IAB DiscussionPaper

Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung

No. 3/2004

# Practial estimation methods for linked employer-employee data

*Martyn Andrews, Thorsten Schank, Richard Upward*

Bundesagentur für Arbeit

# Practical estimation methods for linked employer-employee data

*Martyn Andrews, Thorsten Schank, Richard Upward*

# Practical estimation methods for linked employer-employee data*

*Martyn Andrews, Thorsten Schank, Richard Upward*


## Abstract

Methods for the analysis of linked employer-employee data are not yet available in standard econometrics packages. In this paper, we make the fixed-effects methods developed orginally by Abowd, Kramarz, Margolis and others more accessible, where possible, and show how they can be implemented in Stata. To illustrate these techniques, we give an example using German linked data. There is a caveat: when the number of plants is prohibitively large and the investigator wants to estimate the correlation between the worker and firm unobserved heterogeneities, the regression-based techniques discussed are not feasible.

In this version of the paper, we replace our earlier Two-Step estimator by a Classical Minimum Distance estimator. [109 words]


**Keywords:** linked employee-employer panel data, fixed effects
**New JEL Classification:** C23, C87, J30

**Address for Correspondence:**

Dr. M.J. Andrews
School of Economic Studies
University of Manchester
Manchester, M13 9PL

Email: martyn.andrews@man.ac.uk
Phone: +44-(0)161-275-4874

* The authors thank the IAB (Institut für Arbeitsmarkt– und Berufsforschung, Nürnberg) for kindly supplying the data, in particular, Lutz Bellmann and Stephan Bender. Financial support from the British Academy under Grant SG-35691 is also gratefully acknowledged. The views expressed in this paper are solely those of the authors and are not those of the IAB. Comments from presentations at the Symposium of Multisource Databases, IAB, Nürnberg, July 2004, the 10th Annual Stata Users Group conference, London, 2004, the Institute of Social and Economic Research at Essex, and the Departments of Economics at Aberdeen, Kent, Manchester, and Warwick are gratefully acknowledged. The usual disclaimer applies. All calculations were performed with Stata8/SE and the code is available from http://www.nottingham.ac.uk/economics/staff/details/richard_upward.html.

# 1 Introduction

Labour market outcomes are driven by the decisions of both workers and firms. However, it is only recently that the analysis of both sides of the market has become possible using matched (or linked) employer-employee data. There is a growing literature, whose origins are associated mainly with Abowd, Kramarz and Margolis. In Abowd, Kramarz & Margolis (1999) (hereafter AKM), they re-examine the whole of issue of persistent inter-industry wage differentials. Many other labour-market issues have been analysed, including inter-firm differences in productivity; the effects of hiring, quits and layoffs on productivity; the impact of new technology on wages; job creation and destruction; the effects of training; estimates of the cost of worker displacement; and the effects of unions and collective bargaining.[1]

Most econometric investigations of labour market issues are based on datasets that are either supply-side (individual- or household-level datasets) or demand-side (plant- or firm-level).[2] If worker variables are correlated with firm variables, then any study that ignores information from the other side of the market will produce biased estimates. Biases also occur if the worker heterogeneity or the firm heterogeneity are correlated with the observables. For example, in AKM's (1999) paper 'High wage workers, high wage firms', it is unobservably better workers, in terms of wages, that are assumed to work in unobservably better firms.

Although there is a growing literature, the analysis of linked employer-employee data is not yet routine. There are two reasons why this research agenda has not moved on as quickly as it might. First, matched datasets involve linking together different sources of official information, and there are often technical, logistic and accessibility constraints that hinder progress. Second, there are various econometric issues to overcome, which mean that routine techniques and packages cannot be used. AKM's papers suggest these issues are quite technical. The objective of this paper, therefore, is to make these methods more accessible, where possible, and then show how they can be implemented in Stata. To illustrate these techniques, we give an example using German linked data, from the Institut für Arbeitsmarkt– und Berufsforschung, Nürnberg (hereafter IAB).[3]

A puzzle has emerged, in that the unobserved component of workers' wages appears to be *negatively* correlated with the unobserved component of firms' average wages. Apart from AKM's original study, which reported a positive correlation, all subsequent estimates

[1]See also Abowd & Kramarz (1999) and Haltiwanger, Lane, Spletzer, Theeuwes & Troske (1999) for early surveys of the wide range of issues covered in this literature.

[2]Some datasets ask questions about the other side of the market; for example, a firm identifier and plant-size is available in the BHPS. Also, in what follows, 'workers' and 'individuals' are synonyms.

[3]Hereafter we refer to the data as LIAB: Linked IAB data.

have been negative. Abowd, Creecy & Kramarz (2002) (hereafter ACK) report that this is because the approximation used in their earlier work gives different estimates when the models are re-estimated with the exact solution developed subsequently. ACK report correlations of $-0.283$ for French data and $-0.025$ for data from Washington State. Goux & Maurin (1999) find a correlation ranging from $+0.01$ to $-0.32$ depending on the time period chosen. Gruetter & Lalive (2003) find a correlation of $-0.543$ for Austrian data; Barth & Dale-Olsen (2003) report a correlation of between $-0.47$ and $-0.55$. Our own estimates from German data suggest a correlation of approximately zero.

The paper is organised as follows. In Section 2, we set out the generic model that best represents the econometrics of fixed-effects models using matched employee-employer data, and in Section 3 we describe the various methods that can be used to estimate this generic model. In Section 4, we describe the LIAB data that we use to illustrate these techniques, which are presented in Section 5. Section 6 concludes.

## 2 A generic model

Consider the following model with both employer and employee heterogeneity and employer and employee covariates:

$$y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{jt}\boldsymbol{\gamma} + \mathbf{u}_i\boldsymbol{\eta} + \mathbf{q}_j\boldsymbol{\rho} + \alpha_i + \phi_j + \epsilon_{it} \tag{1}$$

There are $i = 1, \ldots, N$ workers ($N$ is often millions) and $j = 1, \ldots, J$ firms ($J$ is often thousands); $y_{it}$ is the dependent variable; $\mathbf{x}_{it}$ and $\mathbf{u}_i$ are vectors of observable $i$-level covariates; $\mathbf{w}_{jt}$ and $\mathbf{q}_j$ are vectors of observable $j$-level covariates; and $\alpha_i$ and $\phi_j$ are (scalar) unobserved heterogeneities, correlated with observables and each other. Note that both $\alpha_i$ and $\mathbf{u}_i$ are variables that are time-invariant for workers and similarly $\phi_j$ and $\mathbf{q}_j$ are fixed over time for firms. $\mathbf{x}_{it}$, on the other hand, varies across $i$ and $t$, and $\mathbf{w}_{jt}$ varies across $j$ and $t$. (There is more on use of $j$ subscript below.) Equation (1) therefore contains all four possible types of information which a researcher might have about workers and firms. There are $K$ observed covariates in total.

Both workers and firms are assumed to enter and exit the panel, which means we have an unbalanced panel with $T_i$ observations per worker. There are $N^* = \sum_{i=1}^{N} T_i$ observations (worker-years) in total. Workers also change firms. This is crucial, as fixed-effects methods are identified by changers. In this paper, we assume $\epsilon_{it}$ is strictly exogenous, which implies that workers' mobility decisions are independent of $\epsilon_{it}$. It is worth noting that mobility may be a function of the observables and the time-invariant unobservables.

Suppose the investigator only has access to worker (or household) data, and therefore considers estimating

$$y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{u}_i\boldsymbol{\eta} + \alpha_i + \phi_j + \epsilon_{it}.$$

If the investigator does not observe the vector $[\mathbf{w}_{jt}, \mathbf{q}_j]$ then the estimates of $[\boldsymbol{\beta}, \boldsymbol{\eta}]$ are biased if the vector $[\mathbf{x}_{it}, \mathbf{u}_i]$ is correlated with these missing firm-level variables. However, he can still control for $\phi_j$ providing he knows the identity of the firm, as there are multiple observations on workers within the same firm, which means that there are no biases arising from $\phi_j$ being correlated with any of the observables. Now suppose the investigator only has access to a single cross section. Clearly, he can still control for $\phi_j$, but now he cannot control for $\alpha_i$ as it now part of the error term $\alpha_i + \epsilon_i$.

Now suppose the investigator has only firm-level data, and considers estimating:

$$y_{jt} = \mu + \mathbf{w}_{jt}\boldsymbol{\gamma} + \mathbf{q}_j\boldsymbol{\rho} + \phi_j + \alpha_{jt} + \epsilon_{jt}.$$

Now the unit of observation is a firm, which means that $[y_{jt}, \alpha_{jt}, \epsilon_{jt}]$ are averages over each firm's employees. If everything were observed, *including* the vector of worker-level variables $[\mathbf{x}_{jt}, \mathbf{u}_j]$ (eg average age of the firm's employees, or the proportion of males in the firm), then the aggregation of variables would just cause heteroskedasticity. However, not observing $[\mathbf{x}_{jt}, \mathbf{u}_j]$ causes bias if these variables are correlated with the vector $[\mathbf{w}_{jt}, \mathbf{q}_j]$. However, we can control for $\phi_j$ using firm-level fixed effects methods, but we cannot control for $\alpha_{jt}$, because it is part of the error term $\alpha_{jt} + \epsilon_{jt}$. This is the well-known aggregation bias caused by having firm-level rather than worker-level data.[4] To conclude, without linked data, there are obvious biases from not observing observables, and from not controlling for unobservables.

Turning back to Equation (1), we emphasise that it is usual to assume that the heterogeneity terms $\alpha_i$ and $\phi_j$ are correlated with the observables. This means that random effects methods are inconsistent, and so fixed effects methods are needed to estimate the parameters of interest. This, in turn, means that $[\boldsymbol{\rho}, \boldsymbol{\eta}]$, the parameter vector associated with the time-invariant variables, is not identified. Rather than dropping $[\mathbf{u}_i, \mathbf{q}_j]$, it is usual to define

$$\theta_i \equiv \alpha_i + \mathbf{u}_i\boldsymbol{\eta} \tag{2}$$

and

$$\psi_j \equiv \phi_j + \mathbf{q}_j\boldsymbol{\rho} \tag{3}$$

giving

$$y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{jt}\boldsymbol{\gamma} + \theta_i + \psi_j + \epsilon_{it}. \tag{4}$$

[4]Early estimates of the union wage differential in the UK came from plant-level data (WIRS), which typically did not have important information on the employees' backgrounds.

Estimates of $[\boldsymbol{\eta}, \boldsymbol{\rho}]$ can be recovered by making the additional random effects assumptions $\text{Cov}(\mathbf{u}_i, \alpha_i) = \text{Cov}(\mathbf{q}_j, \phi_j) = 0$. Hausman & Taylor (1981) show that it is possible to identify time-varying effects using fixed-effects methods whilst identifying non-time-varying effects using random-effects methods in the same regression. However, some investigators may be unhappy about having different assumptions depending on whether the variable is time-invariant, or otherwise, so in everything that follows, we consider the identification of $[\boldsymbol{\eta}, \boldsymbol{\rho}]$ as an optional extra rather than part of the main story.

# 3 Econometric methods

Equation (4) is the generic model that represents most of the existing literature. A number of fixed-effects methods have been proposed in the literature. In what follows, we describe each. Code that illustrates how each can be implemented in Stata (StataCorp 2003) is available from
`http://www.nottingham.ac.uk/economics/staff/details/richard_upward.html`.

## 3.1 Least squares dummy variables (LSDV)

AKM are the first to propose consistent estimates of the parameters of Equations (1–4). It needs emphasising that they are particularly interested in estimating $\theta_i$ and $\psi_j$, in addition to $[\boldsymbol{\beta}, \boldsymbol{\gamma}]$, for two reasons. The first is that they want to see whether estimates of $\theta_i$ and $\psi_j$ are correlated, hence the title 'High wage workers, high wage firms'. The second is that they want to recover estimates of $\boldsymbol{\rho}$ and $\boldsymbol{\eta}$ using Equations (2) and (3) respectively. Because the heterogeneity variables are assumed to be correlated with the observables, they note that the Least Squares Dummy Variables (LSDV) estimator has the best properties, for the usual reasons. The LSDV estimates of $\alpha_i$ are inconsistent, although unbiased. (See Wooldridge (2002, ch. 10) for assumptions and properties of panel data models.) The properties of the $\psi_j$ are the same as for $[\boldsymbol{\beta}, \boldsymbol{\gamma}]$, the parameters associated with the time-varying covariates $[\mathbf{x}_{it}, \mathbf{w}_{jt}]$.

There are two potential problems with actually computing this LSDV estimator. It is well known that a model with individual and time dummies (Baltagi's Two Way Fixed Effects Model, Section 3.2) gives algebraic solutions for the estimates of the effects of the covariates *and* both sets of dummies. Essentially, there is a matrix that sweeps out both sets of dummies in one go, which means that a regression involving transformed variables is performed. For the model here, there are two important differences. First, in Baltagi the data are balanced, whereas here both workers and firms can enter and exit the panel. Wansbeek & Kapteyn (1989) analyse Baltagi's model for unbalanced data, and obtain inelegant expressions that involve generalised inverses. Second, there is not a regular pattern between the firm and worker dummies as there is between Baltagi's individual and time dummies. It is the second that is the important difference, because it means that there is no algebraic transformation of the observables that sweeps away both heterogeneity terms in one go *and* which allows them to be recovered subsequently. To circumvent this second problem, AKM note that explicitly including dummy variables for the firm heterogeneity, but sweeping out the worker heterogeneity algebraically, gives exactly the same solution as the LSDV estimator.[5]

More precisely, the investigator must generate a dummy variable for each firm:

$$F_{it}^{j} = 1(j(i,t) = j) \quad j = 1, \ldots, J,$$

where 1( ) is the dummy variable indicator function and the function $j(i,t) = j$ maps worker $i$ at time $t$ to firm $j$. Now substitute

$$\psi_{j(it)} = \sum_{j=1}^{J} \psi_j F_{it}^{j} \tag{5}$$

into Equation (4).[6] The $\theta_i$ are removed by time-demeaning (or differencing) over $i$:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (\mathbf{w}_{jt} - \bar{\mathbf{w}}_i)\boldsymbol{\gamma} + \sum_{j=1}^{J} \psi_j (F_{it}^{j} - \bar{F}_i^{j}) + \epsilon_{it}, \tag{6}$$

where $\bar{z}_i = \sum_t z_{it}/T_i$ for any variable $z$. This means that $J$ de-meaned (or differenced) firm dummies actually need creating.[7] To distinguish this estimator from LSDV above, hereafter we label this estimator *FEiLSDVj*. They are identical estimators, but differ in how they are computed. The covariance matrix for FEiLSDVj needs the standard degrees-of-freedom adjustment, the formula for which is given in the next subsection.

We should note that $(F_{it}^{j} - \bar{F}_i^{j})$ will be zero for all $J$ dummies for any worker $i$ who does not change firm. Furthermore, if we have a sample of firms—rather than the population, as in AKM's studies—it will only be non-zero for workers who change from one firm within the sample to another firm in the sample. This means that for samples such as the LIAB, only a tiny proportion of workers have any non-zero terms. Identification of $\psi_j$ is driven

[5]In linear models, there is no distinction between removing the heterogeneity algebraically or adding two full sets of dummy variables, for workers and firms, and so the terminology LSDV applies to both.

[6]Equation (5) shows that it would be better to use non-Greek letter for heterogeneity $\psi_{j(it)}$, because it is a variable, not a parameter.

[7]Differencing is ignored hereafter. There are various reasons why it is easier to implement the covariance transformation. Normally, the decision whether to estimate the model in first differences or use the covariance transform depends on which give the more efficient estimates. Both estimators are consistent. See Wooldridge (2002, Section 10.6.3).

by the total number of such movers in each firm $j$. Some small firms may have no movers, in which case $\psi_j$ is not identified. Other small firms may have only a very few movers, in which case estimates of $\psi_j$ will be very imprecise. This means that it may be not be sensible to estimate $\psi_j$ for small firms, and instead one should group small firms together (this is what AKM and others do.)

To obtain estimates of the heterogeneity, first compute

$$\widehat{\psi}_{j(it)} = \sum_{j=1}^{J} \widehat{\psi}_j F_{it}^j \tag{7}$$

and then

$$\widehat{\theta}_i = \bar{y}_i - \overline{\widehat{\psi}}_i - \bar{\mathbf{x}}_i \widehat{\boldsymbol{\beta}} - \bar{\mathbf{w}}_i \widehat{\boldsymbol{\gamma}} \tag{8}$$

where $\overline{\widehat{\psi}}_i$ averages $\widehat{\psi}_{j(it)}$ over $t$.

There are two potential computational problems with this estimator. The first is the number of firms $J$, because the software needs to invert a matrix of dimension $(K + J) \times (K + J)$. For many applications, the number of firms is sufficiently small that FEiLSDVj is computationally feasible. For example, StataSE inverts 11,000 x 11,000 matrices. In our own empirical work, for reasons explained below, we only need to add approximately 2,000 firm dummies. There are many other situations where the number of firms/schools/doctors is sufficiently small. However, some datasets have tens of thousands of firms, or even hundreds of thousands (for, example, AKM and ACK). The second is the requirement that one must create and store $J$ mean-deviations for $N^*$ observations, meaning that the data matrix is $N^* \times (K + J)$. This may be prohibitively large for software packages which store all data in memory, such as Stata.

Some improvement in the storage efficiency of the $J$ mean-deviated firm dummies can be achieved in Stata by using the lowest common multiple of all values of $T_i$. For example, if the data span a maximum of 5 years then $T_i$ can be any value from $[1, 2, 3, 4, 5]$. Multiplying $F_{it}^j - \bar{F}_i^j$ by the lowest common multiple (in this case 60) yields a set of integers which can be stored in Stata as single bytes rather than 4- or 8-byte fractions.[8]

The memory requirements of the data matrix for the FEiLSDVj estimator are then approximately $(N^*J) + 4[N^*(K + 1)]$ bytes. We require $N^*J$ bytes for the mean-deviated firm dummies and $4[N^*(K + 1)]$ bytes for the remaining $K$ explanatory variables and the dependent variable, assuming each is stored as 4-bytes. In our example we have $N^* = 5,145,098$, $J = 1,821$ and $K = 64$, meaning that we require about 10GB of memory to proceed.

[8]Storing the mean-deviated firm dummies as integers also appears to improve the accuracy of the matrix inversion procedure.

It is worth emphasising that firm dummies are no different from any multi-category dummy, so long as workers can move from one category to another over time (eg region dummies, but not ethnicity dummies). This is why the notation $\mathbf{w}_{jt}$ and $\mathbf{q}_j$ is possibly confusing, since both are defined over every row indexed $it$. (Note that AKM use the notation $\mathbf{J}(i,t)$ to denote the mapping from worker $i$ at time $t$ to the firm $j$ in which they are employed.) This means that the index $j$ refers to the level of aggregation that $w_{jt}$ actually varies over.

### Identifying the unobserved firm effects

An important issue is establishing how many unique unobserved firm effects can be identified. First, effects cannot identified for firms which have no turnover; otherwise, $F_{it}^j - \bar{F}_i^j = 0$. These comprise a total of $J_2$ firms out of a total of $J$. Second, note that the firm dummies, when in mean-deviations, form a collinear set of variables

$$\sum_{j=1}^{J}(F_{it}^j - \bar{F}_i^j) = 0.$$

This is simply a consequence of having a collinear set of firm dummies, which sum to the constant before forming mean-deviations, and therefore sum to zero afterwards. In such a situation, one drops one of the firm dummies.

However, there is an additional identification issue, discussed by ACK. Identification of firm effects is only possible within a 'group', where a group is defined by the movement of workers between firms. A group contains all the workers who have ever worked for any of the firms in that group, and all the firms at which any of the workers were employed. A second (unconnected) group is defined only if no firm in the first group has ever employed any workers in the second, and no firms in the second group have ever employed any workers in the first. If there are $G_1$ separate groups of firms, then it is not possible to identify one firm per group for the reason above. Thus the total number of firms whose effect cannot be estimated is $G = G_1 + J_2$.

ACK conclude that the number of estimable/identified firm effects is $J - G$, and that the number of estimable/identified person effects is $N$, where $N$ is the number of workers observed twice or more. Thus the correct degrees of freedom when estimating Equation (4) is $N^* - K - (J - G) - N$. When estimating Equation (6), the actual correct degrees of freedom are $N^* - K - (J - G)$, and so estimated standard errors, both robust and non-robust, need scaling by

$$\sqrt{\frac{N^* - K - (J - G)}{N^* - K - (J - G) - N}}. \tag{9}$$

A second implication of the grouping of firms is that estimates of $\widehat{\psi}_j$ cannot be directly compared across groups. This is because it is arbitrary which $\psi_j$ is set equal to zero for normalisation in each group. The same issue applies to the resulting $\widehat{\theta}_i$. ACK suggest making the additional assumption that the average firm effect is the same across groups.

We have implemented the grouping algorithm in Stata, available from `http://www.nottingham.ac.uk/economics/staff/details/richard_upward.html`.

### Identifying the effects of time-invariant variables

If the investigator can implement the LSDV estimator on $i$- de-meaned data (FEiLSDVj), or implement one of AKM's other methods (discussed briefly below), AKM suggest that one can recover estimates of $\widehat{\alpha}_i$ and $\widehat{\phi}_j$ by estimating Equations (2, 3) as follows. First, run the auxiliary regressions:

$$\widehat{\theta}_i = \text{const} + \mathbf{u}_i\boldsymbol{\eta} + \text{error} \tag{10}$$

$$\widehat{\psi}_j = \text{const} + \mathbf{q}_j\boldsymbol{\rho} + \text{error} \tag{11}$$

which give consistent estimates of $\boldsymbol{\eta}$, $\boldsymbol{\rho}$ (AKM 1999, Section 3.4.4). Because $\alpha_i$ is dropped from (2), the identifying assumption is that $\text{Cov}(\mathbf{u}_i, \alpha_i) = 0$ or else there is omitted variable bias. Similarly, $\text{Cov}(\mathbf{q}_j, \phi_j) = 0$ is assumed in (3). One only needs $N$ observations to estimate (2) and $J$ observations to estimate (3). AKM estimate these equations by GLS, because of the aggregation to the firm-level. Because there are other causes of heteroskedasticity, one could use OLS and adjust the covariance matrix for clustering at the firm-level. Second, the investigator computes

$$\widehat{\alpha}_i = \widehat{\theta}_i - \mathbf{u}_i\widehat{\boldsymbol{\eta}} \tag{12}$$

$$\widehat{\phi}_j = \widehat{\psi}_j - \mathbf{q}_j\widehat{\boldsymbol{\rho}}. \tag{13}$$

$\theta$ and $\psi$ can be defined at three levels of aggregation:

| | | |
|---|---|---|
| $i, t$ | $\theta_i$ replicated $T_i$ times | $\psi_{j(i,t)}$ |
| $i$ | $\theta_i$ | $\bar{\psi}_i = \sum_{t=1}^{T_i} \psi_{j(it)}/T_i$ |
| $j$ | $\bar{\theta}_j = \sum_{(it)\in j} \theta_i/N_j$ | $\psi_j$ |

($N_j$ is the total number of worker-years observed in firm $j$.) AKM show that statistics based on aggregating $\widehat{\theta}_i$ and $\widehat{\alpha}_i$ to the level of the firm are consistent as $T_i$ goes to infinity (see also Chamberlain 1984). To conclude, one can analyse distributions of $\widehat{\psi}_j$, $\widehat{\theta}_i$, specifically to see whether they are correlated.

### 3.2 AKM's approximate methods

To deal with the large number of firm dummies, AKM propose a number of techniques in their (1999) paper that reduce the dimensionality of the problem. These require imposing further (testable) orthogonality assumptions. We do not discuss these further because ACK have recently developed a numerical solution for the LSDV estimator above.

### 3.3 ACK's Direct Least Squares (DLS)

ACK, in addition to providing a more accessible discussion of their earlier papers, provide a numerical solution to the LSDV estimator of (1). They call it a Direct Least Squares Algorithm. They also make it clear that these methods are only relevant if one wants to estimate the heterogeneities. Finally, they re-estimate their original models on Washington and French data, and show that the AKM approximate methods reported in their (1999) paper give poor results for the French data. Their solution involves an iterative technique that does not look easy to implement in standard software such as Stata.[9] More importantly, it is not regression based. The software is available from Abowd's website `http://instruct1.cit.cornell.edu/~jma7/abowdcv.html`.

### 3.4 Spell Fixed Effects

If one is not interested in the estimates of $\theta_i$ and $\psi_j$ themselves, consistent estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from Equation (4) are straightforward to obtain by taking differences or by time-demeaning within each unique worker-firm combination (or 'spell'). This is because for each spell of a worker within a firm neither $\theta_i$ nor $\psi_j$ vary. Defining $\lambda_s \equiv \theta_i + \psi_j$ as spell-level heterogeneity, which is swept out by subtracting averages at the spell-level, both $\theta_i$ and $\psi_j$ have disappeared:

$$y_{it} - \bar{y}_s = (\mathbf{x}_{it} - \bar{\mathbf{x}}_s)\boldsymbol{\beta} + (\mathbf{w}_{jt} - \bar{\mathbf{w}}_s)\boldsymbol{\gamma} + (\epsilon_{it} - \bar{\epsilon}_s). \tag{14}$$

[9] Gruetter & Lalive (2003) also have an iterative technique, but it does not provide a covariance matrix.

Again, the effects of $\mathbf{u}$ and $\mathbf{q}$ are not identified, because $\mathbf{u}_i - \bar{\mathbf{u}}_s = \mathbf{0}$ and $\mathbf{q}_j - \bar{\mathbf{q}}_s = \mathbf{0}$. In addition, any variable $x_{it}$ or $w_{jt}$ which is constant *within a spell* will also not be identified. One observation per spell is used up in identifying each spell fixed effect.[10]

This is basically the method that AKM discuss in Section 3.3, except they use differences rather than mean-deviations. AKM do not label this technique, so we call it *Spell FE* or *FE(s)*. AKM state that it is consistent, inefficient, and "cannot be used to identify separately the firm intercept ... and the person effect". It is clearly consistent as all the heterogeneity has been removed, and it is not the most efficient estimator because LSDV is. Because one cannot separate the worker and firm heterogeneities, AKM do not pursue this method further.

As when estimating any fixed-effects model, the standard errors may need correcting for the number of spells that the software has 'forgotten' about[11]

$$\sqrt{\frac{N^* - K}{N^* - K - S}}.$$

Unfortunately, given estimates of $\widehat{\lambda}_s$, one cannot recover $\widehat{\theta}_i$ and $\widehat{\psi}_j$. Even if $S > N + J$, so that one could regress $\widehat{\lambda}_s$ on worker and firm dummies, all that has happened is that $\boldsymbol{\beta}$ has been partitioned out of the problem, reducing the size of the problem by just $K$.

It is worth emphasising, however, that for many researchers this 'spell fixed effects' method is a practical and simple solution which does not present any computational difficulty, providing the investigator is not interested in analysing the heterogeneity post-estimation.

Spell FE is trivial to implement in Stata (again see our Stata code).

### Identifying the effects of time-invariant variables: Spell FEIV

We develop this method further to estimate the effects of time-constant variables $\mathbf{u}$ and $\mathbf{q}$, which get swept away being constant within a spell. Consider the standard one-way fixed-effects model (say, using worker-level data only)

$$y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\beta} + \theta_i + u_{it}. \tag{15}$$

The standard FE estimator of $\boldsymbol{\beta}$ can be interpreted as an IV estimator (Verbeek 2004, Section 10.2.5):

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{FE} &= [\Sigma_i\Sigma_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)]^{-1}\Sigma_i\Sigma_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(y_{it} - \bar{y}_i)\\ &= [\Sigma_i\Sigma_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\mathbf{x}_{it}]^{-1}\Sigma_i\Sigma_t(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'y_{it}\end{aligned}$$

[10] If there is just one observation per spell, then $y_{it} - \bar{y}_s = 0$, $\mathbf{x}_{it} - \bar{\mathbf{x}}_s = \mathbf{0}$, $\mathbf{w}_{jt} - \bar{\mathbf{w}}_s = \mathbf{0}$. This 'singleton' result can used to reduce the sample size.

[11] Stata has a command `areg` which does not need this correction. Also, it can correct the standard errors for clustering, which, in this context, should be at the firm level.

$x_{it} - \bar{x}_i$ is an ideal IV for any scalar $x_{it}$ because: (i) it is uncorrelated with the unobservable $\theta_i$, and (ii) it is correlated with $x_{it}$.

This implies one can estimate Equation (15) by IV GLS with $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$ as an IV for $\mathbf{x}_{it}$. The other extreme case uses $\mathbf{x}_{it}$ as an IV, which generates the random effects estimator.

The objective here is to estimate the parameters of Equation (1), not Equation (4). The above argument implies that it is possible to estimate the parameters on the time-varying variables by time-demeaning them, *and* to estimate the parameters of the time-invariant variables using random effects. This approach can be thought of as 'in between' the FE estimator (which cannot estimate the parameters on time-invariant variables) and the RE estimator (which does not allow for any correlation between the time-varying variables and the unobservable heterogeneity). All variables that are correlated with unobservables ($\mathbf{x}_{it}$, $\mathbf{w}_{jt}$) are instrumented by their mean deviations $\mathbf{x}_{it} - \bar{\mathbf{x}}_s$ and $\mathbf{w}_{jt} - \bar{\mathbf{w}}_s$ respectively. This is not possible for the time invariant variables, ($\mathbf{u}$, $\mathbf{q}$), which can only be instrumented by themselves, which means we are assuming that $\text{Cov}(\mathbf{u}_i, \alpha_i) = 0$ and $\text{Cov}(\mathbf{q}_j, \phi_j) = 0$. In other words we are making exactly the same assumptions for $\mathbf{u}$ and $\mathbf{q}$ as we have done throughout, which is why Spell FEIV is a side-issue. This is a special case of Hausman & Taylor's (1981) estimator.

### 3.5 A Classical Minimum Distance (CMD) method

The main problem with the FEiLSDVj estimator is that it requires the inversion of a $(K + J) \times (K + J)$ cross-product matrix. As noted, in some cases $J$ may be only a few thousand, and so the estimator is feasible. This is particularly true where we have a sample of firms, and if we only attempt to identify the firm effects for larger firms. There is another constraint however, which is the sheer number of observations, even when $J$ is sufficiently small. This is because the data matrix is $N^* \times (K + J)$, and might be prohibitively large for software packages that store data in memory rather than on disk. To circumvent this problem, we propose the following method, based on the fact that only movers between firms identify firm effects.[12]

We separate the model into observations for movers, subscripted by "1", and non-movers, subscripted by "2" by sorting the data by $t$ within $i$. There are $N_1^*$ mover-observations and $N_2^*$ non-mover-periods. We then write Equation (4) in matrix notation, where each

[12] This replaces a method we proposed in earlier versions of this paper, which we labelled a Two-step Method. This earlier method ignored the sampling error associated with estimating $\widehat{\psi}$ and so computed standard errors that are too small.

model is estimated separately:[13]

$$\widetilde{\mathbf{y}}_1 = \widetilde{\mathbf{X}}_1\boldsymbol{\beta}_1 + \widetilde{\mathbf{F}}_1\boldsymbol{\psi}_1 + \boldsymbol{\epsilon}_1 \tag{16}$$

$$\widetilde{\mathbf{y}}_2 = \widetilde{\mathbf{X}}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2. \tag{17}$$

Note that $\widetilde{\mathbf{y}}_1$, $\widetilde{\mathbf{y}}_2$, $\widetilde{\mathbf{X}}_1$, $\widetilde{\mathbf{X}}_2$ and $\widetilde{\mathbf{F}}_1$ have all been mean-deviated and defined viz $\widetilde{\mathbf{y}}_1 = \mathbf{M_D}\mathbf{y}_1$ etc, where $\mathbf{M_D} \equiv \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$. Denote the variances of the two error terms as $\sigma^2_{\epsilon 1}$ and $\sigma^2_{\epsilon 2}$. We now drop all columns of $\widetilde{\mathbf{F}}_1$ that are the zero vector, that is the $J_2$ firms that have no turnover. By definition, $\widetilde{\mathbf{F}}_2 \equiv \mathbf{0}$.

Because there are often very few movers, eliminating $\widetilde{\mathbf{F}}_2 \equiv \mathbf{0}$ from the model means that, by estimating the model for movers and non-movers separately, the memory constraints noted above are sided-stepped.[14] The Classical Minimum Distance (CMD) estimator forms a restricted estimator for $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ from $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\psi}_1$. (See Wooldridge (2002, ch. 14.6) for further details.)

In general, denote $\boldsymbol{\pi}$ as the $S \times 1$ unrestricted parameter vector and denote $\boldsymbol{\delta}$ as the $P \times 1$ restricted parameter vector. The constraint is $\boldsymbol{\pi} = \mathbf{h}(\boldsymbol{\delta})$. In CMD estimation, one estimates $\boldsymbol{\pi}$ and then finds a $\boldsymbol{\delta}$ such that the distance between $\hat{\boldsymbol{\pi}}$ and $\mathbf{h}(\boldsymbol{\delta})$ is minimised. An efficient CMD estimator uses any consistent estimator of asymptotic covariance matrix $\mathbf{V}$ to act as weighting matrix for the distance between $\hat{\boldsymbol{\pi}}$ and $\mathbf{h}(\boldsymbol{\delta})$, denoted $\widehat{\mathbf{V}}$. In other words, Efficient CMD solves:

$$\min_{\boldsymbol{\delta}}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\boldsymbol{\delta})]'\widehat{\mathbf{V}}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\boldsymbol{\delta})],$$

whose solution is

$$\hat{\boldsymbol{\delta}} = (\mathbf{H}'\widehat{\mathbf{V}}^{-1}\mathbf{H})^{-1}\mathbf{H}'\widehat{\mathbf{V}}^{-1}\hat{\boldsymbol{\pi}},$$

when the mapping from $\boldsymbol{\delta}$ to $\boldsymbol{\pi}$ is linear: $\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\delta}$. Also, the appropriate estimator of $\widehat{\text{Avar}}(\hat{\boldsymbol{\delta}})$ with which to conduct inference is

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\delta}}) = [\mathbf{H}'\widehat{\text{Avar}}(\hat{\boldsymbol{\pi}})^{-1}\mathbf{H}]^{-1} = [\mathbf{H}'\widehat{\mathbf{V}}^{-1}\mathbf{H}]^{-1}.$$

A test of the validity of the restrictions is given by Wooldridge (2002, Eqn. (14.76)):

$$[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\delta}})]'\widehat{\mathbf{V}}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\delta}})] \sim \chi^2(S - P).$$

For the model at hand, the constraint $\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\delta}$ is written:

$$\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\psi}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_J \\ \mathbf{I}_K & \mathbf{0} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\psi} \end{pmatrix}$$

[13] We dispense with the distinction between $\mathbf{x}_{it}$ variables and $\mathbf{w}_{jt}$ variables.

[14] There are good reasons for treating these as separate models. In models of assortative matching, it is possible that the correlation between observed and/or unobserved components might be higher for movers compared with non-movers.

where $\boldsymbol{\pi}$ is $(2K+J)\times 1$, $\boldsymbol{\delta}$ is $(K+J)\times 1$, and $\mathbf{H}$ is $(2K+J)\times(K+J)$.

The appropriate asymptotic covariance matrix is:

$$\widehat{\mathbf{V}} = \begin{bmatrix} \widehat{\mathbf{V}}_1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{V}}_2 \end{bmatrix} = \begin{bmatrix} \widehat{\sigma}_1^{-2}\begin{pmatrix} \widetilde{\mathbf{X}}_1'\widetilde{\mathbf{X}}_1 & \widetilde{\mathbf{X}}_1'\widetilde{\mathbf{F}}_1 \\ \widetilde{\mathbf{F}}_1'\widetilde{\mathbf{X}}_1 & \widetilde{\mathbf{F}}_1'\widetilde{\mathbf{F}}_1 \end{pmatrix} & \begin{matrix} \mathbf{0} \\ \mathbf{0} \end{matrix} \\ \begin{matrix} \mathbf{0} & \mathbf{0} \end{matrix} & \widehat{\sigma}_2^{-2}(\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2) \end{bmatrix}.$$

Given the general expressions immediately above, it follows that the restricted estimator $\hat{\boldsymbol{\delta}} = [\mathbf{H}'\widehat{\mathbf{V}}^{-1}\mathbf{H}]^{-1}\mathbf{H}'\widehat{\mathbf{V}}^{-1}\hat{\boldsymbol{\pi}}$ is given by:

$$\hat{\boldsymbol{\delta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\psi}} \end{pmatrix} = \left[\widehat{\mathbf{V}}_1^{-1} + \begin{pmatrix} \widehat{\mathbf{V}}_2^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right]^{-1} \left[\widehat{\mathbf{V}}_1^{-1}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\psi}}_1 \end{pmatrix} + \begin{pmatrix} \widehat{\mathbf{V}}_2^{-1}\hat{\boldsymbol{\beta}}_2 \\ \mathbf{0} \end{pmatrix}\right] \tag{18}$$

and that

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\delta}}) = [\mathbf{H}'\widehat{\mathbf{V}}^{-1}\mathbf{H}]^{-1} = \left[\widehat{\mathbf{V}}_1^{-1} + \begin{pmatrix} \widehat{\mathbf{V}}_2^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\right]^{-1}, \tag{19}$$

a $(K+J)\times(K+J)$ matrix. It should be emphasised that these expressions use standard (unrobust) covariance matrices. A robust version of this covariance matrix replaces $\widehat{\mathbf{V}}_1$ and $\widehat{\mathbf{V}}_2$ in Equation (19) by robust equivalents.

A standard criticism is that movers and non-movers are different groups of individuals and so one should model them separately. Before imposing $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, one should test these restrictions, although this rarely happens. Under $H_0$:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\psi}}_1 - \hat{\boldsymbol{\psi}} \end{pmatrix}' \widehat{\mathbf{V}}_1^{-1} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\psi}}_1 - \hat{\boldsymbol{\psi}} \end{pmatrix} + (\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}})'\widehat{\mathbf{V}}_2^{-1}(\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}) \sim \chi^2(K). \tag{20}$$

Finally, given estimates $\widehat{\psi}_{j(i,t)}$, one obtains estimates of $\widehat{\theta}_i$ using Equation (8) above.

It should be emphasised that the *only* price paid with this approach is that one cannot constrain $\sigma^2_{\epsilon 1} = \sigma^2_{\epsilon 2}$. The only difference between this and the LSDV estimator is because $\widehat{\mathbf{V}}_1^{-1}$ and $\widehat{\mathbf{V}}_2^{-1}$ come from separate regressions.

## 3.6 A road map?

To conclude the discussion of the methods discussed in this section, we outline a flow chart that should help the investigator decide which method is appropriate for his needs.

1. Does the investigator want to estimate employer and employee heterogeneity?

   **No** Use Spell-level FE

**Yes** ...

2. Are there too many firm dummies to add 'by hand'?

   **Yes** Use AKM techniques

   **No** ...

3. Is there enough memory?

   **Yes** Use (transformed) firm dummies (FEiLSDVj)

   **No** Use the CMD method

For all methods, one can recover estimates on $\mathbf{u}_i$ and $\mathbf{q}_j$ making standard RE assumptions. The Stata code for estimating all of the models outlined in this section, apart from ACK's Direct Least Squares, is available from
`http://www.nottingham.ac.uk/economics/staff/details/richard_upward.html`.

# 4 The data (LIAB)

## The IAB establishment panel (Betriebspanel)

The IAB (Institut für Arbeitsmarkt– und Berufsforschung) collect their own demand side data: the Betriebspanel is an establishment panel of $\approx 8,000$ establishments located in the former West Germany and $\approx 8,000$ establishments in the former East Germany.[15] It covers the period 1993–present (1996–present for East Germany) and covers 1% of all plants and 7% of all employees in the population. The establishments are selected using a fairly complicated weighting procedure. (See Kölling (2000) for full details on the Betriebspanel.) Information on each establishment includes:[16]

- Total employment (also disaggregated) (`size1-size10`)
- Standard hours (`lhbar`) and overtime hours
- Wage recognition (`B,B1,B2`)
- Output

[15]Because these are establishments, not firms, we dispense with the latter terminology hereafter. Establishments and plants are synonyms.

[16]If variables are used in tables below, their acronyms are also given. For full definitions, see Table 3.

- Exports
- Investment (`inv`)
- Wage bill
- Urbanicity (`urban1-urban10`)
- Geographical location
- Nationality of ownership (foreign in 2000)
- Technology (subjective measure)
- Organisational change (subjective measure)
- Profitability (`profit1-profit5`)
- Age of plant (`vin`) and whether parent is a single-plant firm (`single`)

## The employment statistics register (Beschäftigtenstatistik)

On the other side of the labour market, the IAB has access to the employment statistics register (Beschäftigtenstatistik). It is an administrative panel of all employees who are covered by the social security system (about 80% of total employment), and is collected by the plant. There is at least one compulsory notification during each calendar year. It covers 1975–present for West Germany and 1992-present for East Germany. It contains about 400 million records, covering about 46 million employees. (See Bender, Haas & Klose (2000) for full details on the Beschäftigtenstatistik.) Information on each worker includes:

- Gender (`female`), age (`age`), nationality (`foreign`), marital status (`married`)
- Start and end dates of every employment spell (`mjob` for more than one job)
- Occupation (3-digit) (`occ1-occ6`)
- Daily wages (left truncated and right censored) (`lw`, but see below for more information)
- Qualifications: education/apprenticeship (`qual1-qual6`)
- Industry (`ind1-ind10`)
- Region
- Establishment identification number

## The linked IAB employer-employee data (LIAB)

By using the establishment identification number, the IAB are able to associate each worker in the Beschäftigtenstatistik with an establishment in the IAB panel. Note that it is also possible to aggregate up all workers (not just those employed by establishments in the panel) to the establishment level. The particular dataset we use for this study was created by selecting all employees in the employment register who are employed by the surveyed establishments on June 30th each year.

## Sample used for wage equations

To illustrate the techniques outlined above, we estimate various standard wage equations. The sample we use covers 1993-1997, that is $1 \leq T_i \leq 5$, and is for West Germany only. We also drop observations for apprentices, part-timers, homeworkers and those with a daily wage of less than 10 $DM$. In addition, the data are right-censored.[17] As always, we also drop observations with missing values.

Workers change plants, and in particular, can change between plants that are surveyed in the IAB panel and plants that are not. In this study, we keep only those years (*it* rows) when a worker is working in an IAB-panel plant. This is because we do not observe $\mathbf{w}_{jt}$ or $\mathbf{q}_j$ in those years when a worker is working for a non-IAB plant. Table 1 summarises the data, in exactly the same format used by AKM.

[TABLE 1 ABOUT HERE]

Identification of unobserved plant-effects is driven only by those workers who change plants. Thus an important sub-sample comprises those workers who have two or more spells ($S_i > 1$) in IAB plants ('IAB movers'). In Table 1, workers who return to the same employer after an intervening spell with another employer are coded as starting a new spell. In Section 3.4, a spell is defined as any unique worker/employer combination, and so all periods a worker spends with a given employer are coded as a single spell. This is why there are 1,954,242 spells in Table 1 but only 1,953,774 spells in the regression sample.[18] This, and the sample of IAB movers, is summarised in Table 2.

Is the sample of IAB movers representative of the whole sample? As already discussed, the IAB-panel plants over-represent large plants in the population, and so workers in IAB plants are not a random sub-sample of the population. It is also possible that the 23,393 workers who move between IAB plants may not a random sub-sample of 1,930,260; exactly

[17] In a paper that is concerned with methods, this is not as issue, although one could deal with this in the same way as Gruetter & Lalive (2003, p.6).

[18] The 1,954,242 spells in Table 1 is calculated as 1,906,867 plus 2*22,806 plus 3*385 plus 4*2.

the same issue arises in all panel data models, which rely on movers for identification. (For example, estimates of union wage differentials based on a sample of joiners/quitters.)

Table 3 reports sample means: the first three columns average by workers whereas columns four to six average by plants. For example, in column one, the regression sample, 22.95% of workers are female whereas, on average, each plant employs 34.76% females (column four). These sets of means are often different from each other because of the underlying nature of the plant-size distribution. Workers are much more likely to work for large plants rather than small plants. Because large plants have higher wages, average log earnings are much smaller in column four than in column one. There are also big differences in sample means for whether married, qualifications, industry, union bargaining, investment and the age of the plant.

Column two corresponds to column one, but for the 23,393 workers who move between IAB plants. The difference between columns one and two is in column three. Column five corresponds to column four, but for the 1,821 plants that experience 'IAB turnover', that is employ workers who move between IAB plants. The difference between columns four and five is in column six. As we only identify 1,821 plants out of 4,376, the obvious question is whether these plants are *observably* the same, on average, as the 4,376? The same question applies to whether the 23,393 movers are observably the same as the 1,930,260 workers. In fact, the 1,821 plants pay lots more (0.1678 log-points), employ fewer females, employ more married workers, tend to be bigger firms located in different industries, and invest more (column six). Looking at individual workers, movers only get slightly more pay (0.0327 log-points), are younger, are less likely to be women, are more highly qualified, and are employed at plants with lower investment (column three).

[TABLE 3 ABOUT HERE]

Even if this sub-sample is not random, it does not follow that the estimates of 1,821 $\widehat{\psi}_j$ are inconsistent. This depends on what causes movement. If based on match quality, say $f(\alpha, \phi)$, then estimates are consistent because $\alpha, \phi$ are swept away. However, it is a strong assumption to suggest that movement is independent of $\epsilon$; any shock that affects workers and firms suggests that movement and $\epsilon$ are correlated.

We conclude this discussion on the identification of unobserved plant-effects by counting the number of movers for each plant. Figure 1 plots the cumulative frequency for the number of plants against the number of movers. For example, one plant has 1,886 movers, but 472 plants only have one mover, and 2,555 plants have no movers at all. This is a very skewed distribution, and is a feature of linked employee-employer datasets. The IAB panel is a 1% sample of plants. Even though it is a large sample, the probability of observing a worker moving from one IAB plant to another is very small. Even if one observed the population of plants, very small plants would experience little or no turnover in a five-year period, making estimation of their $\psi_j$ very noisy.

One possible strategy the investigator might adopt is to only identify $\psi_j$ for plants with more than $x$ movers, and group all remaining small plants into one plant (Abowd et al. 2002). Using Figure 1, we set $x = 30$, giving 211 large plants and one small plant (albeit with a lot of employees). To conclude, it is important for the investigator to be aware of how little information is sometimes used to identify each unobserved plant effect, especially if plants are small.

# 5 Results

[TABLE 4 ABOUT HERE]

Table 4 reports three conventional models, so described because they control for heterogeneity from only one side of the market, at best. The first is labelled Pooled OLS, which is Equation (1) where neither $\alpha_i$ nor $\phi_j$ are controlled for, of which there are three variants. The first only includes worker-level covariates, the second only plant-level covariates, and third includes both sets. The idea here is to assess the extent to which estimates on worker-level covariates are affected by the absence of plant-level covariates, and *vice versa*—in other words, to assess the extent to which the two sets of covariates are correlated with each other. A comparison of the estimates shows that the estimates do change, but not by much. The plant-level covariates move more, which is expected, given their standard errors are generally bigger.

The second model is labelled FE(i), ie the worker-level heterogeneity $\theta_i$ is controlled for, but $\phi_j$ becomes part of the model's error term:

$$y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{jt}\boldsymbol{\gamma} + \mathbf{q}_j\boldsymbol{\rho} + \theta_i + (\phi_j + \epsilon_{it})$$

Notice that the effects of the time-invariant worker-level variables $\mathbf{u}_i$ are not identified, namely `foreign` and `female`. The extent to which an estimate moves compared with Pooled OLS (previous column) depends on the extent to which $\theta_i$ is correlated with observed covariates. Here there are some large movements. Notice that Stata reports an estimate of the correlation between the deterministic part of the regression and $\theta_i$ ('`corr(ui,Xb)`'), and there is very strong negative correlation of –0.66, which is a different manifestation of the same thing.[19]

[19] '`corr(ui,Xb)`' varies from model to model. For FE(i), it is the correlation between $\widehat{\theta}_i$ and $\mathbf{x}_{it}\widehat{\boldsymbol{\beta}} + \mathbf{w}_{jt}\widehat{\boldsymbol{\gamma}} + \mathbf{q}_j\widehat{\boldsymbol{\rho}}$.

The third model is labelled FE(j)

$$y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{jt}\boldsymbol{\gamma} + \mathbf{u}_i\boldsymbol{\eta} + \psi_j + (\alpha_i + \epsilon_{it})$$

Now the plant-level heterogeneity $\psi_j$ is controlled for, but $\alpha_i$ becomes part of the model's error term. The effects of the time-invariant plant-level variables $\mathbf{q}_j$ are not identified, namely industry dummies and a dummy for whether the plant is `single`. This is not a model that one would normally estimate, but is useful if the investigator *cannot* control for both $\psi_j$ and $\alpha_i$ simultaneously, because it at least indicates the extent to which $\psi_j$ is correlated with the observed covariates. Here the correlation between $\psi_j$ and the deterministic part of the model is much weaker, and positive, at 0.08.

What is missing, of course, is that we do not control for any correlation of *both* unobserved fixed effects, $\phi_j$ and $\alpha_i$, with observable characteristics. Table 5 reports two models that do exactly this.

[TABLE 5 ABOUT HERE]

The first of these is FE(s), the easy-to-use technique that removes spell-level heterogeneity (Section 3.4 above). The effects of all time-invariant covariates are not identified, but the estimates of the time-varying covariates are consistent. If the investigator is not interested in estimating the worker- and plant- heterogeneities, he can stop here. Comparing these estimates with Pooled OLS and FE(i) in the previous table is of some considerable interest, as these are the better estimates. Notice that the correlation between the deterministic part of the regression and $\lambda_s$ is -0.56, which is approximately equal to the sum of those from FE(i) and FE(j). Given that $\lambda_s = \theta_i + \psi_j$, this is not surprising. The IV version that estimates the effects of time-invariant variables, under the extra assumptions $\text{Cov}(\mathbf{u}_i, \alpha_i) = \text{Cov}(\mathbf{q}_j, \phi_j) = 0$, is reported in the second column. The estimates of the time-varying covariates are virtually identical.

Following the 'road-map' outlined in Section 3.6, the next decision is to ascertain whether there are too many plant dummies to add 'by hand' when estimating Equation (6). This technique, if feasible, is labelled FEiLSDVj. 'By hand' means that dummies for each plant are explicitly added to the regression like any other covariate; that is, cannot be dealt with algebraically. In the models being estimated here, we have 5,145,098 observations, and need $J_1 = 1,821$ plant dummies, these being those plants which have IAB turnover, ie movers to/from another IAB plant. The memory needed is too prohibitive. As discussed on Page 6, we consider implementing the trick whereby we multiply the dummies by 60 so that they are stored as single bytes. This didn't work: we have $N^* = 5,145,098$, $J_1 = 1,821$ and $K = 64$, meaning that we still require about 10GB of memory to proceed.

Thus the only way forward is to use the CMD method outlined in Section 3.5. This is the second model in Table 5. There are $N_1 = 23,393$ movers ($N_1^* = 72,353$ observations) and $N_2 = 1,906,867$ non-movers ($N_2^* = 5,072,845$ observations). Even though there are $J = 4,376$ plants, we can only estimate effects for $J_1 = 1,821$ plants, because the rest experience no IAB-turnover. In addition, another $G_1 = 33$ effects cannot be identified, for one plant in each group. Thus the total number of identifiable plants is $J_1 - G_1 = 1,788$. The restricted estimates (third column) are formed from estimating models for movers and non-movers separately, and computed using Equation (18), together with robust standard-errors. Including 33 normalising restrictions, the $\widehat{\psi}_j$s can only be computed for the 1,821 plants, which represents 1,816,368 workers (4,883,331 observations). Each $\widehat{\psi}_j$ is normalised on the average $\widehat{\psi}_j$ for its group $g$. The last two columns in Table 5 report estimates of the auxiliary regressions shown in Equations (10, 11), whereby estimates of the time-invariant covariates $[\mathbf{u}_i, \mathbf{q}_j]$ are recovered, under the usual assumptions $\text{Cov}(\mathbf{u}_i, \alpha_i) = \text{Cov}(\mathbf{q}_j, \phi_j) = 0$. These are based on the 4,883,331 worker-year observations, ie are not aggregated to either the individual- or plant- level.

All three methods (Spell FE, Spell FEIV, and CMD) give very similar estimates of the time-varying covariates, which illustrates that the CMD method also gives consistent estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. However, the estimates of the time-invariant covariates do differ, probably because the estimates $\widehat{\psi}_{j(i,t)}$ used as a dependent variable in the last column are unreliable, given the discussion on their identification above. The estimates for the $\widehat{\theta}_i$ regression are much closer to Spell FEIV.

As emphasised repeatedly, the advantage of the CMD method over FE(s) is that estimates of $\theta_i$ and $\psi_j$ are obtained. The means of these two distributions are not identified, but estimates of their variances are easily computed, as is the correlation between them. It is the correlation that is particularly interesting, since it estimates the extent to which unobservably 'good' workers are employed in unobservably 'good' plants. The correlation between $\widehat{\psi}_j$, $\widehat{\theta}_i$, $\widetilde{\alpha}_i = \widetilde{\theta}_i - \mathbf{u}_i\widetilde{\boldsymbol{\eta}}$ and $\widehat{\phi}_j = \widehat{\psi}_j - \mathbf{q}_j\widehat{\boldsymbol{\rho}}$ are as follows (see Equations 12 and 13):

| | $\widehat{\theta}$ | $\widehat{\psi}$ | $\widetilde{\alpha}$ | $\widehat{\phi}$ |
|---|---|---|---|---|
| $\widehat{\theta}$ | 1.0000 | | | |
| $\widehat{\psi}$ | –0.1907 | 1.0000 | | |
| $\widetilde{\alpha}$ | 0.9580 | –0.2323 | 1.0000 | |
| $\widehat{\phi}$ | –0.2221 | 0.9486 | -0.2587 | 1.0000 |

Uses 4,883,331 *it* observations.

The important finding is that $\text{corr}(\widehat{\psi}, \widetilde{\theta}) = -0.1907$. This correlation has the wrong sign if one expects that unobservably 'good' workers would be employed in unobservably 'good' plants. However, all of the literature (summarised briefly in the Introduction) finds a negative correlation, which gives rise to the question as to whether this a genuine economic phenomenon or whether we there is a technical issue insofar as this estimate is downwards biased. Our own view is that it is the latter (Andrews, Schank & Upward 2004), and that the size of the bias decreases with the number of movers used in estimating each $\widehat{\psi}_j$.

Under the assumptions of the model, we have now consistent estimates of all the components of the RHS of Equation (4)

$$y_{it} = \mathbf{x}_{it}\widehat{\boldsymbol{\beta}} + \mathbf{w}_{jt}\widehat{\boldsymbol{\gamma}} + \widehat{\theta}_i + \widehat{\psi}_j + \widehat{\epsilon}_{it}$$

where the hat now refers to any consistent estimate (CMD, FEiLSDVj, or AKM's DLS). This allows us to analyse the correlations between the observed and unobserved components of wages, on both sides of the market:

| | $\widehat{\theta}$ | $\widehat{\psi}$ | $\mathbf{x}_{it}\widehat{\boldsymbol{\beta}}$ | $\mathbf{w}_{jt}\widehat{\boldsymbol{\gamma}}$ |
|---|---|---|---|---|
| $\widehat{\theta}$ | 1.0000 | | | |
| $\widehat{\psi}$ | -0.1907 | 1.0000 | | |
| $\mathbf{x}_{it}\widehat{\boldsymbol{\beta}}$ | 0.0827 | 0.0329 | 1.0000 | |
| $\mathbf{w}_{jt}\widehat{\boldsymbol{\gamma}}$ | 0.0211 | -0.3251 | -0.0526 | 1.0000 |

Uses 4,883,331 *it* observations.

| | $\widehat{\theta}$ | $\widehat{\psi}$ | $\mathbf{x}_{it}\widehat{\boldsymbol{\beta}}$ | $\mathbf{w}_{jt}\widehat{\boldsymbol{\gamma}}$ |
|---|---|---|---|---|
| $\widehat{\theta}$ | 1.0000 | | | |
| $\widehat{\psi}$ | -0.1775 | 1.0000 | | |
| $\mathbf{x}_{it}\widehat{\boldsymbol{\beta}}$ | 0.0866 | 0.0470 | 1.0000 | |
| $\mathbf{w}_{jt}\widehat{\boldsymbol{\gamma}}$ | 0.0170 | -0.3289 | -0.0725 | 1.0000 |

Averages to 1,816,368 *i* observations.

| | $\widehat{\theta}$ | $\widehat{\psi}$ | $\mathbf{x}_{it}\widehat{\boldsymbol{\beta}}$ | $\mathbf{w}_{jt}\widehat{\boldsymbol{\gamma}}$ |
|---|---|---|---|---|
| $\widehat{\theta}$ | 1.0000 | | | |
| $\widehat{\psi}$ | -0.4653 | 1.0000 | | |
| $\mathbf{x}_{it}\widehat{\boldsymbol{\beta}}$ | 0.0803 | 0.0823 | 1.0000 | |
| $\mathbf{w}_{jt}\widehat{\boldsymbol{\gamma}}$ | 0.0897 | -0.3641 | 0.0174 | 1.0000 |

Averages to 1,821 *j* observations.

Even though aggregating information to the plant-level means that estimators remain consistent, it is noticeable that correlations get bigger in absolute size. Looking at the *it*-level correlations, they generally make sense, except for those involving $\psi$. In particular, $\text{corr}(\widehat{\psi}, \mathbf{w}\widehat{\boldsymbol{\gamma}}) = -0.3251$ looks somewhat awry, as well as $\text{corr}(\widehat{\psi}, \widetilde{\theta})$ discussed above. The observed components are uncorrelated with each other, $\text{corr}(\mathbf{x}\widehat{\boldsymbol{\beta}}, \mathbf{w}\widehat{\boldsymbol{\gamma}}) = -0.0526$, which means that ignoring information from one side of the market does not affect estimates from the other side. All of the other cross-market correlations are small: $\text{corr}(\widehat{\theta}, \mathbf{w}\widehat{\boldsymbol{\gamma}}) = 0.0211$ and $\text{corr}(\widehat{\psi}, \mathbf{x}\widehat{\boldsymbol{\beta}}) = 0.0329$. Also, the unobserved and observed components of workers' characteristics are correlated, but weakly so, that is $\text{corr}(\widehat{\theta}, \mathbf{x}\widehat{\boldsymbol{\beta}}) = 0.0827$. In short, it is the three correlations that involve $\widehat{\psi}$ that looks wrong, and confirms these estimates of $\psi$ are often 'poor', being identified from plants that have very little turnover.

To investigate this further, we group all but the smallest 211 plants into one plant. Now all the plants are connected, ie $G = 1$. Compared with the figures given in Table 2, there are 20,313 movers (62,668 mover observations), 212 plants, and 40,719 spells. There are fewer movers, because any movement between one small plant and another is now counted as within-plant movement for this newly formed "plant". Moreover, there are now no plants without IAB turnover, and so all the data is used to compute $\widehat{\psi}_j$ and $\widehat{\theta}_i$. The results are reported in Table 6.

One advantage of estimating this model is that we are able to estimate it using FEiLSDVj (first three columns), which is the estimator with the best properties because is it LSDV. This allows us to make two comparisons. The first is to re-estimate the model using the CMD method (final three columns), thereby compare the two estimation methods directly. The second is that this CMD method for a model with 212 plants, in Table 6, can be compared with the same method applied to the model that has 1,821 plants, discussed immediately above and reported in Table 5.

The estimates for the two models are virtually identical to each other, as are the unrobust standard errors (unrobust standard errors are not reported). This illustrates clearly that our CMD method is virtually the same as LSDV. The obvious reason why there are some differences between the two methods is that CMD does not constrain $\sigma_{\epsilon 1} = \sigma_{\epsilon 2}$ across the mover and non-mover regressions. In fact, these are estimated as 0.0861 and 0.0676 respectively. Robust standard errors are reported in the table, and are generally lower than their LSDV counterparts by roughly 30%. This is because they are able to deal with one source of heteroskedasticity, namely $\widehat{\sigma}_{\epsilon 1} > \widehat{\sigma}_{\epsilon 2}$.[20]

The correlation between $\theta_i$ and $\psi_j$ is -0.0172 for both FEiLSDVj and CMD methods. This is much lower than the -0.1907 estimate that was reported above. This confirms the main conclusion from Andrews et al. (2004) that the more movers each plant has, the smaller is the downwards bias in the correlation. Andrews et al. (2004) also develop formulae for calculating the size of this bias. Thus the estimate is a lower bound: what we are not able to say is whether the true correlation is zero or positive, but at least this rules out negative assortative matching.

# 6 Conclusions

The main objective of this paper is to illustrate that the analysis of matched employee-employer datasets is more accessible than the investigators might imagine. We then

[20]Throughout this paper we cluster on $j$. A more conservative approach would have been to cluster on $i, t$, that is report standard robust standard errors.

show how they can be implemented in Stata. We illustrate with examples using linked employer-employee data from Germany (the Linked IAB data).

There are two points worth emphasising. The first is that investigators who are interested in estimating unobserved worker heterogeneity and unobserved worker heterogeneity, and who have a 'large' number of plants, must use ACK's Direct Least Squares algorithm. In this paper we explain how the investigator can make the feasible number of plants as large as possible without having to resort to ACK's algorithm. Our CMD method is virtually identical to the 'correct' FEiLSDVj method, and only differs because the error variances are different in the mover and non-mover regressions.

It is important to emphasise that the estimates of $\widehat{\psi}_j$ rely entirely on workers who change plants, as in any fixed-effects model. If one has a sample of plants, as here, there are very few movers (we have 1.9 million workers, but only 23,000 movers). The estimates on $\widehat{\psi}_j$ need interpreting with caution. Moreover, we suspect that the negative correlation usually found in such studies is biased downwards, and this is caused by standard least-squares sampling error. This issue is investigated in a companion paper (in preparation).

If investigators are not interested in estimating the worker and firm heterogeneities themselves, but merely wish to control for them, Spell-level FE is very straightforward to use.

# References


Abowd, J., Creecy, R. & Kramarz, F. (2002), Computing person and firm effects using linked longitudinal employer-employee data, Technical Paper 2002-06, U.S. Census Bureau, April.

Abowd, J. & Kramarz, F. (1999), The analysis of labor markets using matched employer-employee data, *in* O. Ashenfelter & D. Card, eds, 'Handbook of Labor Economics', Vol. 3B, Elsevier, Amsterdam, chapter 40, pp. 2567–627.

Abowd, J., Kramarz, F. & Margolis, D. (1999), 'High wage workers and high wage firms', *Econometrica* **67**, 251–333.

Andrews, M., Schank, T. & Upward, R. (2004), High wage workers and low wage firms: negative assortative matching or statistical artefact?, Mimeo, University of Manchester, November.

Baltagi, B. (2001), *Econometric Analysis of Panel Data*, second edn, Wiley.

Barth, E. & Dale-Olsen, H. (2003), Assortative matching in the labour market? Stylised facts about workers and plants, Mimeo, Institute for Social Research, Oslo,, February.

Bender, S., Haas, A. & Klose, C. (2000), 'The IAB employment subsample 1975-1995: Opportunities for analysis provided by the anonymised sample', *Schmollers Jahrbuch: Zeitschrift für Wirtschafts- und Sozialwissenschaften* **120**, 649–62.

Chamberlain, G. (1984), Panel data, *in* Z. Griliches & M. Intrilligator, eds, 'Handbook of Econometrics', Vol. 2, Elsevier, Amsterdam, chapter 22, pp. 1247–318.

Goux, D. & Maurin, E. (1999), 'Persistence of interindustry wage differentials: a re-examination using matched worker-firm panel data', *Journal of Labor Economics* **17**, 492–533.

Gruetter, M. & Lalive, R. (2003), Job mobility and industry wage differentials: evidence from matched employer employee data, Mimeo, University of Zurich, October.

Haltiwanger, J., Lane, J., Spletzer, J., Theeuwes, J. & Troske, K., eds (1999), *The creation and analysis of employer-employee matched data*, North-Holland.

Hausman, J. & Taylor, W. (1981), 'Panel data and unobservable individual effects', *Econometrica* **49**, 1377–98.

Kölling, A. (2000), 'The IAB establishment panel', *Schmollers Jahrbuch: Zeitschrift für Wirtschafts- und Sozialwissenschaften* **120**, 291–300.

StataCorp (2003), *Stata Statistical Software: Release 8*, Stata Corporation, College Station, TX.

Verbeek, M. (2004), *A Guide to Modern Econometrics*, second edn, Wiley.

Wansbeek, T. & Kapteyn, A. (1989), 'Estimation of the error components models with incomplete panels', *Journal of Econometrics* **41**, 341–60.

Wooldridge, J. (2002), *Econometric analysis of cross section and panel data*, MIT Press.

# Tables

Table 1: Reproducing AKM's table

| Years in | Number of Employers | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample | 1 | 1a | 2 | 3 | 4 | 5 | Total | Percent |
| 1 | 532,875 | 489,896 | | | | | 532,875 | 27.6 % |
| | 1 | 1 | | | | | | |
| 2 | 479,653 | 448,502 | 7,604 | | | | 487,257 | 25.2 % |
| | 2 | 2 | 11 | | | | | |
| 3 | 282,599 | 268,095 | 8,102 | 197 | | | 290,898 | 15.1 % |
| | 3 | 3 | 21 | 111 | | | | |
| 4 | 325,833 | 312,517 | 5,082 | 220 | 0 | | 331,135 | 17.2 % |
| | 4 | 4 | 22 | 112 | 1111 | | | |
| 5 | 285,907 | 273,965 | 2,018 | 168 | 2 | 0 | 288,095 | 14.9 % |
| | 5 | 5 | 23 | 122 | 1121 | 11111 | | |
| Total | 1,906,867 | 1,792,975 | 22,806 | 585 | 2 | 0 | 1,930,260 | 100.0 |
| Percent | 98.8% | 92.9% | 1.2 % | 0.0 % | 0.0 % | 0.0 % | 100.0 % | |

*Format of this table copied from Table 1 in Abowd et al. (1999). We report the most common employment configurations for each cell, which are described in terms of the number of consecutive years spent with each of the worker's employers (e.g. configuration 113 means that the worker spent 1 year with his first employer, then 1 year with his second employer and finally 4 years with his third employer). Column 1a refers to the subset of workers with only one employer whose employing plant had at least one other worker who had changed plants at least once in his career.

Table 2: The regression samples

| | $N^*$ | $N$ | $J$ | $S$ |
|---|---|---|---|---|
| Whole sample | 5,145,098 | 1,930,260 | 4,376 | 1,953,774 |
| | | | | |
| Workers who move to other IAB plants | 72,253 | 23,393 | 1,821 | 46,907 |
| Workers who don't move to other IAB plants | 5,072,845 | 1,906,867 | 4,376 | 1,906,867 |
| Workers in plants with movement to other IAB plants | 4,883,331 | 1,816,368 | 1,821 | 1,839,882 |
| | | | | |
| *Group all plants with fewer than 30 movers into one plant* | | | | |
| Workers who move to other IAB plants | 62,668 | 20,313 | 212 | 40,719 |
| Workers who don't move to other IAB plants | 5,082,430 | 1,909,947 | 212 | 1,909,947 |

Table 3: Sample means by individual and plant

| | all | IAB movers | diff | all plants | IAB turnover | diff |
|---|---|---|---|---|---|---|
| log daily wage in Pfennige (`lw`) | 9.7423 | 9.7750 | -0.0327 | 9.5150 | 9.6828 | -0.1678 |
| non-German nationality (`foreign`) | 0.1086 | 0.1118 | -0.0032 | 0.0876 | 0.0990 | -0.0114 |
| female (`female`) | 0.2295 | 0.1665 | 0.0629 | 0.3476 | 0.2739 | 0.0737 |
| married (`married`) | 0.6064 | 0.6165 | -0.0101 | 0.4911 | 0.5726 | -0.0815 |
| interaction (`marr*fem`) | 0.1044 | 0.0614 | 0.0430 | 0.1394 | 0.1196 | 0.0198 |
| age (`age`) | 39.1673 | 37.1191 | 2.0482 | 38.0608 | 39.1977 | -1.1369 |
| age-squared/100 (`age2/100`) | 16.5498 | 14.7014 | 1.8484 | 15.7159 | 16.5339 | -0.8180 |
| age-cubed/10000 (`age3/10000`) | 7.4582 | 6.1754 | 1.2829 | 6.9683 | 7.4291 | -0.4608 |
| with appren'ship, without A-levels (`qual2`) | 0.6419 | 0.6387 | 0.0033 | 0.7310 | 0.6660 | 0.0650 |
| without appren'ship, with A-levels (`qual3`) | 0.0085 | 0.0063 | 0.0022 | 0.0086 | 0.0075 | 0.0011 |
| with appren'ship, with A-levels (`qual4`) | 0.0400 | 0.0400 | 0.0000 | 0.0372 | 0.0393 | -0.0021 |
| technical college degree (`qual5`) | 0.0474 | 0.0720 | -0.0246 | 0.0298 | 0.0421 | -0.0123 |
| university degree (`qual6`) | 0.0527 | 0.0710 | -0.0183 | 0.0341 | 0.0432 | -0.0091 |
| skilled blue collar (`occ2`) | 0.1786 | 0.1424 | 0.0362 | 0.2034 | 0.1792 | 0.0242 |
| highly skilled blue collar (`occ3`) | 0.1433 | 0.2034 | -0.0601 | 0.0914 | 0.1248 | -0.0334 |
| unskilled white collar (`occ4`) | 0.1303 | 0.1144 | 0.0160 | 0.1947 | 0.1579 | 0.0368 |
| skilled white collar (`occ5`) | 0.1929 | 0.1599 | 0.0330 | 0.2607 | 0.2272 | 0.0335 |
| highly skilled white collar (`occ6`) | 0.0552 | 0.0525 | 0.0028 | 0.0752 | 0.0638 | 0.0114 |
| 1 if person has more than one job (`mjob`) | 0.0014 | 0.0017 | -0.0003 | 0.0019 | 0.0014 | 0.0006 |
| 1 if stand-alone plant (`single`) | 0.2577 | 0.2475 | 0.0102 | 0.6248 | 0.4305 | 0.1942 |
| 500.000+(outskirt) (`urban2`) | 0.0486 | 0.0285 | 0.0201 | 0.0809 | 0.0714 | 0.0095 |
| 100.000-499.999 (center) (`urban3`) | 0.1901 | 0.1477 | 0.0424 | 0.1245 | 0.1428 | -0.0182 |
| 100.000-499.999 (outskirt) (`urban4`) | 0.0316 | 0.0133 | 0.0183 | 0.0491 | 0.0357 | 0.0134 |
| 50.000-99.999 (center) (`urban5`) | 0.0211 | 0.0152 | 0.0059 | 0.0286 | 0.0313 | -0.0027 |
| 50.000-99.999 (outskirt) (`urban6`) | 0.0208 | 0.0151 | 0.0058 | 0.0194 | 0.0236 | -0.0042 |
| 20.000-49.999 (`urban7`) | 0.1006 | 0.0912 | 0.0093 | 0.1211 | 0.1104 | 0.0107 |
| 5.000-19.999 (`urban8`) | 0.1117 | 0.2473 | -0.1357 | 0.1538 | 0.1395 | 0.0143 |
| 2.000-4.999 (`urban9`) | 0.0170 | 0.0076 | 0.0094 | 0.0434 | 0.0335 | 0.0099 |
| below 2.000 (`urban10`) | 0.0078 | 0.0031 | 0.0047 | 0.0286 | 0.0181 | 0.0104 |
| electricity, gas & water, mining & quarrying (`ind2`) | 0.0526 | 0.0967 | -0.0441 | 0.0190 | 0.0302 | -0.0112 |
| manufacturing (without construction) (`ind3`) | 0.7002 | 0.7502 | -0.0500 | 0.3814 | 0.5662 | -0.1848 |
| construction (`ind4`) | 0.0145 | 0.0059 | 0.0086 | 0.0738 | 0.0412 | 0.0326 |
| wholesale and retail trade (`ind5`) | 0.0404 | 0.0230 | 0.0174 | 0.1616 | 0.0983 | 0.0633 |
| transport and communciation (`ind6`) | 0.0375 | 0.0301 | 0.0073 | 0.0484 | 0.0406 | 0.0078 |
| financial intermediation (`ind7`) | 0.0701 | 0.0437 | 0.0264 | 0.0526 | 0.0708 | -0.0183 |
| other services (`ind8`) | 0.0782 | 0.0451 | 0.0331 | 0.2315 | 0.1411 | 0.0904 |
| non-profit org's and private h'holds (`ind9`) | 0.0017 | 0.0015 | 0.0003 | 0.0089 | 0.0038 | 0.0051 |
| regional bodies and social security (`ind10`) | 0.0034 | 0.0034 | -0.0001 | 0.0048 | 0.0033 | 0.0015 |
| log weekly standard hours (excl'g overtime) (`lHbar`) | 3.6070 | 3.6048 | 0.0022 | 3.6412 | 3.6209 | 0.0202 |
| firm or sector bargaining (`B`) | 0.4723 | 0.4898 | -0.0175 | 0.2976 | 0.3765 | -0.0789 |
| sector bargaining (1995ff) (`B1`) | 0.4458 | 0.4450 | 0.0009 | 0.4296 | 0.4714 | -0.0418 |
| firm bargaining (1995ff) (`B2`) | 0.0510 | 0.0506 | 0.0005 | 0.0745 | 0.0692 | 0.0053 |
| investment in DM, divided by median inv (`inv`) | 8.5643 | 3.4075 | 5.1568 | 0.5149 | 1.1743 | -0.6594 |
| log concentrarion index (for emply't) (`lconc`) | -4.6199 | -4.3273 | -0.2926 | -6.8160 | -5.7049 | -1.1111 |
| plant size [1-4) (`size1`) | 0.0005 | 0.0000 | 0.0004 | 0.1045 | 0.0009 | 0.1036 |
| plant size [5,10) (`size2`) | 0.0012 | 0.0003 | 0.0010 | 0.1231 | 0.0090 | 0.1141 |
| plant size [10,20) (`size3`) | 0.0029 | 0.0009 | 0.0021 | 0.1102 | 0.0176 | 0.0926 |
| plant size [20,50) (`size4`) | 0.0092 | 0.0050 | 0.0042 | 0.1440 | 0.0628 | 0.0812 |
| plant size [50,100) (`size5`) | 0.0132 | 0.0101 | 0.0031 | 0.0871 | 0.0830 | 0.0041 |
| plant size [100,200) (`size6`) | 0.0267 | 0.0240 | 0.0026 | 0.0896 | 0.1146 | -0.0250 |
| plant size [200,500) (`size7`) | 0.0827 | 0.1019 | -0.0192 | 0.1261 | 0.2286 | -0.1025 |
| plant size [500,1000) (`size8`) | 0.1056 | 0.1455 | -0.0399 | 0.0706 | 0.1485 | -0.0779 |
| plant size [1000,5000) (`size9`) | 0.4848 | 0.5316 | -0.0468 | 0.1308 | 0.3013 | -0.1706 |
| profit 'good' (`profit2`) | 0.2037 | 0.1297 | 0.0739 | 0.2565 | 0.2337 | 0.0227 |
| profit 'satisfactory' (`profit3`) | 0.3772 | 0.4155 | -0.0382 | 0.3550 | 0.3455 | 0.0095 |
| profit 'just ok' (`profit4`) | 0.2197 | 0.2463 | -0.0266 | 0.2162 | 0.2273 | -0.0112 |
| profit 'bad' (`profit5`) | 0.1755 | 0.1791 | -0.0035 | 0.1386 | 0.1658 | -0.0272 |
| vtg*(1-vtgcen) [a](`vin`) | 0.7743 | 0.9064 | -0.1320 | 2.4125 | 1.6574 | 0.7551 |
| vin*vin (`vinsq`) | 8.2214 | 5.9894 | 2.2320 | 24.1205 | 17.6922 | 6.4283 |
| vtg*(1-vtgcen) (`cvin`) | 15.4860 | 12.5803 | 2.9058 | 9.3528 | 12.6862 | -3.3333 |
| cvin*cvin (`cvin2`) | 275.7671 | 223.0398 | 52.7273 | 1.6772 | 2.2738 | -0.5966 |
| 1 if 1994 (`year2`) | 0.2679 | 0.2438 | 0.0242 | 0.2085 | 0.2375 | -0.0290 |
| 1 if 1995 (`year3`) | 0.1956 | 0.1271 | 0.0685 | 0.2074 | 0.2199 | -0.0124 |
| 1 if 1996 (`year4`) | 0.1713 | 0.1859 | -0.0147 | 0.2243 | 0.2147 | 0.0096 |
| 1 if 1997 (`year5`) | 0.1505 | 0.1934 | -0.0429 | 0.2163 | 0.1653 | 0.0510 |
| No. of obs | 1,930,260 | 23,393 | | 4,376 | 1,821 | |

[a]Where `vtg` is age of the plant and `vtgcen` is 1 if age is censored, at 20 years.

## Table 4: Conventional models*

| | Pooled OLS | | | | |
|---|---|---|---|---|---|
| | w/o $[\mathbf{q}_j, \mathbf{w}_{jt}]$ | w/o $[\mathbf{u}_i, \mathbf{x}_{it}]$ | | FE(i) | FE(j) |
| foreign | -0.0163 (0.0053) | | -0.0183 (0.0038) | | -0.0209 (0.0021) |
| female | -0.1568 (0.0049) | | -0.1436 (0.0040) | | -0.1241 (0.0026) |
| married | 0.0401 (0.0050) | | 0.0375 (0.0031) | 0.0058 (0.0019) | 0.0419 (0.0015) |
| marr*fem | -0.0723 (0.0041) | | -0.0689 (0.0032) | -0.0037 (0.0027) | -0.0641 (0.0023) |
| age | 0.0688 (0.0033) | | 0.0727 (0.0029) | 0.1093 (0.0043) | 0.0715 (0.0027) |
| age2/100 | -0.1272 (0.0079) | | -0.1397 (0.0068) | -0.1660 (0.0082) | -0.1394 (0.0061) |
| age3/10000 | 0.0775 (0.0063) | | 0.0893 (0.0053) | 0.1070 (0.0064) | 0.0903 (0.0046) |
| qual2 | 0.1009 (0.0046) | | 0.0977 (0.0041) | 0.0140 (0.0050) | 0.0858 (0.0031) |
| qual3 | 0.1098 (0.0104) | | 0.0944 (0.0097) | -0.0512 (0.0261) | 0.0599 (0.0091) |
| qual4 | 0.1577 (0.0065) | | 0.1290 (0.0066) | 0.0217 (0.0119) | 0.1071 (0.0051) |
| qual5 | 0.2385 (0.0064) | | 0.2221 (0.0062) | 0.0467 (0.0089) | 0.1957 (0.0046) |
| qual6 | 0.2687 (0.0073) | | 0.2479 (0.0088) | 0.0577 (0.0127) | 0.2089 (0.0059) |
| occ2 | 0.0491 (0.0056) | | 0.0530 (0.0042) | 0.0008 (0.0033) | 0.0455 (0.0036) |
| occ3 | 0.2277 (0.0050) | | 0.2327 (0.0051) | 0.0357 (0.0042) | 0.2249 (0.0059) |
| occ4 | 0.0213 (0.0071) | | 0.0465 (0.0055) | -0.0046 (0.0030) | 0.0370 (0.0043) |
| occ5 | 0.1942 (0.0063) | | 0.1992 (0.0042) | 0.0185 (0.0045) | 0.1835 (0.0039) |
| occ6 | 0.2230 (0.0075) | | 0.2660 (0.0058) | 0.0351 (0.0043) | 0.2731 (0.0053) |
| mjob | -0.0672 (0.0090) | | -0.0580 (0.0072) | -0.0193 (0.0042) | -0.0509 (0.0083) |
| single | | -0.0242 (0.0082) | -0.0206 (0.0064) | -0.0081 (0.0091) | |
| ind2 | | 0.1888 (0.0347) | 0.1303 (0.0349) | 0.0441 (0.0793) | |
| ind3 | | 0.1610 (0.0307) | 0.1282 (0.0326) | 0.0622 (0.0776) | |
| ind4 | | 0.2346 (0.0329) | 0.1845 (0.0341) | -0.0016 (0.0815) | |
| ind5 | | 0.0563 (0.0351) | 0.0486 (0.0347) | 0.0060 (0.0784) | |
| ind6 | | 0.1036 (0.0326) | 0.0631 (0.0341) | 0.0004 (0.0815) | |
| ind7 | | 0.2577 (0.0312) | 0.1781 (0.0327) | 0.0449 (0.0788) | |
| ind8 | | 0.1258 (0.0326) | 0.0517 (0.0337) | 0.0039 (0.0780) | |
| ind9 | | 0.0635 (0.0479) | -0.0190 (0.0419) | 0.0392 (0.0803) | |
| ind10 | | 0.1658 (0.0499) | 0.0818 (0.0423) | 0.0410 (0.0817) | |
| lHbar | | -0.4735 (0.1448) | -0.3831 (0.0927) | -0.0605 (0.2583) | -0.0721 (0.2001) |
| B | | -0.0343 (0.0292) | -0.0076 (0.0203) | -0.0053 (0.0093) | -0.0016 (0.0054) |
| B1 | | -0.0506 (0.0379) | -0.0245 (0.0250) | 0.0004 (0.0078) | 0.0085 (0.0059) |
| B2 | | -0.0481 (0.0384) | -0.0182 (0.0257) | -0.0185 (0.0102) | -0.0100 (0.0077) |
| inv | | -0.0002 (0.0002) | -0.0002 (0.0001) | 0.0001 (0.0000) | 0.0001 (0.0000) |
| lconc | | 0.0005 (0.0032) | 0.0035 (0.0023) | -0.0044 (0.0027) | -0.0021 (0.0024) |
| size1 | | -0.5932 (0.0351) | -0.5149 (0.0276) | -0.0818 (0.0225) | -0.0280 (0.0243) |
| size2 | | -0.4246 (0.0265) | -0.3513 (0.0213) | -0.0702 (0.0176) | -0.0030 (0.0180) |
| size3 | | -0.3038 (0.0237) | -0.2517 (0.0184) | -0.0565 (0.0163) | 0.0073 (0.0156) |
| size4 | | -0.2341 (0.0201) | -0.1964 (0.0158) | -0.0455 (0.0139) | 0.0143 (0.0126) |
| size5 | | -0.1992 (0.0203) | -0.1633 (0.0155) | -0.0328 (0.0115) | 0.0199 (0.0111) |
| size6 | | -0.1665 (0.0173) | -0.1334 (0.0134) | -0.0228 (0.0102) | 0.0117 (0.0096) |
| size7 | | -0.1217 (0.0154) | -0.0948 (0.0125) | -0.0107 (0.0096) | 0.0130 (0.0087) |
| size8 | | -0.1017 (0.0157) | -0.0809 (0.0127) | -0.0039 (0.0082) | 0.0072 (0.0072) |
| size9 | | -0.0681 (0.0139) | -0.0607 (0.0115) | 0.0004 (0.0074) | 0.0045 (0.0064) |
| profit2 | | -0.0350 (0.0153) | -0.0163 (0.0111) | -0.0015 (0.0042) | -0.0033 (0.0031) |
| profit3 | | -0.0448 (0.0168) | -0.0260 (0.0116) | -0.0067 (0.0046) | -0.0082 (0.0034) |
| profit4 | | -0.0259 (0.0174) | -0.0134 (0.0129) | -0.0078 (0.0051) | -0.0091 (0.0038) |
| profit5 | | -0.0440 (0.0180) | -0.0400 (0.0125) | -0.0094 (0.0061) | -0.0114 (0.0045) |
| vin | | -0.0093 (0.0052) | -0.0035 (0.0037) | -0.0054 (0.0021) | 0.0159 (0.0020) |
| vinsq | | 0.0003 (0.0003) | 0.0001 (0.0002) | 0.0001 (0.0001) | 0.0002 (0.0001) |
| cvin | | -0.0093 (0.0041) | -0.0051 (0.0029) | -0.0018 (0.0020) | -0.0058 (0.0278) |
| cvin2 | | 0.0004 (0.0002) | 0.0002 (0.0001) | 0.0000 (0.0001) | 0.0007 (0.0008) |
| year2 | 0.0198 (0.0079) | 0.0225 (0.0062) | 0.0164 (0.0052) | -0.0047 (0.0048) | 0.0012 (0.0030) |
| year3 | 0.0467 (0.0105) | 0.0734 (0.0312) | 0.0656 (0.0207) | 0.0123 (0.0037) | 0.0155 (0.0031) |
| year4 | 0.0673 (0.0105) | 0.0894 (0.0318) | 0.0777 (0.0212) | 0.0060 (0.0023) | 0.0094 (0.0018) |
| year5 | 0.0793 (0.0103) | 0.1035 (0.0327) | 0.0866 (0.0217) | | |
| cons | 8.3740 (0.0533) | 11.5001 (0.5287) | 9.7534 (0.3449) | 7.5951 (0.9940) | 8.5528 (0.8486) |
| psihat | | | | | |
| No. of obs | 5,145,098 | 5,145,098 | 5,145,098 | 5,145,098 | 5,145,098 |
| No. of workers | 1,930,260 | 1,930,260 | 1,930,260 | 1,930,260 | 1,930,260 |
| No. of plants | 4,376 | 4,376 | 4,376 | 4,376 | 4,376 |
| No. of spells | | | | | |
| ‘`corr(ui,Xb)`’ | not applic | not applic | not applic | -0.6591 | 0.0773 |
| $\sigma_\theta$ or $\sigma_\psi$ | not applic | not applic | not applic | 0.3529 | 0.2968 |
| $\sigma_\epsilon$ | 0.2015 | 0.2610 | 0.1895 | 0.0680 | 0.1687 |

*6 urbanicity dummies also included (not reported to save space). For all regressions, we report robust standard errors adjusted for clustering on firms.

## Table 5: Double heterogeneity models*

| | | | CMD method | | |
|---|---|---|---|---|---|
| | Spell FE | Spell FEIV | Restricted | $\widehat{\theta}$ | $\widehat{\psi}$ |
| foreign | | -0.1091 (0.0024) | | -0.1140 (0.0054) | |
| female | | -0.1360 (0.0020) | | -0.1407 (0.0059) | |
| married | 0.0056 (0.0020) | 0.0056 (0.0002) | 0.0056 (0.0020) | | |
| marr*fem | -0.0036 (0.0028) | -0.0036 (0.0004) | -0.0036 (0.0026) | | |
| age | 0.1035 (0.0045) | 0.1035 (0.0004) | 0.0871 (0.0327) | | |
| age2/100 | -0.1643 (0.0084) | -0.1643 (0.0009) | -0.1649 (0.0077) | | |
| age3/10000 | 0.1060 (0.0065) | 0.1057 (0.0008) | 0.1061 (0.0059) | | |
| qual2 | 0.0108 (0.0053) | 0.0108 (0.0008) | 0.0113 (0.0052) | | |
| qual3 | -0.0615 (0.0313) | -0.0615 (0.0025) | -0.0571 (0.0153) | | |
| qual4 | 0.0157 (0.0160) | 0.0157 (0.0018) | 0.0174 (0.0156) | | |
| qual5 | 0.0518 (0.0118) | 0.0518 (0.0017) | 0.0479 (0.0080) | | |
| qual6 | 0.0632 (0.0158) | 0.0632 (0.0020) | 0.0586 (0.0115) | | |
| occ2 | 0.0010 (0.0034) | 0.0010 (0.0004) | 0.0012 (0.0029) | | |
| occ3 | 0.0359 (0.0043) | 0.0359 (0.0005) | 0.0363 (0.0038) | | |
| occ4 | -0.0035 (0.0031) | -0.0035 (0.0005) | -0.0035 (0.0026) | | |
| occ5 | 0.0193 (0.0045) | 0.0193 (0.0006) | 0.0194 (0.0041) | | |
| occ6 | 0.0354 (0.0043) | 0.0354 (0.0008) | 0.0358 (0.0034) | | |
| mjob | -0.0208 (0.0044) | -0.0208 (0.0015) | -0.0202 (0.0045) | | |
| single | | -0.0652 (0.0022) | | | -0.0256 (0.0081) |
| ind2 | | 0.1903 (0.0213) | | | 0.1243 (0.0706) |
| ind3 | | 0.2140 (0.0208) | | | 0.1393 (0.0694) |
| ind4 | | 0.1685 (0.0217) | | | 0.1182 (0.0731) |
| ind5 | | 0.0958 (0.0209) | | | 0.0278 (0.0703) |
| ind6 | | 0.1749 (0.0212) | | | 0.0855 (0.0721) |
| ind7 | | 0.3407 (0.0212) | | | 0.1269 (0.0703) |
| ind8 | | 0.1631 (0.0207) | | | 0.0490 (0.0709) |
| ind9 | | 0.0823 (0.0273) | | | -0.0188 (0.0786) |
| ind10 | | 0.2525 (0.0246) | | | 0.0489 (0.0856) |
| lHbar | -0.0564 (0.2645) | -0.0564 (0.0016) | -0.0563 (0.1063) | | |
| B | -0.0064 (0.0090) | -0.0064 (0.0004) | -0.0064 (0.0082) | | |
| B1 | -0.0002 (0.0078) | -0.0002 (0.0004) | -0.0001 (0.0067) | | |
| B2 | -0.0187 (0.0101) | -0.0187 (0.0004) | -0.0187 (0.0086) | | |
| inv | 0.0001 (0.0000) | 0.0001 (0.0000) | 0.0001 (0.0000) | | |
| lconc | -0.0059 (0.0030) | -0.0059 (0.0001) | -0.0059 (0.0025) | | |
| size1 | -0.0721 (0.0232) | -0.0721 (0.0052) | -0.0719 (0.0220) | | |
| size2 | -0.0614 (0.0184) | -0.0614 (0.0034) | -0.0613 (0.0166) | | |
| size3 | -0.0516 (0.0171) | -0.0516 (0.0019) | -0.0514 (0.0148) | | |
| size4 | -0.0391 (0.0146) | -0.0391 (0.0013) | -0.0389 (0.0138) | | |
| size5 | -0.0272 (0.0124) | -0.0272 (0.0010) | -0.0272 (0.0114) | | |
| size6 | -0.0192 (0.0112) | -0.0192 (0.0007) | -0.0193 (0.0102) | | |
| size7 | -0.0099 (0.0105) | -0.0099 (0.0005) | -0.0099 (0.0096) | | |
| size8 | -0.0026 (0.0090) | -0.0026 (0.0004) | -0.0026 (0.0079) | | |
| size9 | 0.0021 (0.0080) | 0.0021 (0.0003) | 0.0021 (0.0069) | | |
| profit2 | -0.0020 (0.0043) | -0.0020 (0.0002) | -0.0019 (0.0039) | | |
| profit3 | -0.0072 (0.0047) | -0.0072 (0.0002) | -0.0072 (0.0045) | | |
| profit4 | -0.0082 (0.0053) | -0.0082 (0.0003) | -0.0082 (0.0050) | | |
| profit5 | -0.0099 (0.0062) | -0.0099 (0.0003) | -0.0098 (0.0058) | | |
| vin | | | 0.0165 (0.0324) | | |
| vinsq | 0.0001 (0.0001) | 0.0001 (0.0000) | 0.0001 (0.0001) | | |
| cvin | -0.0160 (0.0354) | -0.0160 (0.0021) | | | |
| cvin2 | 0.0005 (0.0010) | 0.0005 (0.0001) | 0.0005 (0.0009) | | |
| year2 | -0.0032 (0.0037) | -0.0032 (0.0002) | -0.0032 (0.0031) | | |
| year3 | 0.0140 (0.0041) | 0.0140 (0.0003) | 0.0141 (0.0036) | | |
| year4 | 0.0073 (0.0024) | 0.0073 (0.0002) | 0.0073 (0.0024) | | |
| year5 | | | | | |
| cons | 7.8941 (1.1414) | 7.7591 (0.0257) | | 8.3224 (0.0062) | -0.1118 (0.0695) |
| No. of obs. | 5,145,098 | 5,145,098 | 5,145,098 | 4,883,331 | 4,883,331 |
| No. of workers | 1,930,260 | 1,930,260 | 1,930,260 | 1,816,368 | 1,816,368 |
| No. of plants | 4,376 | 4,376 | 4,376 | 1,821 | 1,821 |
| No. of spells | 1,953,774 | 1,953,774 | | | |
| `corr(ui,Xb)` | -0.5625 | | | | |
| $\sigma_\lambda$ | 0.3220 | | | | |
| $\sigma_\epsilon$ | 0.0675 | | | 0.2212 | 0.1122 |

*6 urbanicity dummies also included (not reported to save space). For Spell FE, Pooled CMD, and the auxiliary regressions for $\widehat{\theta}_i$ and $\widehat{\psi}_j$, we report robust standard errors adjusted for clustering on plants. Stata does not give robust standard errors for its IV GLS routine. See text for how robust standard errors are computed for Pooled CMD.

Table 6: Models with only 212 large plants*

| | FEiLSDVj | | | CMD method | | |
|---|---|---|---|---|---|---|
| | | thetahat | psihat | Restricted | $\widehat{\theta}$ | $\widehat{\psi}$ |
| foreign | | -0.1106 (0.0076) | | | -0.1105 (0.0076) | |
| female | | -0.1148 (0.0068) | | | -0.1147 (0.0068) | |
| married | 0.0057 (0.0020) | | | 0.0057 (0.0014) | | |
| marr*fem | -0.0036 (0.0028) | | | -0.0036 (0.0021) | | |
| age | 0.1068 (0.0042) | | | 0.1066 (0.0036) | | |
| age2/100 | -0.1656 (0.0082) | | | -0.1653 (0.0075) | | |
| age3/10000 | 0.1067 (0.0064) | | | 0.1064 (0.0057) | | |
| qual2 | 0.0130 (0.0049) | | | 0.0128 (0.0042) | | |
| qual3 | -0.0516 (0.0261) | | | -0.0529 (0.0137) | | |
| qual4 | 0.0195 (0.0122) | | | 0.0196 (0.0063) | | |
| qual5 | 0.0459 (0.0089) | | | 0.0475 (0.0064) | | |
| qual6 | 0.0560 (0.0123) | | | 0.0577 (0.0084) | | |
| occ2 | 0.0011 (0.0033) | | | 0.0011 (0.0022) | | |
| occ3 | 0.0358 (0.0042) | | | 0.0358 (0.0034) | | |
| occ4 | -0.0046 (0.0029) | | | -0.0044 (0.0020) | | |
| occ5 | 0.0183 (0.0044) | | | 0.0186 (0.0033) | | |
| occ6 | 0.0348 (0.0041) | | | 0.0348 (0.0032) | | |
| mjob | -0.0196 (0.0042) | | | -0.0196 (0.0041) | | |
| single | | | -0.0169 (0.0062) | | | -0.0170 (0.0062) |
| ind2 | | | 0.0141 (0.0108) | | | 0.0143 (0.0108) |
| ind3 | | | 0.0251 (0.0108) | | | 0.0253 (0.0108) |
| ind4 | | | 0.0073 (0.0091) | | | 0.0073 (0.0092) |
| ind5 | | | -0.0044 (0.0107) | | | -0.0044 (0.0107) |
| ind6 | | | 0.0058 (0.0139) | | | 0.0060 (0.0139) |
| ind7 | | | 0.0140 (0.0109) | | | 0.0139 (0.0109) |
| ind8 | | | 0.0085 (0.0111) | | | 0.0084 (0.0111) |
| ind9 | | | 0.0017 (0.0096) | | | 0.0016 (0.0097) |
| ind10 | | | -0.0038 (0.0304) | | | -0.0036 (0.0304) |
| lHbar | -0.0594 (0.2618) | | | -0.0589 (0.0655) | | |
| B | -0.0053 (0.0093) | | | -0.0054 (0.0076) | | |
| B1 | 0.0006 (0.0078) | | | 0.0006 (0.0065) | | |
| B2 | -0.0182 (0.0102) | | | -0.0181 (0.0073) | | |
| inv | 0.0001 (0.0000) | | | 0.0001 (0.0000) | | |
| lconc | -0.0045 (0.0028) | | | -0.0047 (0.0012) | | |
| size1 | -0.0811 (0.0226) | | | -0.0802 (0.0148) | | |
| size2 | -0.0702 (0.0179) | | | -0.0693 (0.0127) | | |
| size3 | -0.0567 (0.0167) | | | -0.0564 (0.0112) | | |
| size4 | -0.0453 (0.0142) | | | -0.0450 (0.0104) | | |
| size5 | -0.0322 (0.0120) | | | -0.0321 (0.0090) | | |
| size6 | -0.0220 (0.0107) | | | -0.0219 (0.0079) | | |
| size7 | -0.0112 (0.0101) | | | -0.0111 (0.0074) | | |
| size8 | -0.0030 (0.0088) | | | -0.0030 (0.0067) | | |
| size9 | 0.0021 (0.0079) | | | 0.0021 (0.0058) | | |
| profit2 | -0.0019 (0.0043) | | | -0.0019 (0.0029) | | |
| profit3 | -0.0072 (0.0047) | | | -0.0072 (0.0034) | | |
| profit4 | -0.0081 (0.0052) | | | -0.0081 (0.0035) | | |
| profit5 | -0.0098 (0.0062) | | | -0.0097 (0.0041) | | |
| vin | -0.0029 (0.0021) | | | -0.0031 (0.0017) | | |
| vinsq | 0.0001 (0.0001) | | | 0.0001 (0.0001) | | |
| cvin | -0.0003 (0.0020) | | | -0.0005 (0.0014) | | |
| cvin2 | 0.0000 (0.0001) | | | 0.0000 (0.0001) | | |
| year2 | -0.0047 (0.0048) | | | -0.0046 (0.0019) | | |
| year3 | 0.0124 (0.0036) | | | 0.0123 (0.0027) | | |
| year4 | 0.0060 (0.0023) | | | 0.0060 (0.0015) | | |
| year5 | | | | | | |
| cons | | 7.7096 (0.0057) | -0.0066 (0.0103) | | 7.7092 (0.0057) | -0.0064 (0.0104) |
| No. of obs. | 5,145,098 | 5,145,098 | 5,145,098 | 5,145,098 | 5,145,098 | 5,145,098 |
| No. of workers | 1,930,260 | 1,930,260 | 1,930,260 | 1,930,260 | 1,930,260 | 1,930,260 |
| No. of plants | 212 | 212 | 212 | 212 | 212 | 212 |
| $\sigma_\epsilon$ | 0.0675 | 0.3145 | 0.0501 | not app | 0.3105 | 0.0463 |

*6 urbanicity dummies also included (not reported to save space). For the regressions we report robust standard errors adjusted for clustering on plants.

# Figures

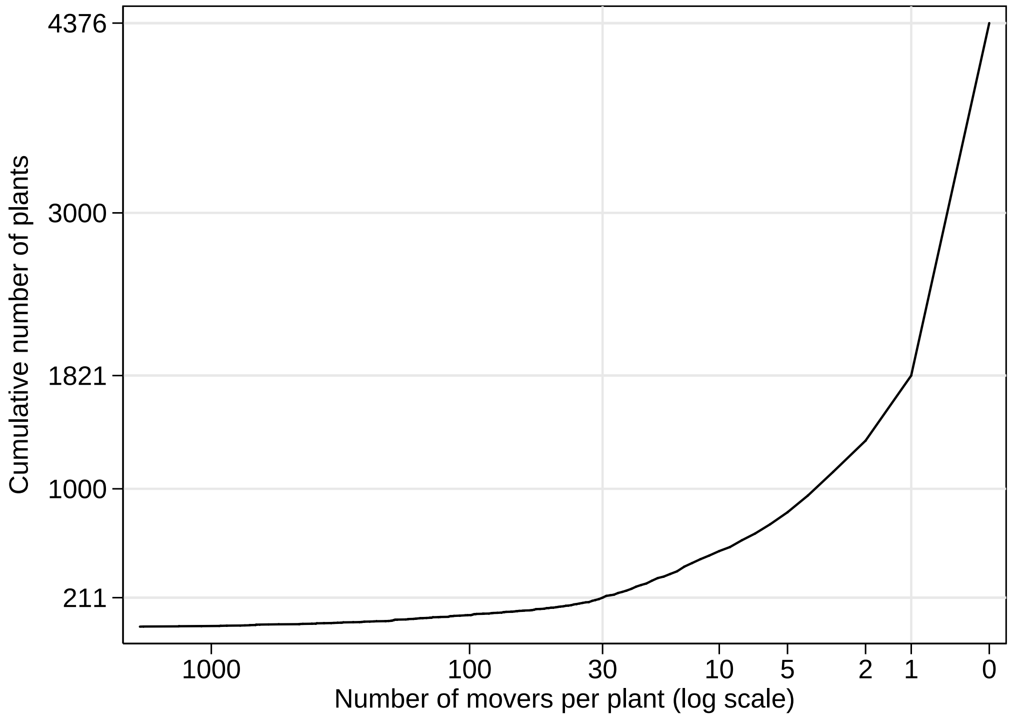


Figure 1: Distribution function of numbers of movers for each plant

## In dieser Reihe sind zuletzt erschienen

## Recently published

| No. | Author(s) | Title | Date |
|---|---|---|---|
| 1/2004 | Bauer, Th. K., Bender, St., Bonin, H. | Dismissal Protection and Worker Flows in Small Establishments | 7/2004 |
| 2/2004 | Achatz, J., Gartner, H., Glück, T. | Bonus oder Bias? Mechanismen geschlechts-spezifischer Entlohnung | 7/2004 |

## Impressum

**IABDiscussionPaper**
**No. 3 / 2004**

**Herausgeber**
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit
Weddigenstr. 20-22
D-90478 Nürnberg

**Redaktion**
Regina Stoll, Jutta Palm-Nowak

**Technische Herstellung**
Jutta Sebald





**Bezugsmöglichkeit**
Volltext-Download dieses DiscussionPaper unter:
http://doku.iab.de/discussionpapers/2004/dp0304.pdf

**IAB im Internet**
http://www.iab.de