Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data

John M. Abowd, Cornell University, United States Census Bureau, CREST, and NBER

Robert H. Creecy, United States Census Bureau

and

Francis Kramarz, CREST-INSEE, CEPR, IZA

March 2002

The authors gratefully acknowledge the research assistance of Paul Lengermann. Charles E. McCulloch provided valuable assistance in relating our research to the statistical breeding literature. The authors also acknowledge funding from the NSF (SES 9978093 to Cornell University). The data used in this paper are confidential but the authors' access is not exclusive. For further information contact the Cornell University Restricted Access Data Center, CISER, 201 Caldwell Hall, Ithaca, NY 14853. No United States Census Bureau confidential data were used in this paper. The authors' conclusions do not represent the official policy of the any government agency.

Computing Person and Firm Effects
Using Linked Longitudinal Employer-Employee Data

Abstract

In this paper we provide the exact formulas for the direct least squares estimation of statistical models that include both person and firm effects. We also provide an algorithm for determining the estimable functions of the person and firm effects (the identifiable effects). The computational techniques are also directly applicable to any linear two-factor analysis of covariance with two high-dimension non-orthogonal factors. We show that the application of the exact solution does not change the substantive conclusions about the relative importance of person and firm effects in the explanation of log real compensation; however, the correlation between person and firm effects is negative, not weakly positive, in the exact solution. We also provide guidance for using the methods developed in earlier work to obtain an accurate approximation.

John M. Abowd
259 Ives Hall
Cornell University
Ithaca, NY 14853-3901
United States of America

mailto: John_Abowd@cornell.edu

Robert H. Creecy
Statistical Research Division
United States Census Bureau
FOB 4-3207
Suitland, MD 20746
United States of America

mailto: Robert.H.Creecy@census.gov

Francis Kramarz
CREST-INSEE
15, bd Gabriel Péri
92245 Malakoff
France

mailto: kramarz@ensae.fr

1. Introduction

Two related articles Abowd, Kramarz and Margolis (AKM, 1999) and Abowd, Finer and Kramarz (AFK, 1999) provided a basic statistical framework for decomposing wage rates into components due to individual heterogeneity (measured and unmeasured) and firm heterogeneity (measured and unmeasured). The first of these articles, AKM, analyzes French data. The second of these articles, AFK, analyzes data from the State of Washington.

Both AKM and AFK used statistical approximations to estimate the decomposition of wage differentials into individual and employer components. In this article we show new methods that provide the exact solution to the estimation problem. We analyze the same French data as AKM and the same American data as AFK. The exact results fully confirm the approximate results for the State of Washington but slightly change the explanation for wage differentials for France. The reason for the difference in the French results is that the computations for the approximation in AKM were limited by the capacity of the computers on which they were generated. The approximation was not sufficiently accurate. The same approximation, using more terms in the conditioning set, worked well for the analysis of the State of Washington.

Section 2 summarizes the basic statistical model. Section 3 provides the details for identification and estimation by fixed-effect methods. Section 4 presents the data analysis comparing the original approximate results with the exact results. Section 5 concludes.

2. Basic Statistical Model

The dependent variable is the natural logarithm of the rate of compensation per unit of time, y_{it}, observed for individual *i* at date *t*, expressed as a function of individual heterogeneity, firm heterogeneity, and measured time-varying characteristics:

$$y_{it} = \theta_i + \psi_{\mathrm{J}(i,t)} + x_{it}\beta + \varepsilon_{it} . \qquad (1)$$

where $i = 1,\dots,N$, $t \in \{n_{i1},\dots,n_{iT_i}\}$, and there is no intercept included in x_{it}.[1] The function J(*i,t*) indicates the employer *j* of *i* at date *t* where $j = \mathrm{J}(i,t)$ and $j = 1,\dots,J$. There are T_i observations per individual and $N^* = \sum_i T_i$ total observations. The first component of equation (1) is the individual effect, θ_i. The second component is the firm effect, $\psi_{\mathrm{J}(i,t)}$. The third component is the effect of measured time-varying characteristics, $x_{it}\beta$. The fourth component is the statistical residual, ε_{it}, with the assumptions that $\mathrm{E}[\varepsilon_{it}|i,t,x] = 0$, $\mathrm{Var}[\varepsilon_{it}|i,t,x] < \infty$, and orthogonal to all

[1] The dating is complicated by the fact that individuals may have incomplete work histories. The time subscript ranges over all available dates for the individual with each n_{is} being between the start date and the end date for the sample. The time subscripts are arranged in increasing order. See AKM (1999) for additional discussion of the notation.

other effects in the model.[2] In the statistics literature equation (1) is known as a two-factor analysis of covariance with two high-dimensional factors and an unbalanced (non-orthogonal) design (see Searle *et al.*, 1992, chapter 5). Although we focus on the application of these methods to models developed for linked longitudinal employer-employee data, we note that models for doctor-patient and school-student outcomes have the same statistical structure. Our methods would also be directly useful for these applications.

In order to state the basic statistical relations more clearly we restate equation (1) in matrix format. All vectors/matrices have row dimensionality equal to the total number of observations, N^*. The data are sorted by *i* and ordered chronologically for each person. This gives the following equation for the stacked system:

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

where D $\left(N^* \times N\right)$ is the design matrix for the person effect: columns equal to the number of unique person IDs; F $\left(N^* \times J\right)$ is the design matrix for the firm effect: columns equal to the number of unique firm IDs; and X $\left(N^* \times K\right)$ is the stacked matrix of time-varying characteristics: columns equal to the number of regressors in *X*.

3. Identification and Estimation by Fixed-effect Methods

The normal equations for least squares estimation of fixed person, firm, and characteristic effects are of very high dimension. Estimation of the full model by fixed-effect methods requires special algorithms to deal with the high dimensionality of the problem. After completing work on AFK and AKM, which use statistical approximations, we developed new algorithms that permit the exact least squares estimation of all the effects in equation (2). These algorithms, which are based on the iterative conjugate gradient method, deal with the high dimensionality of the data by using sparse matrices. Our methods have some similarity to those used in the animal and plant breeding literature.[3] Because of the way these algorithms work, conventional methods for assuring that the effects are identified (estimable) do not work. Thus, we also developed appropriate, new, methods for computing the estimable functions of interest based on equation (3) below.

Least Squares Normal Equations

The full least squares solution to the estimation problem for equation (2) solves the following normal equations for all estimable effects:

[2] See Abowd and Kramarz (1999a and 1999b) for a more complete discussion of the exogeneity assumption for the residual.

[3] See Abowd and Kramarz (1999a) for a longer discussion 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 effect that is equivalent to our employer effect is parameterized.

$$\begin{bmatrix} X'X & X'D & X'F \\ D'X & D'D & D'F \\ F'X & F'D & F'F \end{bmatrix} \begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} X'y \\ D'y \\ F'y \end{bmatrix} \tag{3}$$

In both of our estimation samples, the cross-product matrix on the left-hand side of equation (3) is too high-dimensional to use 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 (3)). AKM present a set of approximate solutions to (3) based on the use of different conditioning effects, *Z*. AFK applies the best of these approximations with a much higher-dimension *Z*.

Identification of Individual and Firm Effects

Many interesting economic applications of equation (2) make use of the estimated person and firm effects. Estimation requires a method for determining the identified effects.[4] The usual technique of sweeping out singular row/column combinations from the normal equations (3) is not applicable to the method described in this paper because equation (3) is solved without the computation of a generalized inverse. Hence, identification of the person and firm effects for estimation by direct least squares requires finding the conditions under which equation (3) 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 methods from graph theory 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 (see equation (4) below) 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.

[4] Standard statistical references, for example Searle *et al.* (1992), provide general methods for finding the estimable functions of the parameters of equation (3). These methods also require the solution of a very high dimension linear system and are, therefore, impractical for our purposes.

```
For $g$ = 1, ..., repeat until no firms remain:[5]
    The first firm not assigned to a group is in group $g$.
    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. The number of individuals in each group is N_g. The number of employers in each group is 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. 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. Figure 1 illustrates the graph that identifies person and firm effects for a simple example.

Firm	Person	Group
1	1	1
1	2	1
2	1	1
2	3	1
3	3	1
3	4	1
4	5	2
5	5	2

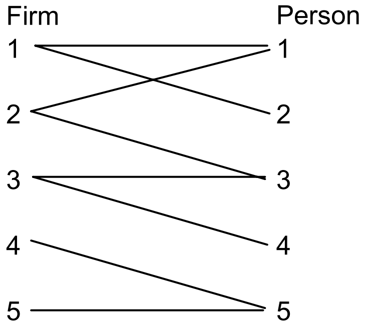

Figure 1. The graph on the right corresponds to the table of firm-person pairs on the left. The grouping algorithm finds the two connected sub-graphs shown.

Within each group g, the group mean of y and $N_g - 1 + J_g - 1$ person and firm effects are identified, where the number of individuals in each group is N_g and the number of employers in each group is 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. After the construction of the G groups, exactly $N + J - G$ effects are estimable. See the proof in Appendix 1.[6]

[5] The implemented algorithm is equivalent to the one described here but is somewhat more complicated because of bookkeeping that tracks which firms and persons have already been added to groups in order to make it more efficient.

[6] The grouping algorithm that we use identifies the "main effect" contrasts due to persons and firms in our model within each group. In the linear models literature our "groups" are called "connected data." See Searle (1987), chapter 5, section 3, pp. 139-145 for a discussion of connected data. See Weeks and Williams (1964) for the general algorithm in analysis of variance models.

Normal Equations after Group Blocking

Our identification argument can be clarified by considering the normal equations after reordering the person and firm effects so that those associated with each group are placed in the design matrix in ascending order. For simplicity, let the arbitrary equation determining the unidentified effect simply set that effect equal to zero, *i.e,* set one person or firm effect equal to zero in each group. Thus, the column associated with this effect can be removed from the reorganized design matrix and the column associated with the group mean is suppressed (recall that there is no constant in *X*). The resulting normal equations are:

$$\left[\begin{array}{c|cc|cc|c|cc}
X'X & X'D_1 & X'F_1 & X'D_2 & X'F_2 & \cdots & X'D_G & X'F_G \\
\hline
D_1'X & D_1'D_1 & D_1'F_1 & 0 & 0 & \cdots & 0 & 0 \\
F_1'X & F_1'D_1 & F_1'F_1 & 0 & 0 & \cdots & 0 & 0 \\
\hline
D_2'X & 0 & 0 & D_2'D_2 & D_2'F_2 & \cdots & 0 & 0 \\
F_2'X & 0 & 0 & F_2'D_2 & F_2'F_2 & \cdots & 0 & 0 \\
\hline
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\hline
D_G'X & 0 & 0 & 0 & 0 & \cdots & D_G'D_G & D_G'F_G \\
F_G'X & 0 & 0 & 0 & 0 & \cdots & F_G'D_G & F_G'F_G
\end{array}\right]
\left[\begin{array}{c}
\beta \\ \hline \theta_1 \\ \psi_1 \\ \hline \theta_2 \\ \psi_2 \\ \hline \cdots \\ \hline \theta_G \\ \psi_G
\end{array}\right]
=
\left[\begin{array}{c}
X'y \\ \hline D_1'y \\ F_1'y \\ \hline D_2'y \\ F_2'y \\ \hline \cdots \\ \hline D_G'y \\ F_G'y
\end{array}\right] \tag{4}$$

The normal equations have a sub-matrix with block diagonal components. This matrix is of full rank and the solution for the parameter vector is unique. We do not solve equation (4) directly. Rather, we apply the technique discussed below to estimate the identifiable effects.

Characteristics of the Groups

Table 1 shows the results of applying our grouping algorithm to the French and Washington State data. Notice that the largest group in both data sets contains the overwhelming majority of all the identifiable person and firm effects. We could apply our estimation methods directly to group 1 alone without much change in the statistical results. We cannot, however, use conventional methods to estimate the person and firm effects group by group because the cross-product matrix for group 1 is essentially the same size as the full set of normal equations (3) and because we are also interested in the β coefficients.

		Second	*Average of*	
	Largest	*largest*	*all other*	*Total of all*
	group	*group*	*groups*	*groups*
French Data				
Observations	4,682,420	51	4.4	5,305,108
Persons	974,985	31	1.4	1,166,305
Firms	334,637	1	1.3	521,180
Groups	1	1	141,550	141,552
Estimable effects	1,309,621	31		1,545,933
State of Washington Data				
Observations	3,999,598	276	15.0	4,036,171
Persons	292,945	33	1.6	296,801
Firms	81,107	3	2.0	85,864
Groups	1	1	2,426	2,428
Estimable effects	374,051	35		380,237

Notes: The ranking of the largest and second largest groups is based on the number of persons in the group. Sources: Authors' calculations based on INSEE and State of Washington UI data.

Table 1
Results of Applying the Grouping Algorithm to Both Data Sets

Estimation by Direct Solution of the Least Squares Problem

Appendix 2 shows the exact algorithm used to solve equation (3), a variant of the conjugate gradient algorithm for which we customized the sparse representation of equation (3) so that very large problems with many X variables would be practical. In practice, we apply this algorithm to the full set of persons, firms and characteristics shown in the design matrices of equations (2) and (3). Unlike equation (4), the design matrix in equation (3) is not of full rank. Although the algorithm we use converges to a least squares solution, the parameter estimates are not unique. The output from the algorithm provides a non-unique set of effects to which we subsequently apply the identification procedure. To make the effects unique for each group, we eliminate one person effect by setting the group mean person effect to zero. We also set the overall mean person and firm effects equal 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.[7]

4. Some Results Comparing AKM, AFK, and Direct Least Squares

Summary of Data Sources

The French data are based on a collection of employer payroll reports called the Déclaration annuelles des données sociales. These consist of a 1/25th sample of the French work

[7] 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.

force with the individual and employing firm identified for the years 1976-1987 (1981 and 1983 are not available). There are approximately 1.2 million individuals, 500,000 firms and 5.3 million observations. The time varying characteristics consist of labor force experience (quartic), time period (annual), and region of France all fully interacted with sex. See AKM for a full description of the methods used to create the data and for summary statistics.

The State of Washington data are derived from unemployment insurance wage records, which are also employer reports. We use a 1/10th sample of State of Washington employment with the individual and the taxable employing entity identified for the years 1984-1993 (quarterly). There are approximately 293,000 individuals, 80,000 firms and 4.3 million observations used. The time-varying characteristics consist of labor force experience (quartic) and time period (annual and quarter) both fully interacted with sex and race. See AFK for a full description of the methods used to create the data and for summary statistics.

		Correlations								
	Standard						$x\beta$	θ	ψ	ε
	Deviation	*lny*	$x\beta$	θ	ψ	ε	*(approx.)*	*(approx.)*	*(approx.)*	*(approx.)*
French Data										
Log real annual full-time compensation	0.519	1.000	0.141	0.704	0.201	0.169	0.261	0.840	0.213	0.459
Time-varying characteristics	0.135	0.141	1.000	-0.068	0.023	0.000	0.731	-0.051	0.016	-0.057
Person effect	0.455	0.704	-0.068	1.000	-0.283	0.000	-0.017	0.836	0.021	0.044
Firm effect	0.285	0.201	0.023	-0.283	1.000	0.000	0.036	0.217	0.184	-0.022
Residual	0.206	0.169	0.000	0.000	0.000	1.000	-0.005	0.000	0.048	0.359
Time-varying characteristics (approximate)	0.146	0.261	0.731	-0.017	0.036	-0.005	1.000	0.001	0.019	-0.052
Person effect (approximate)	0.425	0.840	-0.051	0.836	0.217	0.000	0.001	1.000	0.097	0.016
Firm effect (approximate)	0.065	0.213	0.016	0.021	0.184	0.048	0.019	0.097	1.000	0.007
Residual (approximate)	0.238	0.459	-0.057	0.044	-0.022	0.359	-0.052	0.016	0.007	1.000
State of Washington Data										
Log real hourly compensation	0.527	1.000	0.304	0.511	0.518	0.306	0.323	0.585	0.478	0.331
Time-varying characteristics	0.380	0.304	1.000	-0.530	0.143	0.000	0.998	-0.485	0.172	0.000
Person effect	0.476	0.511	-0.530	1.000	-0.025	0.000	-0.512	0.960	0.020	0.000
Firm effect	0.231	0.518	0.143	-0.025	1.000	0.000	0.153	0.155	0.769	0.114
Residual	0.161	0.306	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.922
Time-varying characteristics (approximate)	0.361	0.323	0.998	-0.512	0.153	0.000	1.000	-0.469	0.181	0.000
Person effect (approximate)	0.470	0.585	-0.485	0.960	0.155	0.000	-0.469	1.000	0.050	0.000
Firm effect (approximate)	0.163	0.478	0.172	0.020	0.769	0.000	0.181	0.050	1.000	0.000
Residual (approximate)	0.175	0.331	0.000	0.000	0.114	0.922	0.000	0.000	0.000	1.000

Notes: The column headers use the symbols from the text while the row headers provide short definitions. All approximations are based on AKM (1999), persons first, formulas.

Sources: Authors' calculations based on the INSEE and State of Washington UI data.

Table 2

Correlations between Exact and Approximate Components of the Log Wage Rate Estimated from French and State of Washington Data

Table 2 shows the correlation structure of the estimated components of the real rate of compensation for France and the State of Washington. Two kinds of results are of interest. First, note that the person effects account for a much larger percentage of the variation in the French data than in the State of Washington data, where the two components have nearly identical importance. This was the substantive conclusion of the original research and it remains correct. In the French data, the exact solution for the person effect is correlated 0.836 with the approximate solution; whereas in the Washington data this correlation is 0.960. The approximation clearly worked very well in the Washington data and acceptably well in the French data. On the other hand, the exact firm effect is only correlated 0.184 with the approximate firm effect in the French data; whereas the two effects are correlated 0.769 in the

Washington data. The approximation did not work well for the French firm effects—probably because AKM were not able to include nearly as many control variables, *Z* in their notation, in the French model. The approximate method worked acceptably well for firm effects in Washington.

One very important substantive difference between the approximate and exact solutions for the person and firm effects is seen in the correlation between the two effects. In the approximate French and Washington solutions, this correlation is positive but very small. In the exact solution for France, the correlation between the person and firm effects is substantially negative, -0.283. In the exact solution for Washington this correlation is also negative, -0.025, but still of quite small magnitude.

As regards standard errors of estimated person and firm effects, we note that the formula for the standard errors of the coefficients of the time-varying characteristics (AKM, section 3) is directly applicable to the estimated β coefficients from the direct solution of equation (3). The approximate sampling variance formulas for the estimated person and firm effects are given here:

$$\mathrm{Var}\left[\hat{\theta}_i\right] \approx \frac{\mathrm{Var}\left[\varepsilon_{it}\right]}{T_i}, \mathrm{Var}\left[\hat{\psi}_j\right] \approx \frac{\mathrm{Var}\left[\varepsilon_{it}\right]}{N_j}$$

where N_j is the number of observations in firm *j*.

Finally, for those with access to standard linear modeling software, we provide the following recommendations for constructing the conditioning variables (*Z* in AKM) to get an accurate approximation of θ and ψ from the persons-first formulas in AKM. Order the firms by size. Create a partial design, *F*, using as many of the large firms as feasible in the linear modeling package (the results shown above for the State of Washington include the largest 1,700 firms in the partial *F*). Include in *Z* a complete set of industry by firm size indicators for firms too small to be included in the partial *F*. If additional estimation capacity exists, include interactions of the person-average characteristics with the industry by firm size indicators in *Z*.

5. Conclusions

As the availability of linked longitudinal labor market data becomes more widespread, as for example in the Longitudinal Employer-Household Dynamics program at the US Census Bureau (Abowd, Lane and Prevost, 2000), there is a clear need to develop computable solutions to statistical models like equation (1). We have provided here two contributions to this effort. First, the identification and estimation procedures described in this article enhance the modeling techniques that were used in AKM and AFK. All of the formulas in those papers that are based upon estimated person and firm effects (including firm-level analyses and all of the standard error formulas) can be used directly with the exact solution provided in this paper. Second, for those with access to standard linear modeling software, we have provided instructions for obtaining an accurate approximation using the persons-first formulas from AKM.

References

Abowd, John M. and Francis Kramarz, “Econometric Analysis of Linked Employer-Employee Data,” *Labour Economics,* 6 (March 1999(a)), pp. 53-74.

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*, Volume 3(B), Chapter 26 (Amsterdam: North Holland, 1999(b)), pp. 2629-2710.

Abowd, John M., Hampton Finer and Francis Kramarz, “Individual and Firm Heterogeneity in Compensation: An Analysis of Matched Longitudinal Employer-Employee Data for the State of Washington” in J. Haltiwanger *et al.* (eds.) *The Creation and Analysis of Employer-Employee Matched Data*, (Amsterdam: North Holland, 1999), pp. 3-24.

Abowd, John M., Francis Kramarz and David Margolis, “High Wage Workers and High Wage Firms,” *Econometrica* (March 1999), pp. 251-333.

Abowd, John M., Julia I. Lane and Ronald C. Prevost, “Design and Conceptual Issues in Realizing Analytical Enhancements through Data Linkages of Employer and Employee Data” in *Proceedings of the Federal Committee on Statistical Methodology* (November 2000).

Dongarra Jack J., Iain S. Duff, Danny C. Sorensen and Henk A.Van der Vorst, *Solving Linear Systems on Vector and Shared Memory Computers*, (Philadelphia: SIAM, 1991).

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

Neumaier, Arnold and Eildert Groeneveld, “Restricted Maximum Likelihood Estimation of Covariance in Sparse Linear Models,” *Genet. Sel. Evol.* 30 (1998), pp. 3-26.

Robinson, George K., “That BLUP is a Good Thing: The Estimation of Random Effects,” *Statistical Science,* 6 (1991), pp. 15-51.

Searle, Shayle R, *Linear Models for Unbalanced Data*, (New York: John Wiley, 1987), 536 pages.

Searle, Shayle R., George Casella and Charles E. McCulloch, *Variance Components,* (New York: John Wiley, 1992), 501 pages.

Weeks, David L. and Donald R. Williams, “A Note on the Determination of Connectedness in and N-way Cross Classification,” *Technometrics* 6, No. 3 (August 1964): 319-324.

Appendix 1: Proof of Necessity and Sufficiency of the Grouping Conditions

Once the persons and firms have been divided into *G* groups, we want to show that exactly $N+J-G$ linear functions of the group means, person and firm effects are estimable (identified). The grouping conditions imply that *G* group means are identified. Then, within each group *g*, at most $N_\mathrm{g} - 1$ linear functions of the person effects and $J_\mathrm{g} - 1$ linear functions of the firm effects are estimable. Thus, the maximum number of estimable linear functions of the effects is:

$$N+J-G=G+\sum_{g}\left(N_g-1+J_g-1\right).$$

This establishes that the grouping conditions are necessary for identification (see Searle 1987, p. 139). The *G* group means have no obvious economic interpretation. Hence, we use these *G* estimable effects to increase the number of estimable linear functions of the person and firm effects in our application.

To establish that the grouping conditions are sufficient, consider a sample with *J* firms and *N* workers. As above, denote by $\mathrm{E}[y_{it}]$ the projection of worker *i*'s wage at date *t* on the column space generated by the person and firm identifiers. For simplicity, suppress the effects of observable variables *X* and write $\mathrm{E}[y_{it}]$ as:

$$\mathrm{E}\left[y_{it}\right]=\theta_i+\psi_{\mathrm{J}(i,t)}$$

With the persons and firms connected into *G* groups, sufficiency requires that we prove that the group mean of *y* is estimable and that the effects θ_i and ψ_j, in group *g* are estimable up to constraints of the form:

$$\sum_{i\in\{\text{group } g\}} w_i\theta_i=0 \text{ and } \sum_{j\in\{\text{group } g\}} w_j\psi_j=0$$

where the w_j are arbitrary weights at least one of which is nonzero.[8]

The proof is by induction. Suppose that there are 2 firms and *N* workers. Assume that the two firms are connected; so, $G=1$. Then at least one worker, denoted as individual 1, is employed in both firms over the sample period. Denote the projection of this worker's wage as $\mathrm{E}\left[y_{1t_1}\right]=\theta_1+\psi_1$ at date 1 and $\mathrm{E}\left[y_{1t_2}\right]=\theta_1+\psi_2$ at date 2. There must exist weights w_1 and w_2, at least one of which is non-zero, such that the equations

$$w_1\psi_1+w_2\psi_2=0 \text{ and } \mathrm{E}\left[y_{1t_1}\right]-\mathrm{E}\left[y_{1t_2}\right]=\psi_1-\psi_2$$

[8] In the language of linear models, sufficiency requires that we show that the group mean of *y*, N_g-1 linearly independent contrasts of the θ_i, and J_g-1 linearly independent contrasts of the ψ_j are estimable.

can be solved exactly. Clearly $w_1 = 1$ and $w_2 = 0$ satisfy the conditions on the weights. Thus, exactly one linear function of the firm effects is estimable.

An exactly analogous argument applies to the person effects, implying that $N-1$ person effects are estimable. Since $G=1$, the group mean and the grand mean are identical and clearly estimable. Thus, the connectedness of the single group was sufficient to identify $N+J-G$ = $N+2-1$ linear functions of the grand mean, person, and firm effects.

Next, suppose there is a connected group *g* with J_g firms, exactly J_g-1 linear functions of the firm effects identified, N_g persons, and exactly $N_g - 1$ linear functions of the person effects identified. Consider the addition of one more connected firm to such a group, adding exactly one person who was not in the original group. Because the new firm is connected to the existing J_g firms in the group there exists at least one individual, say worker 1 who works for a firm in the identified group, say firm J_g, at date 1 and for the supplementary firm at date 2. Then, there must exist a set of $J_g + 1$ weights $\{w_j\}$that satisfy the relations

$$\sum_{j \le J_g} w_j \psi_j + w_{J_g+1} \psi_{J_g+1} = 0 \text{ and } \mathrm{E}\lfloor y_{1t_1} \rfloor - \mathrm{E}\lfloor y_{1t_2} \rfloor = \psi_{J_g} - \psi_{J_g+1}$$

to identify ψ_{J_g+1} given the hypothesized identification of $J_g - 1$ linear functions of the firm effects in the original group *g*. Clearly, $w_{J_g+1} = 0$ in combination with the original identification restrictions summarized by $w_1, \ldots, w_{J_g}$, at least one of which was non-zero, allows the estimation of the one additional linear function of the ψ_j involving ψ_{J_g+1}.

Once again, an exactly analogous argument applies to the person effects. If N_g persons are in the original connected group g, then $N_g - 1$ linear functions of the person effects are estimable by hypothesis. Let person 1 be one of the individuals who connects firm $J_g + 1$ to the original firms in group *g* (as above). Let person $N_g + 1$ be the individual in firm $J_g + 1$ who never worked for any of the original J_g firms in group *g*. (If such an individual does not exist then no new person effects are introduced by the addition of firm $J_g + 1$ to the connected group.) Let t_1 be the period in which these two individuals both work for firm $J_g + 1$. It suffices to show that there exists a set of $N_g + 1$ weights $\{w_i\}$, at least one of which is not zero, that permit solution of the two relations

$$\sum_{i \le N_g} w_i \theta_i + w_{N_g+1} \theta_{N_g+1} = 0 \text{ and } \mathrm{E}\lfloor y_{1t_1} \rfloor - \mathrm{E}\lfloor y_{N_g+1,t_1} \rfloor = \theta_1 - \theta_{N_g+1}$$

that identify an additional linear function of the θ_i involving θ_{N_g+1}, given the hypothesized identification of $N_g - 1$ linear functions of the person effects in the original group *g*. Clearly, the original $w_1, \ldots, w_{N_g}$ and $w_{N_g+1} = 0$ satisfy this condition.

The group mean of y is clearly estimable in group g. Thus, if $N_g + J_g - 1$ linear functions of the effects were estimable in the original group g, exactly two more linear functions of the effects are identifiable after the addition of firm $J_g + 1$ and individual $N_g + 1$ to the group.

Finally, consider the addition of group $g+1$ with N_{g+1} individuals and J_{g+1} firms. By construction none of the individuals or firms is contained in any of the existing g groups; so, the identification of the effects in the existing g groups is unaffected by the addition of group $g+1$. This completes the induction argument. The grouping conditions are, thus, both necessary and sufficient for the identification of $N + J - G$ linear functions of the group means and the individual and firm effects. ■

Appendix 2: Direct Least Squares Algorithm

We solve the normal equations (3) using the standard conjugate gradient (CG) algorithm with preconditioning as described in Dongarra *et al.* (1991). To map our problem into their notation, which solves the equation $Ax = b$ for *x*, with a positive definite symmetric coefficient matrix *A*, we define:

$$A = \begin{bmatrix} X'X & X'D & X'F \\ D'X & D'D & D'F \\ F'X & F'D & F'F \end{bmatrix},\ x = \begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix}, \text{and } b = \begin{bmatrix} X'y \\ D'y \\ F'y \end{bmatrix}.$$

We actually solve a preconditioned problem in order to speed convergence, with a preconditioning matrix:

$$K = \begin{bmatrix} X'X & 0 & 0 \\ 0 & D'D & 0 \\ 0 & 0 & F'F \end{bmatrix} = U'U \text{ and } U = \begin{bmatrix} u & 0 & 0 \\ 0 & d^{\frac{1}{2}} & 0 \\ 0 & 0 & f^{\frac{1}{2}} \end{bmatrix}$$

where *u* is the upper triangular matrix obtained by the Cholesky decomposition of $X'X$, $d^{\frac{1}{2}}$ is the diagonal square root matrix of $D'D$, the person counts, and $f^{\frac{1}{2}}$ is the diagonal square root matrix of $F'F$, the firm counts. The preconditioned problem we solve is $\tilde{A}\tilde{x} = \tilde{b}$ where

$$\tilde{A} = U'^{-1} AU^{'},\ \tilde{x} = U'^{-1} x, \text{ and } \tilde{b} = U'^{-1} b$$

Upon completion of the CG algorithm, we obtain the parameter estimates by transforming back:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\theta} \\ \hat{\psi} \end{bmatrix} = \hat{x} = U'\tilde{x}$$

The matrix $\tilde{A}$ now has a relatively simple form, with identity matrices on the diagonal:

$$\tilde{A} = \begin{bmatrix} I & u'^{-1} X'Dd^{-\frac{1}{2}} & u'^{-1} X'Ff^{-\frac{1}{2}} \\ d^{-\frac{1}{2}}D'Xu^{-1} & I & d^{-\frac{1}{2}}D'Ff^{-\frac{1}{2}} \\ f^{-\frac{1}{2}}F'Xu^{-1} & f^{-\frac{1}{2}}F'Dd^{-\frac{1}{2}} & I \end{bmatrix}$$

and, by symmetry, only the three components $u'^{-1} X'Dd^{-\frac{1}{2}}$, $u'^{-1} X'Ff^{-\frac{1}{2}}$, and $d^{-\frac{1}{2}}D'Ff^{-\frac{1}{2}}$ need to be stored. The matrix $d^{-\frac{1}{2}}D'Ff^{-\frac{1}{2}}$ is potentially enormous, of dimension $N \times J$, but by using a sparse representation we can store it in a linear array of size *ncells*, which is the number of distinct person-firm pairs in the data set.

The computation time for the CG algorithm is dominated by a matrix-vector product at each iteration. By representing $\widetilde{A}$ as we have, we minimize the time needed to compute that matrix-vector product and we also reduce the number of iterations by effective preconditioning.

Below is the CG algorithm reproduced from Dongarra *et al.* page 146, modified slightly to leave out the preconditioning matrix since the problem is already preconditioned.

Algorithm CG

```
$x_0 = 0$ (or initial guess)
$r_0 = b - Ax_0$
$p_{-1} = 0$
$\beta_{-1} = 0$
$w_0 = r_0$
$\rho_0 = (r_0 ' w_0)$
for $i = 0,1,2,3,\ldots$
    $p_i = w_i + \beta_{i-1} p_{i-1}$
    $q_i = Ap_i$
    $\alpha_i = \rho_i / (p_i ' q_i)$
    $x_{i+1} = x_i + \alpha_i p_i$
    $r_{i+1} = r_i - \alpha_i q_i$
    if $x_{i+1}$ accurate enough, then stop
    else do
        $w_{i+1} = r_{i+1}$
        $\rho_{i+1} = (r_{i+1} ' w_{i+1})$
        $\beta_i = \rho_{i+1} / \rho_p$
    end
end
```

The algorithm works by generating a set of search directions p_i and improved minimizers x_i with corresponding residuals $r_i = Ax_i - b$. At each iteration, the quantity α_i is found such that for $x_{i+1} = x_i + \alpha_i p_i$ the function $x_{i+1}'(Ax_{i+1} - b)$ is minimized. At the minimizer $\hat{x}$, the gradient $A\hat{x} - b$ is zero, which is the goal. The algorithm terminates when the desired level of accuracy is reached. For our application, termination occurs when the relative error of the residual is small, *i.e.,* when $\|r\|/\|b\| < 10^{-7}$.

Originally, we implemented a variant known as least squares conjugate gradient (LSCG), which is sometimes more numerically stable than the ordinary or preconditioned CG algorithm for least squares problems. We discovered through experimentation that the preconditioned CG algorithm is both numerically stable and faster than LSCG for the problem analyzed in this paper.

Fortran 90 code implementing the CG and grouping algorithms is available from the authors.