Econometric Analyses of Linked Employer-Employee Data

John M. Abowd and Francis Kramarz[1]

December 1998

1. Introduction

There has been an explosion in the use of linked employer-employee data that we have documented in our *Handbook of Labor Economics* chapter (Abowd and Kramarz, 1999).[2] In this article we address the econometric issues associated with analyses of these data, in particular with longitudinal linked employer-employee data. The key feature of such data is that individuals and employing firms are both identified and followed over time. Measured characteristics of the individual are collected at multiple points in time and measured characteristics of the employing firm are also measured at multiple points in time. We will refer to the unit of observation on the employer as a firm, although the economic entity might be a corporation, business, business unit or establishment. Many of the models that we consider in this article are generalizations of the specification we used in Abowd, Kramarz and Margolis (1999, AKM hereafter).

We make a special effort to relate the techniques used by econometricians, in particular various fixed-effects estimators and related approximations, to other techniques popular in the variance components literature–in particular, mixed-effects estimators (see Searle, Casella and McCulloch, 1992). The benefit to this synthesis is that one can see that both the design of the linked data–which individuals and firms occur in the data–as well as the statististical properties of the effect– fixed or random–give rise to distinct computational issues. The interpretation of the realized effects, however, is the same for all of the techniques we consider. We begin, in section 2, with a specification for linear statistical models relating linked employer and employee data to outcomes measured at the individual level. In this 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 3, by defining a variety of effects that are functions of the basic person and firm effects in the main model. Section 4 considers the estimation of the person and firm effects by fixed-effects models. Section 5 discusses the use of mixed-effects models and the relation between various cor-

[1] 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 and Maître de Conférences at Ecole Polytechnique. Abowd acknowledges financial support from the NSF (SBER 96-18111).

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

related random-effects specifications. In section 6 we discuss the important heterogeneity biases that arise when either the person or firm effects are missing or specified incompletely in the basic model. Section 7 discusses the analysis of outcomes at the firm level. In this section the important consideration is the use of information estimated from a sample of the employees, as provided by the statistical analysis of the individual-level linked data. We briefly discuss the consequences of endogenous mobility in section 8. Finally, we conclude in section 9.

2. Linear Statistical Models with Person and Firm Effects

The basic linear statistical model we will discuss is specified as:

$$(y_{it} - \mu_y) = (x_{it} - \mu_x)\beta + \theta_i + \psi_{\mathrm{J}(i,t)} + \varepsilon_{it} \tag{1}$$

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 vector x_{it} contains P time-varying, exogenous characteristics of individual i; θ_i is the pure person effect; $\psi_{\mathrm{J}(i,t)}$ is the pure firm effect for the firm at which worker i is employed at date t (denoted by $\mathrm{J}(i,t)$); μ_y is the grand mean of y_{it}; μ_x is the grand mean of x_{it}; 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

$$\begin{aligned}
y_i &= \begin{bmatrix} y_{i,n_{i1}} - \mu_y \\ \cdots \\ y_{i,n_{iT_i}} - \mu_y \end{bmatrix}, \\
X_i &= \begin{bmatrix} x_{i,n_{i1},1} - \mu_{x1} \cdots x_{i,n_{i1},P} - \mu_{xP} \\ \cdots \\ x_{i,n_{iT_i},1} - \mu_{x1} \cdots x_{i,n_{iT_i},P} - \mu_{xP} \end{bmatrix}, \\
\varepsilon_i &= \begin{bmatrix} \varepsilon_{i,n_{i1}} \\ \cdots \\ \varepsilon_{i,n_{iT_i}} \end{bmatrix}
\end{aligned}$$

where y_i and ε_i are $T_i \times 1$ and X_i is $T_i \times P$.

We will assume that a simple random sample of N individuals is observed for a maximum of T periods.[3] Thus, ε_i has the following properties:

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

and

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

[3] Many linked data sets began as simple random samples of administrative files. See Abowd and Kramarz (1999) for a typology.

where the set $\{\mathrm{J}(i,.)\}$ means $\{\mathrm{J}(i,n_{i1}),\ldots,\mathrm{J}(i,n_{iT_i})\}$, the set of all employers of individual i and $\{\Sigma_{T_i}\}_i$ 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 $\{n_{i1}, n_{i2}, \ldots, n_{iT_i}\}$. In full matrix notation we have

$$y = X\beta + D\theta + F\psi + \varepsilon \tag{2}$$

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 indicators variables for the individual; F is the $N^* \times J$ design matrix of firm indicators variables for the firm effects 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 the vector y is ordered according to individuals as

$$y = \begin{bmatrix} y_1 \\ \cdots \\ y_N \end{bmatrix} \tag{3}$$

and X, D, F 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 (1) and (2) 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. Identification of the effects is accomplished by the imposition of a zero sample mean for θ_i and $\psi_{\mathrm{J}(i,t)}$ taken over all (i,t) for fixed-effects techniques and by the assumption of zero mean and finite variance for all random-effects techniques.

3. Definition of Effects of Interest

Many familiar models are special cases of the linear model in equations (1) and (2). In this section we define a variety of effects using 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 illustrate many of the issues we raise using the example of the estimation of inter-industry differentials, called industry effects, on the basis of the specification in equation (2).

3.1 Person Effects and Unobservable Personal Heterogeneity

The person effect in equation (1) combines the effects of observable time-invariant personal characteristics and unobserved personal heterogeneity. We decompose these two parts of the person effect as

$$\theta_i = \alpha_i + u_i\eta \tag{4}$$

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 decompostion in equation (4) 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). The design matrix D in equation (2) 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 (4).

3.2 Firm Effects and Unobservable Firm Heterogeneity

The firm effect in equation (1) combines the effects of observable 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. We illustrate each of those possibilities in this subsection. The definition of the firm effect ψ_j used in equations (1) and (2) is time-invariant. We decompose this effect into observable and unobservable components as

$$\psi_j = \phi_j + q_j \rho \tag{5}$$

where ϕ_j is the unobservable firm heterogeneity, q_j is a vector of time-invariant firm characteristics, and ρ is a vector of effects associated with the time-invariant firm characteristics. Time-varying firm characteristics and time-varying employer-employee match characteristics require a re-definition of the simple firm effect as ψ_{jit}, where the addition of the i and t subscripts allows for the possibility of time-varying firm effects and employer-employee match effects. Now let the firm characteristics be time-varying, q_{jt}, and consider the match characteristics s_{jit}, then the firm effect can be expressed as

$$\psi_{jit} = \phi_j + q_{jt} \rho + s_{jit} \gamma_j \tag{6}$$

where γ_j is a vector of parameters associated with the match characteristics. Statistical analysis of the effects defined by equation (6) is accomplished by augmenting the columns of F to reflect the data in q_{jt} and s_{jit}. The formulas shown in the estimation sections below can then be applied to the augmented design matrix. We will not show the general formulas necessary to perform this estimation. The reader should see AKM for the fixed-effects methods. Generalization to the mixed-effects models is based on Searle, Casella and McCulloch (1992).

3.3 Firm-Average Person Effect

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

$$\bar{\theta}_j \equiv \bar{\alpha}_j + \bar{u}_j \eta = \frac{\sum_{\{(i,t)|\, \mathrm{J}(i,t)=j\}} \theta_i}{N_j} \tag{7}$$

where the function $1(A)$ takes the value 1 if A is true and 0 otherwise and

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

The importance of the effect defined in equation (7) 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, employment duration weighted) 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.

3.4 Person-Average Firm Effect

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

$$\bar{\psi}_i \equiv \bar{\phi}_i + \bar{q}_i\rho + \bar{s}_i\bar{\gamma}_i = \frac{\sum_t \psi_{\mathrm{J}(i,t)it}}{T_i}. \tag{8}$$

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, potentially measured by the effect defined in equation (7), and the effect of heterogeneous firms on an indivdual's career employment outcomes, potentially measured by the effect in equation (8).

3.5 Industry Effects[4]

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 (1) and (2), 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 (2) 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 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 (1), then the

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

equation becomes

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

or, in matrix notation as in equation (2),

$$y = X\beta + D\theta + FA\kappa + (F\psi - FA\kappa) + \varepsilon \tag{9}$$

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 (9) reveals that the vector κ, $K \times 1$, may be interpreted as the following weighted average of the pure firm effects:

$$\kappa \equiv (A'F'FA)^{-1}A'F'F\psi. \tag{10}$$

and the effect $(F\psi - FA\kappa)$ may be re-expressed as $M_{FA}F\psi$, where the column null space of an arbitrary matrix, Z, is denoted $M_Z \equiv I - Z\left(Z'Z\right)^{-}Z$, and $()^{-}$ is a computable g-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 (9). Only $\mathrm{rank}(F'M_{FA}F)$ 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 (9), using either of the class of methods discussed below. Equation (9) 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 decomoposition is orthogonal, the presence of X and D in equation (9) greatly complicates the estimation using either fixed-effects or mixed-effects techniques.

3.6 Other Firm Characteristic Effects

Through careful specification of the firm effect in equation (6), we can estimate the average effect associated with any firm characteristic, q_{jt}, or any interaction of firm and personal characteristics, s_{jit}, while allowing for unobservable firm and personal heterogeneity.

3.7 Occupation Effects and Other Person $\times$ 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. Occupation effects are often interpreted as an interaction between person and firm effects (Groshen 1991a 1991b, 1996, implicitly). It is possible to expand the model in equation (2) to include the interaction of person and firm effects. Such a formulation might be used to compute estimated occupation effects or to study more general job matching models. The feasible estimation techniques appropriate for such models are discussed in the mixed-model estimation section below.

4. Estimation by Fixed-effects Methods

4.1 Conditional methods

The normal equations for exact fixed-effects estimation of equation (2) by least squares are

$$\begin{bmatrix} X'X & X'D & X'F \\ D'X & D'D & D'F \\ F'X & F'D & F'F \end{bmatrix} \begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} X'y \\ D'y \\ F'y \end{bmatrix} \tag{11}$$

Estimation by fixed-effects methods is complicated by the high dimensionality of equation (11). AKM showed that many common methods of approximating the solution can be derived by considering an augmented version of the equation in which ancillary effects, λ, defined in conjunction with interactions of observable characteristics of persons and firms denoted by Z are inserted into the model as

$$y = X\beta + D\theta + F_1\psi_1 + Z\lambda + (F_2\psi_2 - Z\lambda) + \varepsilon$$

where the matrix F has been partitioned into included, F_1, and excluded, F_2, parts and the effect ψ has been conformably partitioned. The ancillary parameter is defined as

$$\lambda \equiv (Z'Z)^{-1}Z'F_2\psi_2$$

Conditional, fixed-effects, estimation consists of solving the normal equations:

$$\begin{bmatrix} X'X & X'D_1 & X'F_1 & X'Z & X'M_ZF_2 \\ D_1'X & D_1'D_1 & D_1'F_1 & D_1'Z & D_1'M_ZF_2 \\ F_1'X & F_1'D_1 & F_1'F_1 & F_1'Z & F_1'M_ZF_2 \\ Z'X & Z'D_1 & Z'F_1 & Z'Z & Z'M_ZF_2 \\ F_2'M_ZX & F_2'M_ZD_1 & F_2'M_ZF_1 & F_2'M_ZZ & F_2'M_ZF_2 \end{bmatrix} \begin{bmatrix} \beta \\ \theta \\ \psi_1 \\ \lambda \\ \psi_2 \end{bmatrix} = \begin{bmatrix} X'y \\ D'y \\ F_1'y \\ Z'y \\ F_2'M_Zy \end{bmatrix} \tag{12}$$

in two steps imposing the design orthogonality hypotheses:

$$\begin{bmatrix} X'M_ZF_2 \\ D_1'M_ZF_2 \\ F_1'M_ZF_2 \end{bmatrix} = 0 \tag{13}$$

The maintained hypotheses in equation (13) impose orthogonality between $\begin{bmatrix} X & D & F_1 \end{bmatrix}$ and F_2 given Z. The effect of the maintained hypotheses is to reduce the computations to a sequence of computable solutions to the normal equations. There are several potential solutions to the conditional normal equations under the maintained orthogonality hypotheses. AKM examine these solutions empirically and show that the solution

$$\begin{bmatrix} \widehat{\beta} \\ \widehat{\psi}_1 \\ \widehat{\lambda} \end{bmatrix} = \begin{bmatrix} X'M_DX & X'M_DF_1 & X'M_DZ \\ F_1'M_DX & F_1'M_DF_1 & F_1'M_DZ \\ Z'M_DX & Z'M_DF_1 & Z'M_DZ \end{bmatrix}^{-} \begin{bmatrix} X'M_Dy \\ F_1'M_Dy \\ Z'M_Dy \end{bmatrix} \tag{14}$$

$$\widehat{\theta} = (D'D)^{-1}D'\left(y - X\widehat{\beta} - F_1\widehat{\psi}_1 - Z\widehat{\lambda}\right) \tag{15}$$

and

$$\widehat{\psi}_2 = (F_2'F_2)^{-}\left(y - X\widehat{\beta} - F_1\widehat{\psi}_1 - D\widehat{\theta}\right) \tag{16}$$

has the best properties among the computable solutions for the French data they were using.

In words, the solution in equations (14) to (16) can be described as follows. Compute

the within-persons solution for β, ψ_1, and λ. Next, compute the fixed person effects, θ, from the average deviation of y from its conditional mean for each individual given the coefficients from (14). Next, compute the remaining firm effects, ψ_2 on a firm-by-firm basis using the conditional mean within each firm given the coefficients from (14) and (15), excluding $Z\hat{\lambda}$.

AKM also present a variety of specification tests for the maintained hypotheses (13). However, they note that given the very large sample sizes inherent in linked longitudinal employer-employee data these hypotheses are usually rejected.

4.2 Consistent Methods for β and γ

In this subsection we show how to obtain consistent estimates of β and γ_j using the within-individual-firm differences of the data. This method provides us with our most robust statistical method for these effects in the sense that we use no additional statistical assumptions beyond those specified in equation (1) and the definitions (6) and (1).[5] We should note, however, that this estimation technique relies heavily on the assumption of no interaction between X and F. Consider the first differences:

$$y_{i,n_{it}} - y_{in_{it-1}} = (x_{in_{it}} - x_{in_{it-1}})\beta + \gamma_{\mathrm{J}(i,n_{it})}(s_{in_{it}} - s_{in_{it-1}}) + \varepsilon_{in_{it}} - \varepsilon_{in_{it-1}} \quad (17)$$

for all observations for which $\mathrm{J}(i,n_{it}) = \mathrm{J}(i,n_{it-1})$, which we represent in matrix form as:

$$\triangle y = \triangle X\beta + \widetilde{F}\gamma + \triangle\varepsilon \quad (18)$$

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 $\mathrm{J}(i,n_{it}) = \mathrm{J}(i,n_{it-1})$. The matrix $\widetilde{F}$ is the rows of F_1 that correspond to the person-years (i,t) for which the condition $\mathrm{J}(i,n_{it}) = \mathrm{J}(i,n_{it-1})$ is satisfied. Then,

$$\widetilde{\beta} = (\triangle X' M_{\widetilde{F}} \triangle X)^{-1} \triangle X' M_{\widetilde{F}} \triangle y \quad (19)$$

and

$$\widetilde{\gamma} = (\widetilde{F}'\widetilde{F})^{-1}\widetilde{F}'(\triangle y - \triangle X\widetilde{\beta}). \quad (20)$$

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

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

where

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

[5] We have excluded $q_{jt}\rho$ from the firm effect we consider in section.

and

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

It is understood that only the rows of $\triangle\varepsilon$ that satisfy the condition $\mathrm{J}(i,n_{it}) = \mathrm{J}(i,n_{it-1})$ are used in the calculation of $\widetilde{}$, 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 (1).

4.3 Fixed-effects Estimators for θ and ϕ Based on the Consistent Method

To simplify the notation in what follows, we use the time subscripts s and t, which should interpreted as varying over the index set $\{n_{i1},\ldots,n_{iT_i}\}$.

A candidate estimator for ϕ (or ψ, if the firm effect has no covariates) is based upon the consistent estimator defined by equations (19) and (20). Consider the consistent estimator of the difference δ_{it} defined as

$$\widetilde{\delta}_{it} \equiv y_{it} - x_{it}\widetilde{\beta} - s_{it}\tilde{\gamma}_{\mathrm{J}(i,t)} - \frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}} y_{is} - x_{is}\tilde{\beta} - s_{is}\tilde{\gamma}_{\mathrm{J}(i,s)}}{\sum_s 1\left[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\right]}. \tag{21}$$

Then, a consistent estimator of the average δ_j in each firm is

$$\widetilde{\delta}_j \equiv \frac{\sum_{(i,t)\in\{\mathrm{J}(i,t)=j\}} \widetilde{\delta}_{it}}{\sum_{(i,t)} 1\left[\mathrm{J}(i,t)=j\right]}. \tag{22}$$

We next consider the relation between δ_j and ϕ_j (or ψ_j, if the firm effect excludes other covariates). Note that, before the substitution of estimators for β and γ, δ_{it} can be written as a function of θ_i and $\phi_{\mathrm{J}(i,s)}$

$$\begin{aligned} \delta_{it} &= \theta_i + \phi_{\mathrm{J}(i,t)} + \varepsilon_{it} - \frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}} \theta_i + \phi_{\mathrm{J}(i,s)} + \varepsilon_{is}}{\sum_s 1\left[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\right]} \\ &= \left(\varepsilon_{it} - \frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}} \varepsilon_{is}}{\sum_s 1\left[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\right]}\right) + \left(\phi_{\mathrm{J}(i,t)} - \frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}} \phi_{\mathrm{J}(i,s)}}{\sum_s 1\left[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\right]}\right) \end{aligned}$$

where we have made the substitution of equation (1) for y_{it}. Under the linear model assumptions stated below equation (2), we have

$$\mathrm{E}\left[\delta_{it}\,|X,D,F\right] = \mathrm{E}\left[\left(\varepsilon_{it} - \frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}} \varepsilon_{is}}{\sum_s 1\left[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\right]}\right)\middle| X,D,F\right] \tag{23}$$

$$+\mathrm{E}\left[\phi_{\mathrm{J}(i,t)}-\frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}}\phi_{\mathrm{J}(i,s)}}{\sum_{s}1\left[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\right]}\middle|X,D,F\right]$$

$$=\mathrm{E}\left[\phi_{\mathrm{J}(i,t)}-\frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}}\phi_{\mathrm{J}(i,s)}}{\sum_{s}1\left[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\right]}\middle|X,D,F\right]$$

and

$$\mathrm{E}\left[\delta_j\left|X,D,F\right.\right]=\mathrm{E}\left[\phi_j-\frac{\sum_{(i,t)\in\{\mathrm{J}(i,t)=j\}}\left(\frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}}\phi_{\mathrm{J}(i,s)}}{\sum_{s}1[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)]}\right)}{\sum_{(i,t)}1\left[\mathrm{J}(i,t)=j\right]}\middle|X,D,F\right]. \quad (24)$$

As equations (23) and (24) show, the classical fixed-effects assumptions are not sufficient to identify ϕ_j (or ψ_j, if the firm effect has no covariates), given estimators for β and γ. However, there is an ancillary assumption that works—namely, the assumption that, in addition to exogenous mobility (ε orthogonal to X, D, and F), there is a form of random mobility denoted by

$$\mathrm{E}\left[\frac{\sum_{s\in\{\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\}}\phi_{\mathrm{J}(i,s)}}{\sum_{s}1\left[\mathrm{J}(i,s)\neq\mathrm{J}(i,t)\right]}\middle|X,D,F\right]=0. \quad (25)$$

It is important to understand the meaning of assumption (25) as well as why it is stronger than the original assumption of exogenous mobility. The random mobility assumption means that every time a person separates from one employer and goes to work at another, the value of ϕ_j (or ψ_j, if the firm effect has no covariates) has a zero expected value. Exogenous mobility means that the value of ϕ_j (or ψ_j, if the firm effect has no covariates) is uncorrelated with ε. Statistically, exogenous mobility means that ε is uncorrelated with F, whereas random mobility means that the columns of $\widetilde{F}$ are orthogonal to D and X, in addition to ε. As we saw when we discuss other conditional methods, imposing extra orthogonality conditions on the data, ones that cannot be tested, is one way out of the computational burden of these models.

5. Estimation by Mixed-effects Methods

5.1 The Mixed-Model Equations

A mixed model formulation occurs when some, or all, of the personal characteristics, person effects and firm effects in equation (2) are taken as random, rather than fixed, effects. There is considerable confusion about the comparison of random and fixed effects specifications in the literature on linked employer-employee data, so we take some pains in this section

to define our terms in a manner that is coherent with the enormous statistical literature on these types of models.

We begin by specifying the structure on all the statistical error as well as the random person and firm effects. Let

$$\mathrm{E}\left[\varepsilon_i\right] = 0$$

and

$$\mathrm{V}\left[\varepsilon_i\right] = \Sigma_i$$

where Σ_i is the appropriate selection of rows and columns from the covariance matrix Σ, a $\max[T_i] \times \max[T_i]$ positive definite covariance matrix. Define

$$\Lambda = \begin{bmatrix} \Sigma_1 & 0 & \cdots & & 0 \\ \cdots & \cdots & & & \cdots \\ 0 & \cdots & \Sigma_i & \cdots & 0 \\ \cdots & & & \cdots & \cdots \\ 0 & & \cdots & 0 & \Sigma_N \end{bmatrix} \tag{26}$$

Let

$$\mathrm{E}\begin{bmatrix} \theta \\ \psi \end{bmatrix} | X = 0$$

and

$$\mathrm{V}\begin{bmatrix} \theta \\ \psi \end{bmatrix} | X = \tag{27}$$

The system of equations known in the statistics literature as the mixed-model equations (Searle, Casella and McCulloch 1992) for this model is

$$\begin{bmatrix} X'\Lambda^{-1}X & X'\Lambda^{-1}\left[D\,|F\right] \\ \begin{bmatrix} D' \\ F' \end{bmatrix}\Lambda^{-1}X & \begin{bmatrix} D' \\ F' \end{bmatrix}\Lambda^{-1}\left[D\,|F\right] + \quad {}^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} X'\Lambda^{-1}y \\ \begin{bmatrix} D' \\ F' \end{bmatrix}\Lambda^{-1}y \end{bmatrix} \tag{28}$$

Assuming that the vectors ε and $\begin{bmatrix} \theta' & \psi' \end{bmatrix}$ follow independent multivariate normal distributions with mean zero and the covariance matrices given by equations (26) and (27), the solution to equation (28) for $\hat{\theta}$ and $\hat{\psi}$, using maximum likelihood estimates for Λ and , is known as the estimated realized person and firm effects (Robinson, 1991). These estimated realized effects can be used as a basis for firm-level analyses using formulas that are comparable to those based upon the fixed-effects estimates.

5.2 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 (28), with nondiagonal, when they refer to a correlated random-effects model.

There are some important links between the two model specifications. First, we show that the usual fixed person and firm effects estimator for linked longitudinal employer-

employee data is a special case of the mixed model equation (28). When

$$\Lambda = \sigma_{\varepsilon}^2 I_{N^*}$$

and

$$| \quad | \rightarrow \infty$$

equation (28) approaches equation (11); hence the fixed-effects estimator of $\begin{bmatrix} \beta' & \theta' & \psi' \end{bmatrix}$ that solves (7) is a special case of the random effects estimator that solves (28).

It is important to distinguish between correlated random-effects models based on the mixed-effects equation (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 4 and the difficulty associated with solving the mixed-model equations, in general (see Robinson, 1991, Searle, et al. 1992, Groeneveld and Neumaier, 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:

$$x_{it} = \upsilon_i + \varsigma_{it} \tag{29}$$

where

$$\begin{aligned} \mathrm{Corr}[\upsilon_i, \theta_i] &\neq 0 \\ \mathrm{V}[\varsigma_{it}] &= \Delta \end{aligned}$$

and

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

This specification implies that $\mathrm{Corr}[\upsilon_i, \psi_{\mathrm{J}(i,t)}] \neq 0$ as long as 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

$$y_i = x_i \Pi + \nu_i \tag{30}$$

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 (30) require modification. The most direct way to accomplish the extension of Chamberlain's methods is to substitute equation (29) into equation (1), then restate the system of equations as a mixed model. For

each individual i in period t we have

$$\begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix} = \begin{bmatrix} \tau_i + \psi_{\mathrm{J}(i,t)} + \xi_{it} \\ \upsilon_i + \varsigma_{it} \end{bmatrix}. \tag{31}$$

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 (31) can be written as

$$m = D_1\tau + D_2\upsilon + F_3\psi + \nu \tag{32}$$

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 $\tau, \upsilon,$ 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 (32) has been developed by Groeneveld (1998) and some applications are discussed in Robinson (1991) and Tanner (1996). The methods exploit the sparce stucture of $D_1, D_2,$ and F_3 and use analytic derivatives to solve (28). Robert and Casella (1998) and Tanner (1996) provide algorithms based on simulated data techniques.

6. Heterogeneity biases in incomplete models

The analyses in this section are based upon the exact fixed-effects estimator for model (2) given by the solution to (11).

6.1 Omission of the firm effects

When the estimated version of equation (2) 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:

$$\theta^* = \theta + (D'M_XD)^{-1}D'M_XF\psi. \tag{33}$$

Hence, if X were orthogonal to D and F, so that $D'M_XD = D'D$ and $D'M_XF = D'F$, 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:

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

6.2 Omission of the person effects

Omitting the pure person effects (θ) from the estimated version of equation (2) 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:

$$\psi^{**} = \psi + (F'M_X F)^{-1} F'M_X D\theta. \tag{34}$$

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

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

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 (2), β, and an omitted variable bias that depends upon the covariance of X and D, given F:

$$\beta^{**} = \beta + (X'M_F X)^{-1} X'M_F D\theta. \tag{35}$$

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

6.3 Inter-industry wage differentials

We showed above that industry effects are an example of an effect that aggregates firm effects and 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 (9). Only $\mathrm{rank}(F'M_{FA}F)$ fixed firm effects can be separately identified; however, the mixed-effects model can produce estimates of all realized industry and firm effects. Using the fixed-effects estimator,

there is neither an omitted variable nor an aggregation bias in the estimates based upon equation (2).

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:

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

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 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 (2), 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:

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

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 heterogeniety biases in the estimation of industry effects, AKM propose a decompostion of the raw industry effect into the part due to individual heterogeneity and the part due to firm heterogeneity. Their formulas apply to directly to the fixed-effects estimator of equation (2) and can be extended to the estimated realized effects in a mixed-effects model. When equation (9) 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:

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

which can be restated as

$$\kappa^{**} = (A'F'M_XFA)^{-1}A'F'M_XF\psi + (A'F'M_XFA)^{-1}A'F'M_XD\theta, \qquad (36)$$

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 $A'F'M_XFA = A'F'FA$, $A'F'M_XF = A'F'F$, and $A'F'M_XD = A'F'D$, 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

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

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.

7. Data Analysis at the Firm Level

Linked employer-employee data are also useful for the study of outcomes at the firm level. Such investigations usually require the calculation of firm-level summary statistics based upon either fixed-effects or mixed-effects estimates of equation (2). We show here that firm-average person effects and similar moment-based statistics, estimated by either fixed-effects or mixed-effects techniques, can be used as a part of analyses of linear models at the firm level.

Consider the firm-average person effect defined in equation (7). The asymptotic distribution for $\left[\begin{array}{cc} \hat{\theta}_j & \hat{\psi}_j \end{array}\right]$, for either fixed-effects or mixed-effects estimation is

$$\left[\begin{array}{c} \hat{\theta}_j \\ \hat{\psi}_j \end{array}\right] \rightarrow \mathrm{N}\left(\left[\begin{array}{c} \bar{\theta}_j \\ \bar{\psi}_j \end{array}\right], \Sigma_j^*\right), \text{ as } N_j \rightarrow \infty \tag{37}$$

where $\left|\Sigma_j^*\right| \rightarrow 0$ as $N_j \rightarrow \infty$, holding constant the distribution of firm sizes. Thus as $N, N_j \rightarrow \infty$,we assume that their ratio N_j/N goes to a non-zero constant. In particular, $\sigma^2_{\bar{\theta}_j} \rightarrow 0$ as $N_j \rightarrow \infty$, so that firm-average statistics based on either fixed-effects or mixed-effects estimation of equation (2) have valid asymptotic sampling distributions that can be used as the basis for the firm-level analysis even though the number of time periods that an individual is observed does not grow.

Next, consider next the statistical relation between firm-level outcomes, firm-average person effects and firm effects. The basic model is

$$f_j = \left[\begin{array}{ccc} \bar{\theta}_j & \psi_j & q_j \end{array}\right]\rho + \nu_j \tag{38}$$

where f_j is any firm-level outcome, $\left[\begin{array}{cc} \bar{\theta}_j & \psi_j \end{array}\right]$ is a vector of firm-level measures based on equation (2), ρ is a vector of parameters, and ν_j is a zero-mean statistical error. The firm-level variables based on estimation of equation (2) are derived from the basic linked employer-employee sample. Thus, they are are estimated regressors. Consequently, we

must allow for the estimation errors in $\widehat{\theta}_j$,and $\widehat{\psi}_j$. Thus, equation (38) becomes

$$\begin{array}{rl} f_j = & \left[\begin{array}{ccc} \widehat{\theta}_j & \widehat{\psi}_j & q_j \end{array} \right] \rho + \\ & \left(\left[\begin{array}{ccc} \bar{\theta}_j & \psi_j & q_j \end{array} \right] - \left[\begin{array}{ccc} \widehat{\theta}_j & \widehat{\psi}_j & q_j \end{array} \right] \right) \rho + \nu_j \end{array} \tag{39}$$

where $\left(\left[\begin{array}{ccc} \bar{\theta}_j & \psi_j & q_j \end{array} \right] - \left[\begin{array}{ccc} \widehat{\theta}_j & \widehat{\psi}_j & q_j \end{array} \right] \right) \rho$ is the error associated with the first-step estimation of the firm-level measures.[6]

In order to derive the error covariance matrix for equation (39), let

$$F_j' \left(\widehat{\delta}_j \right) \equiv \left[\begin{array}{ccc} \widehat{\theta}_j & \widehat{\psi}_j & q_j \end{array} \right]$$

Where F_j is not related to the F matrices defined for equation (2), δ_j is not related to the effects defined in equation (24) and

$$\widehat{\delta}_j' \equiv \left[\begin{array}{cc} \widehat{\theta}_j & \widehat{\psi}_j \end{array} \right].$$

Now, equation (39) can be re-expressed in a first order approximation around δ_j as:

$$f_j = F_j' \left(\delta_j \right) \rho + \omega_j \tag{40}$$

where

$$\omega_j \equiv (\widehat{\delta}_j - \delta_j)' \frac{\partial F_j' \left(\delta_j \right)}{\partial \delta_j} \rho + \nu_j$$

The variance of the regression error term for equation (40) consists of the component due to the estimation error in F_j plus the component due to ξ_j:

$$\mathrm{Var} \left[\omega_j \right] = \rho' \frac{\partial F_j}{\partial \delta_j'} \mathrm{Var} \left[\widehat{\delta}_j \right] \frac{\partial F_j'}{\partial \delta_j} \rho + \mathrm{Var} \left[\nu_j \right] \tag{41}$$

where the components of $\mathrm{Var} \left[\widehat{\delta}_j \right]$ are defined based on functions of the estimated effects from equation (2), as shown in equation (37), whether estimated by fixed-effects or mixed-effects methods. One can estimate equation (40) using generalized least squares based upon the error variance in equation (41). See AKM for details.

The equation defining the firm-average person effect, equation (7), as well as the definition of the firm effect itself, ψ, can be modified to permit time-varying firm-average person effects and firm effects. Characteristics of the firm that are measured at multiple points in time can be incorporated into equation (38). None of these extensions affects the general structure of equations (40) and (41). Computational problems, on the other hand, remain severe for time-varying effects, even considering the wealth of new techniques summarized in Tanner (1996).

[6] We adopt the model of Pagan (1984); namely, that the regression of interest relates functions of the individual-level data and estimated firm-level effects to the other measured firm-level outcomes. We account for the estimation error $\left[\begin{array}{ccc} \bar{\theta}_j & \psi_j & q_j \end{array} \right] - \left[\begin{array}{ccc} \widehat{\theta}_j & \widehat{\psi}_j & q_j \end{array} \right]$ explicitly, but we do not add an additional measurement error. Thus, for example, we assert that the outcome f_j depends upon θ_j and not upon $\theta_j + \zeta_j$, where ζ_j is an independent measurement error.

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 components of the person and firm effects. A complete treatment of this problem is beyond the scope of this paper; however, it is worth noting that the interpretation of equations (1) and (2) 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) 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 5; however, to our knowledge the application of such techniques to this type of endogeneous mobility model has not been attempted using linked employer-employee data.

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 or the use of conditional, simpler to compute, solutions.

We do not mean to suggest that these estimation strategies are complete. Indeed, many of the techniques suggested in this paper 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. and Francis Kramarz "The Analysis of Labor Markets Using Matched Employer-Employee Data," in O. Ashenfelter and D. Card, eds. *Handbook of Labor Economics*, 3 (Amsterdam, North Holland, 1999), forthcoming.

[2] Abowd, John M, Francis Kramarz and David N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica* (January 1999): forthcoming.

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

[4] Cramér, H. *Mathematical Models of Statistics* (Princeton, NJ: Princeton University Press, 1946).
[5] Dickens, William T. and Lawrence F. Katz. (1987): "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.
[6] Gibbons, Robert and Lawrence Katz (1992): "Does Unmeasured Ability Explain Inter-Industry Wage Differentials?" *Review of Economic Studies*, 59, 515-535.
[7] Groeneveld, Eildert *VCE4 User's Guide and Reference Manual* (Höltystrass, Germany: Institute of Animal Husbandry and Animal Behavior, 1998).
[8] Groshen, Erica "Sources of Intra-Industry Wage Dispersion: How Much do Employers Matter?" *Quarterly Journal of Economics*, 106, (1991a): 869-884.
[9] Groshen, Erica (1991b): "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, 457-72.
[10] Groshen, Erica (1996): "American Employer Salary Surveys and Labor Economics Research: Issues and Contributions," *Annales d'économie et de statistique*, 41/42, 413-442.
[11] Henderson, C.R. "Estimation of Variance and Covariance Components," *Biometrics* 9 (1953): 226-252.
[12] Jovanovic, Bojan (1979): "Job Matching and the Theory of Turnover," *Journal of Political Economy*, 87, 972-990.
[13] Krueger, Alan B. and Lawrence H. Summers (1987): "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.
[14] Krueger, Alan B. and Lawrence H. Summers (1988): "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica*, 56, 259-293.
[15] Lane, Julia, Simon Burgess and Jules Theeuwes (1997): "The Uses of Longitudinal Matched Worker/Employer Data in Labor Market Analysis," *American Statistical Association Papers and Proceedings*.
[16] Mundlak, Yair (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46.
[17] Murphy, Kevin M. and Robert H. Topel (1987): "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.
[18] Neumaier, Arnold and Eildert Groeneveld, "Restricted Maximum Likelihood Estimation of Covariance in Sparse Linear Models," working paper, Institut für Mathematik, Wien University, Austria, 1996.
[19] Pagan, Adrian (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 21-47.
[20] Robert, Christian P. and George Casella, *Monte Carlo Statistical Methods*, CREST working draft, 1998.
[21] Robinson, G.K. "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6, (1991): 15-51.
[22] Searle, Shayle R., George Casella and Charles E. McCulloch *Variance Components* (New York: John Wiley and sons, 1992).

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