equal to 1 if the individual i participates at date t. m_{it} is an indicator function that takes the following values:

	$y_{it+1} = 1$	$y_{it+1} = 0$
$y_{it} = 1$	$m_{it} = 1$ if $J(i,t+1) \neq J(i,t)$ $m_{it} = 0$ if $J(i,t+1) = J(i,t)$	m_{it} censored
$y_{it} = 0$	$m_{it} = 0$ p.s.	$m_{it} = 0$ p.s.

Table 1: Mobility

The variable w_{it} denotes the logarithm of the annualized total labor costs. The variables X are the observable time-varying as well as the time-invariant characteristics for individuals at the different dates. Here, θ^I and θ^E denote the random effects specific to, respectively, individuals or firms in each equation. u, v and ϵ are the error terms. There are J firms and N individuals in the panel of length T. All stochastic assumptions are described now.

8.3 Stochastic Assumptions

In order to specify the stochastic assumptions for the person and firm-effects, BKKR first rewrite their system of equations as:

$$
\begin{aligned}
z_{it} &= y_{it-1}\left((1-m_{it-1})z_{it-1}+m_{it-1}\widetilde{\eta}_{it}\right)+(1-y_{it-1})\eta \\
y_{it} &= \mathbb{I}\left(\underbrace{\gamma^M m_{it-1}+\gamma^Y y_{it-1}+X_{it}^Y\delta^Y+\Omega_{z_{it}}^E\theta^{Y,E}+\Omega_{it}^I\theta^{Y,I}+v_{it}}_{y_{it}^*}>0\right) \\
w_{it} &= y_{it}\left(X_{it}^W\delta^W+\Omega_{z_{it}}^E\theta^{W,E}+\Omega_{it}^I\theta^{W,I}+\epsilon_{it}\right) \\
m_{it} &= y_{it}.\mathbb{I}\left(\underbrace{\gamma m_{it-1}+X_{it}^M\delta^M+\Omega_{z_{it}}^E\theta^{M,E}+\Omega_{it}^I\theta^{M,I}+u_{it}}_{m_{it}^*}>0\right)
\end{aligned}
$$

for each $t>1$, where Ω_{it}^E is a design matrix of firm effects for the couple (i,t). Hence, it is a $1\times J$ matrix composed of $J-1$ zeros and of a 1 at column $z_{i,t}$. Similarly, Ω_{it}^I is a $1\times N$ matrix composed of $N-1$ zeros and of a 1 at column i. The model includes two dimensions of heterogeneity. This double dimension crucially affects the statistical structure of the likelihood function. The presence of firm effects makes the likelihood non-separable (person by person). Indeed, the outcomes of two individuals employed at the same firm, not necessarily at the same date, are not independent.

The next equations present the stochastic assumptions for the person and firm effects:

$$
\begin{aligned}
\theta^E &= \left(\alpha^{Y,E},\alpha^{M,E},\theta^{Y,E},\theta^{W,E},\theta^{M,E}\right) \quad \text{of dimension} \quad [5J,1] \\
\theta^I &= \left(\theta^{Y,I},\theta^{W,I},\theta^{M,I}\right) \quad \text{of dimension} \quad [3N,1].
\end{aligned}
$$

Moreover,

$$\theta^E|\Sigma^E\sim\mathcal{N}(0,D_0^E) \tag{48}$$

$$\theta^I|\Sigma^I\sim\mathcal{N}(0,D_0^I) \tag{49}$$

$$D_0^E=\Sigma^E\otimes I_J \tag{50}$$

$$D_0^I=\Sigma^I\otimes I_N \tag{51}$$

where Σ^E (resp. Σ^I) is a symmetric positive definite matrix $[5,5]$ (resp. $[3,3]$) with mean zero. Notice that these assumptions imply that correlations between the wage, the mobility, and the participation equations come from both person and firm heterogeneity (in addition to that coming from the idiosyncratic error terms). Furthermore, these assumptions exclude explicit correlation between different firms (for instance, the authors could have considered a non-zero correlation of the firm effects within an industry, a non-tractable assumption).

Notice though that BKKR could have included in the wage equation, for instance, the lagged firm effects of those firms at which a worker was employed in her career. This is difficult, but feasible in this framework.

Finally, they assume that the idiosyncratic error terms follow:

$$\begin{pmatrix} v_{it} \\ \epsilon_{it} \\ u_{it} \end{pmatrix} \sim_{iid} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{yw}\sigma & \rho_{ym} \\ \rho_{yw}\sigma & \sigma^2 & \rho_{wm}\sigma \\ \rho_{ym} & \sigma\rho_{wm} & 1 \end{pmatrix} \right).$$

Notice that experience and seniority are complex and highly non linear functions of the participation and mobility equations. Because all these person and firm effects are correlated **between equations**, the presence of experience and seniority in the wage equation induces a correlation between these two variables and the person and the firm effect in the same equation. Indeed, in the terminology introduced above, the BKKR model exhibits correlated random effects.

BKKR estimate this model on French data using Monte-Carlo Markov Chains methods (Gibbs sampling and the Hastings-Metropolis algorithms).

9 Conclusion

We have presented a relatively concise tour of econometric issues surrounding the specification of linear models that form the basis for the analysis of linked longitudinal employer-employee data. Our discussion has focused on the role of person and firm effects in such models, because these data afford analysts the first opportunity to separately distinguish effects in the context of a wide variety of labor market outcomes. We have shown that identification and estimation strategies depend upon the observed sample of persons and firms (the design of the person and firm effects) as well as on the amount of prior information one imposes on the problem, in particular, the choice of full fixed-effects or mixed-effects estimation.

We do not mean to suggest that these estimation strategies are complete. Indeed, many of the methods described in this chapter have been used by only a few analysts and some have not been used at all in the labor economics context. We believe that future analyses of linked employer-employee data will benefit from our attempt to show the relations among the various techniques and to catalogue the potential biases that arise from ignoring either personal or firm heterogeneity.

References

[1] Abowd, John M., Ronald Creecy, and Francis Kramarz , "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," Cornell University working paper, (2002).

[2] Abowd, John M., Hampton Finer and Francis Kramarz , "Individual and Firm Heterogeneity in Compensation: An Analysis of Matched Longitu-

dinal Employer and Employee Data for the State of Washington," in J. Haltiwanger, J. Lane, J. Spletzer, K. Troske eds, *The Creation and Analysis of Employer-Employee Matched Data*, North-Holland, (1999): 3-24.

[3] Abowd, John M. and Francis Kramarz "The Analysis of Labor Markets Using Matched Employer-Employee Data," in O. Ashenfelter and D. Card, eds. *Handbook of Labor Economics*, chapter 26, volume 3B, (Amsterdam, North Holland, 1999a): 2629-2710..

[4] Abowd, John M. and Francis Kramarz "The Analysis of Labor Markets Using Matched Employer-Employee Data," *Labour Economics* 6, (1999b): 53-74.

[5] Abowd, John M, Francis Kramarz and David N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica* 67, 2, (1999): 251-333.

[6] Abowd, John M, Francis Kramarz, Paul Lengermann, and Sébastien Perez-Duarte, "Are Good Workers Employed by Good Firms ? A test of a simple assortative mating model for France and the United States," Crest working paper (2003).

[7] Abowd, John M. and Martha H. Stinson, "Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Survey and SSA Administrative Data," US Census Bureau LEHD Program Working Paper (2003).

[8] Abowd, John M. and Lars Vilhuber, "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers." *Technical Paper TP-2002-17*, LEHD, U.S. Census Bureau (2002).

[9] Abowd, John M. and Simon Woodcock, "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds.), (Amsterdam: North Holland, 2001), 215-277.

[10] Buchinsky Moshe, Denis Fougère, Francis Kramarz, and Rusty Tchernis, « Interfirm Mobility, Wages, and the Returns to Seniority and Experience in the U.S., » Crest working paper (2003).

[11] Beffy Magali, Thierry Kamionka, Francis Kramarz, and Christian P. Robert, "Job Mobility and Wages with Worker and Firm Heterogeneity", Crest working paper.

[12] Chamberlain, Gary "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches and M.D. Intrilligator. Amsterdam: North Holland, chapter 22, (1984): 1248-1318.

[13] Cramér, H. *Mathematical Models of Statistics* (Princeton, NJ: Princeton University Press, 1946).

[14] Dickens, William T. and Lawrence F. Katz. "Inter-Industry Wage Differences and Industry Characteristics," in *Unemployment and the Structure of Labor Markets*, ed. by Kevin Lang and Jonathan S. Leonard. Oxford: Basil Blackwell(1987).

[15] Dostie, Benoit *Essays on Mobility: Applications to Job Turnover, Returns to Seniority, and Poverty,* Cornell University Ph.D. Thesis (2001).

[16] Gibbons, Robert and Lawrence Katz "Does Unmeasured Ability Explain Inter-Industry Wage Differentials?" *Review of Economic Studies*, 59, (1992): 515-535.

[17] Gilmour, A.R., R. Thompson, and B.R. Cullis "Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models." *Biometrics*, 51, (1995): 1440-1450.

[18] Goldberger, Arthur S. "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, Vol. 57 (1962): 369-75.

[19] Groeneveld, Eildert *VCE4 User's Guide and Reference Manual* (Höltystrass, Germany: Institute of Animal Husbandry and Animal Behavior, 1998).

[20] Groshen, Erica "Sources of Intra-Industry Wage Dispersion: How Much do Employers Matter?" *Quarterly Journal of Economics,* 106, (1991a): 869-884.

[21] Groshen, Erica "The Structure of the Female/Male Wage Differential: Is it Who You Are, What You Do, or Where You Work?" *Journal of Human Resources*, 26, (1991b): 457-72.

[22] Groshen, Erica "American Employer Salary Surveys and Labor Economics Research: Issues and Contributions," *Annales d'économie et de statistique*, 41/42, (1996): 413-442.

[23] Heckman Jim J. , "Heterogeneity and State Dependence", in *Studies in Labor Market*, Rosen S. ed., (University of Chicago Press, 1981).

[24] Henderson, C.R. "Estimation of Variance and Covariance Components," *Biometrics* 9 (1953): 226-252.

[25] Henderson, C.R., Oscar Kempthorne, Shayle.R. Searle and C.M. Von Krosigk, "The Estimation of Environmental and Genetic Trends from Records Subject to Culling," *Biometrics*, 15, 2, (1959): 192-218

[26] Jovanovic, Boyan "Job Matching and the Theory of Turnover," *Journal of Political Economy,* 87, (1979): 972-990.

[27] Krueger, Alan B. and Lawrence H. Summers "Reflections on the Inter-industry Wage Structure," in *Unemployment and the Structure of Labor Markets*, ed. by Kevin Lang and Jonathan S. Leonard, New York: Basil Blackwell, (1987) .

[28] Krueger, Alan B. and Lawrence H. Summers "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica*, 56, (1988): 259-293.

[29] Lengermann, Paul A. *Essays on the Dynamic Interaction of Workers and Firms: New Evidence Using Linked Employer-Employee Data*, University of Maryland Ph.D Thesis (2002).

[30] Lillard, L.A. "Job Turnover Heterogeneity and Person-Job-Specific Time-Series Wages," *Annales D'Économie et de Statistique*, 55-56 (1999): 183-210.

[31] Lillard, L.A. and C.W. Panis. *aML Multilevel Multiprocess Statistical Software*, Release 1.0 EconWare, Los Angeles, California (2000).

[32] McCulloch Charles E. and Shayle R. Searle, *Generalized, Linear, and Mixed Models*, John Wiley and Sons (2001).

[33] Mundlak, Yair "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46 (1978):

[34] Murphy, Kevin M. and Robert H. Topel "Unemployment, Risk, and Earnings: Testing for Equalizing Wage Differences in the Labor Market" in *Unemployment and the Structure of the Labor Market*, ed. by Kevin Lang and Jonathan S. Leonard. New York: Basil Blackwell, (1987)

[35] Nerlove, Marc "A Note on Error Components Models," *Econometrica*, 39, (1971): 383-396

[36] Neumaier, Arnold and Eildert Groeneveld, "Restricted Maximum Likelihood Estimation of Covariance in Sparse Linear Models," working paper, Institut für Mathematik, Wien University, Austria, (1996).

[37] Robert Christian P. *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation* , Springer-Verlag, New York, (2001).

[38] Robinson, G.K. "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6, (1991): 15-51.

[39] Schall Robert , "Estimation in Generalized Linear Models with Random Effects," *Biometrika*, 78, 4, (1991): 719-727}

[40] Searle, Shayle R. , *Linear Models for Unbalanced Data*, John Wiley and Sons, 1987.

[41] Searle, Shayle R., George Casella and Charles E. McCulloch , *Variance Components* (New York: John Wiley and sons, 1992).

[42] Stram, Daniel O. and Jae Won Lee , "Variance Component Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 50, (1994): 1171-1177.

[43] Stinson, Martha H. *Topics in Health and Labor Economics: The Influence of Heterogeneity and Measurement Error in Models of Individual Choice*, Cornell University Ph.D. thesis (2002).

[44] Tanner, Martin A. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (New York: Springer, 1996).

[45] Weeks, D.L. and D.R. Williams, "A Note on the Determination of Connectedness in an N-way cross classification." *Technometrics*, 6, (1964): 319-324.

[46] Woodcock, Simon D. "Heterogeneity and Learning in Labor Markets", in *Essays on Labor Market Dynamics and Longitudinal Linked Data*, Cornell University Ph.D. Thesis (2003).