

# High wage workers and low wage firms: Negative assortative matching or statistical artefact?\*

**M J Andrews**

University of Manchester

**T Schank**

Universität Erlangen-Nürnberg

Institut für Arbeitsmarkt und Berufsforschung

**R Upward**

University of Nottingham

28th October, 2004

## **Abstract**

The empirical literature on the estimation of firm and worker heterogeneity using linked employer-employee data has thrown up a puzzle. Unobserved worker quality appears to be *negatively* correlated with unobserved firm quality. Following a suggestion made by Barth & Dale-Olsen (2003), we investigate the possibility that this is simply the result of the statistical model used, and is caused by sampling error. We develop formulae that show that the estimated correlation is downwards biased if there is true positive assortative

---

\*The authors thank the IAB (Institut für Arbeitsmarkt und Berufsforschung, Nürnberg) for funding this research, in particular, Stephan Bender and Lutz Bellmann. The views expressed in this paper are solely those of the authors and are not those of the IAB. The comments of Len Gill and participants at various presentations are gratefully acknowledged. These include the IAB, the Institute of Social and Economic Research at Essex, the 2004 Annual Conference of the European Association of Labour Economists (Lisbon), and the Departments of Economics at Erlangen-Nürnberg, Manchester and Warwick. The usual disclaimer applies. All calculations were performed using Stata 8/SE and all code is available on request.

matching. These formulae can be used to sign the magnitude of the bias. We also simulate a data generation process which exhibits positive assortative matching, and we show that standard estimation methods do indeed yield biased estimates of the correlation between worker and firm effects.

# 1 Introduction

There is a rapidly-growing empirical literature which uses linked employer-employee data to estimate the contribution of worker and firm heterogeneity to outcomes in the labour market. Much of this literature stems from Abowd, Kramarz & Margolis (1999) (henceforth AKM) and related papers.<sup>1</sup> An important issue in the literature is the relationship between the unobserved worker- and firm-components of wages. Models of assignment imply positive assortative matching and therefore a positive correlation between worker and firm productivities. In the words of AKM: “high-wage workers and high-wage firms” go together.

However, 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 between  $\theta_i$  and  $\psi_j$ , all subsequent work has reported negative correlations. Abowd, Creedy & Kramarz (2002) 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. Abowd, Creedy & Kramarz 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 (Andrews, Schank & Upward 2004) suggest a correlation of approximately zero.

In other words, when focussing on unobserved components, low wage workers work in high wage firms, and *vice versa*. This seems counter-intuitive both in the light of theories of assortative matching. Following a suggestion made by Barth & Dale-Olsen (2003), we investigate the possibility that the observed negative correlation is simply the result of the statistical method used. To do this, we simulate a data generation process which creates an artificial linked employer-employee dataset which exhibits positive assortative matching. We implement the standard method for estimating the parameters of the model, including both unobserved components of wages. We then demonstrate that for many reasonable parameter values and simulation designs the estimated correlation between the worker and firm unobservables are severely downwards biased. It is therefore possible that all the negative estimates obtained

---

<sup>1</sup>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.

thus far in the literature are consistent with positive assortative matching.<sup>2</sup>

The structure of the paper is as follows. In Section 2 we outline a generic model where wages are a function of observed and unobserved worker and firm characteristics. In Section 5 we describe the design of the simulation and in Section 3 we explain the methods used to estimate the parameters of the underlying model. Section 6 presents our results and Section 7 concludes.

## 2 The generic model

Consider a model of wages with both employer and employee heterogeneity and employer and employee covariates:<sup>3</sup>

$$y_{it} = \mu + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{jt}\boldsymbol{\gamma} + \mathbf{u}_i\boldsymbol{\eta} + \mathbf{q}_j\boldsymbol{\rho} + \alpha_i + \phi_j + \varepsilon_{it}. \quad (1)$$

There are  $i = 1, \dots, N$  individuals and  $j = 1, \dots, J$  firms.  $y_{it}$  is the dependent variable (in this case wages);  $\mathbf{x}_{it}$  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.  $\alpha_i$  and  $\phi_j$  are (scalar) unobserved heterogeneities. It is usual to assume that both are correlated with the observable components of wages. Models of positive assortative matching would also imply that they are positively correlated with each other. Note that both  $\alpha_i$  and  $\mathbf{u}_i$  are variables that are time-invariant for individuals. 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  $t$ .<sup>4</sup> Equation (1) therefore contains all four possible types of information which a researcher might have about workers and firms.

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

---

<sup>2</sup>Abowd, Kramarz, Lengermann & Perez-Duarte (2004) also investigate this issue, but focus mainly on economic explanations.

<sup>3</sup>Wherever possible we use AKM's notation, although we explicitly define firm-level and worker-level time-varying covariates ( $\mathbf{w}_{jt}$  as well as  $\mathbf{x}_{it}$ ).

<sup>4</sup>The notation  $\mathbf{w}_{jt}$  and  $\mathbf{q}_j$  is possibly confusing, since both are defined over every row indexed  $it$ . 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.

Indeed, positive assortative matching *requires* that worker mobility is non-random with respect to  $\alpha_i$  and  $\phi_j$ .

As shown by AKM, in the presence of any correlations across the two sides of the market, that is correlations between unobserved/observed worker characteristics and unobserved/observed firm characteristics, there will be obvious biases which arise when estimating Equation (1) using data from only one side of the market. It is usual to assume that the heterogeneity terms  $\alpha_i$  and  $\phi_j$  are correlated with the observables from the same side of the market. This means that random effects methods are inconsistent, and so fixed effects methods are needed to estimate the parameters of interest. This 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

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

Estimates of  $[\boldsymbol{\eta}, \boldsymbol{\rho}]$  can be recovered by making the additional random effects assumptions  $\text{Cov}(\mathbf{u}_i, \alpha_i) = \text{Cov}(\mathbf{q}_j, \phi_j) = 0$  (as AKM do). However, some may be unhappy identifying time-varying effects using fixed-effects methods whilst identifying non-time-varying effects using random-effects methods in the same regression (in the spirit of Hausman & Taylor (1981)), 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.

Equation (4) is the generic model that represents most of the existing literature. The particular focus of this paper is on the estimation of the worker and firm fixed effects,  $\theta_i$  and  $\psi_j$ , and their correlation with each other.

### 3 Estimation

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, and so differences or deviations removes both terms. However, we are interested in the

estimates of  $\theta_i$  and  $\psi_j$  themselves, so this solution is not useful because it allows us to recover only the sum  $\theta_i + \psi_j$  after estimation, and not the individual components (see AKM). It is worth noting, however, that for many researchers this “spell fixed effects” (Spell FE) method is a practical and simple solution which does not present any computational difficulty.

As noted by AKM, the Least Squares Dummy Variable (LSDV) estimator of Equation (1) requires the estimation of  $N$  individual effects and (approximately)  $J$  firm effects.  $N$  is often in the order of millions, and  $J$  is often in the order of thousands, or tens of thousands. For most realistic values of  $N$  and  $J$  this is not a practical solution. In standard linear panel data models the LSDV estimator gives identical results to models where the heterogeneity is removed algebraically, by taking deviations from the mean of all variables in Equation (4). However, there appears to be no algebraic transformation of the observables that sweeps away both terms, nor which allows them to be recovered subsequently. This is because of the lack of patterning between workers and the firms they work for.<sup>5</sup>

To circumvent this 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.

More precisely, generate a dummy variable for each firm:

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

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

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

into Equation (4). The  $\theta_i$  are removed by time-demeaning (or differencing):

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (\mathbf{w}_{jt} - \bar{\mathbf{w}}_i)\boldsymbol{\gamma} + \sum_{j=1}^J \psi_j (F_{it}^j - \bar{F}_i^j) + \varepsilon_{it}, \quad (5)$$

where  $\bar{z}_i = (\sum_t z_{it})/T_i$  for any variable  $z$ .<sup>6</sup> To distinguish this estimator from the standard LSDV estimator, hereafter we label this estimator as “FEiLSDVj”. They

---

<sup>5</sup>More precisely, sort the data by individuals, and the firm dummies are unpatterned; sort the data by firms, and the individual dummies are unpatterned.

<sup>6</sup>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 under the assumptions of our model. See Wooldridge (2002, Section 10.6.3).

are identical estimators, but differ in how they are computed. The covariance matrix for FEiLSDVj needs the standard degrees-of-freedom adjustment.

To obtain estimates of the heterogeneity, first compute

$$\hat{\psi}_{j(it)} = \sum_{j=1}^J \hat{\psi}_j F_{it}^j \quad (6)$$

and then

$$\hat{\theta}_i = \bar{y}_i - \bar{\hat{\psi}}_i - \bar{\mathbf{x}}_i \hat{\boldsymbol{\beta}} - \bar{\mathbf{w}}_i \hat{\boldsymbol{\gamma}} \quad (7)$$

where  $\bar{\hat{\psi}}_i$  averages  $\hat{\psi}_{j(it)}$  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. 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. See Andrews et al. (2004) for fuller details.

An important issue is establishing how many unique unobserved firm effects can be identified. First, effects cannot be identified for firms which have no turnover; otherwise  $F_{it}^j - \bar{F}_i^j = 0$ . 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 Abowd, Creecy & Kramarz (2002). 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  $G$  separate groups of firms, then it is not possible to identify one firm per group for the reason above.

A second implication of the grouping of firms is that estimates of  $\hat{\psi}_j$  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  $\hat{\theta}_i$ . Abowd, Creecy & Kramarz suggest making the additional assumption that the average firm effect is the same across groups.

Equation (7) gives the intuition as to why there is an observed negative correlation between  $\hat{\theta}$  and  $\hat{\psi}$  (as noted by Barth & Dale-Olsen (2003) and Abowd et al. (2004)). The  $\psi_j$  are estimated by LSDV, and are subject to the usual sampling variation (the firm dummies are no different from any other observed covariate). Once estimated, each  $\hat{\psi}_j$  generates a number of  $\hat{\theta}_i$ , via Equation (7). If  $\psi_j$  is over-estimated, then, on average, the corresponding  $\theta_i$  are under-estimated, and *vice versa*. This implies that the estimated correlation between  $\theta_j$  and  $\psi_j$  is biased downwards. Since the sampling variation is greater for plants with small numbers of movers, the bias is worse for these plants. An expression for this bias is formulated in the next section.

## 4 The bias

Equation (4) is the generic model that characterises the literature which uses linked employer-employee data. To keep the notation simple, in what follows, we also drop the two vectors of observed covariates. It is also useful to write the model in matrix notation:

$$\mathbf{y} = \mathbf{D}\boldsymbol{\theta} + \mathbf{F}\boldsymbol{\psi} + \boldsymbol{\varepsilon} \quad (8)$$

where  $\mathbf{y}$  and  $\boldsymbol{\varepsilon}$  are  $N^* \times 1$  vectors,  $\mathbf{D}$  is a  $N^* \times N$  matrix of individual dummies,  $\mathbf{F}$  is a  $N^* \times J$  matrix of firm dummies,  $\boldsymbol{\theta}$  is a  $N \times 1$  parameter vector, and  $\boldsymbol{\psi}$  is a  $J \times 1$  parameter vector. In this section, to keep things simple, we assume we have a balanced panel,  $N^* \equiv NT$ .

We now assume that the model can be estimated by LSDV. This means that any unidentified firm effects have been dropped, and  $J$  redefined accordingly. This avoids use of generalised inverses for expressions involving  $\mathbf{F}'\mathbf{F}$ . After estimation, one computes the sample variance over all  $N$  estimates of  $\theta_i$ , the sample variance over all  $J$  estimates of  $\psi_j$  (where identified), and the covariance between these two unobserved components:

$$\begin{aligned}\text{EstVar}(\hat{\theta}) &= \frac{1}{N^* - 1} \sum_{it} (\hat{\theta}_i - \bar{\hat{\theta}})^2 \\ \text{EstVar}(\hat{\psi}) &= \frac{1}{N^* - 1} \sum_{it} (\hat{\psi}_j - \bar{\hat{\psi}})^2 \\ \text{EstCov}(\hat{\theta}, \hat{\psi}) &= \frac{1}{N^* - 1} \sum_{it} (\hat{\theta}_i - \bar{\hat{\theta}})(\hat{\psi}_j - \bar{\hat{\psi}}),\end{aligned}$$

where  $\hat{\theta}_i$  is the  $it$ -th row of  $\mathbf{D}\hat{\theta}$  and  $\hat{\psi}_j$  is the  $it$ -th row of  $\mathbf{F}\hat{\psi}$ .  $\bar{\hat{\theta}}$  averages  $\hat{\theta}_i$  over all of individual  $i$ 's observations and similarly  $\bar{\hat{\psi}}$  averages  $\hat{\psi}_j$  over all of firm  $j$ 's observations. Notice that each of  $\text{EstVar}(\hat{\theta})$ ,  $\text{EstVar}(\hat{\psi})$  and  $\text{EstCov}(\hat{\theta}, \hat{\psi})$  is computed over  $N^*$  observations, that is, a given  $\hat{\theta}_i$  is summed over  $N$  times and a given  $\hat{\psi}_j$  is summed over for as many worker-periods the firm is observed in the data. These could be computed over  $N$  individual-level observations or  $J$  firm-level observations, with appropriate weighted averages being used, but we do not develop these formulae here.

We write the three estimated components of interest as

$$\text{EstVar}(\hat{\theta}) = \hat{\theta}'\mathbf{D}'\mathbf{A}_{N^*}\mathbf{D}\hat{\theta} \quad \text{EstVar}(\hat{\psi}) = \hat{\psi}'\mathbf{F}'\mathbf{A}_{N^*}\mathbf{F}\hat{\psi} \quad \text{EstCov}(\hat{\theta}, \hat{\psi}) = \hat{\theta}'\mathbf{D}'\mathbf{A}_{N^*}\mathbf{F}\hat{\psi},$$

where  $\mathbf{A}_{N^*} \equiv \mathbf{I}_{N^*} - \mathbf{i}_{N^*}\mathbf{i}_{N^*}'$  and  $\mathbf{i}_{N^*}$  is a  $N^* \times 1$  vector of ones.

The vectors  $\hat{\theta}$  and  $\hat{\psi}$  suffer standard least-squares estimation error, and so we compare the means of the sampling distributions of these three components with their true values (algebra available on request)

$$\begin{aligned}\text{E}[\text{EstVar}(\hat{\theta})] &= \frac{\text{E}(\hat{\theta}'\mathbf{D}'\mathbf{A}_{N^*}\mathbf{D}\hat{\theta})}{N^* - 1} = \frac{\theta'\mathbf{D}'\mathbf{A}_{N^*}\mathbf{D}\theta}{N^* - 1} \\ &\quad + \frac{\sigma_\varepsilon^2}{N^* - 1} \text{tr} [(\mathbf{F}'\mathbf{M}_D\mathbf{F})^{-1}\mathbf{F}'\mathbf{A}_{N^*}\mathbf{F}] + \frac{N - 1}{N^* - 1}\sigma_\varepsilon^2\end{aligned}$$

$$\begin{aligned}\text{E}[\text{EstVar}(\hat{\psi})] &= \frac{\text{E}(\hat{\psi}'\mathbf{F}'\mathbf{A}_{N^*}\mathbf{F}\hat{\psi})}{N^* - 1} = \frac{\psi'\mathbf{F}'\mathbf{A}_{N^*}\mathbf{F}\psi}{N^* - 1} \\ &\quad + \frac{\sigma_\varepsilon^2}{N^* - 1} \text{tr} [(\mathbf{F}'\mathbf{M}_D\mathbf{F})^{-1}\mathbf{F}'\mathbf{A}_{N^*}\mathbf{F}]\end{aligned}$$

$$\begin{aligned}\text{E}[\text{EstCov}(\hat{\theta}, \hat{\psi})] &= \frac{\text{E}(\hat{\theta}'\mathbf{D}'\mathbf{A}_{N^*}\mathbf{F}\hat{\psi})}{N^* - 1} = \frac{\theta'\mathbf{D}'\mathbf{A}_{N^*}\mathbf{F}\psi}{N^* - 1} \\ &\quad - \frac{\sigma_\varepsilon^2}{N^* - 1} \text{tr} [(\mathbf{F}'\mathbf{M}_D\mathbf{F})^{-1}\mathbf{F}'\mathbf{A}_{N^*}\mathbf{P}_D\mathbf{F}]\end{aligned}$$

where  $\mathbf{P}_D \equiv \mathbf{D}'(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}$  and  $\mathbf{M}_D \equiv \mathbf{I} - \mathbf{P}_D$ .

The bias in estimating these three components are given by the terms involving  $\sigma_\varepsilon^2/(N^* - 1)$ . All three biases are unambiguously signed.

Two points are worth noting. First, as expected, (see the end of the last section), both  $\text{EstVar}(\hat{\theta})$  and  $\text{EstVar}(\hat{\psi})$  are overestimated whereas  $\text{EstCov}(\hat{\theta}, \hat{\psi})$  is underestimated. This means that if the true covariance is positive, that is, there is positive assortative matching, the estimated correlation will always be too small, and could be negative. On the other hand, if the true covariance is negative, the estimated correlation could either be more or less negative.

Second, the three biases, in absolute terms, are a (complicated) decreasing function of the number of movers between firms. Intuitively, this is because  $\text{tr}(\mathbf{F}'\mathbf{M}_D\mathbf{F})$  increases as the number of movers increases, and the expression  $(\mathbf{F}'\mathbf{M}_D\mathbf{F})^{-1}$  is found in each of the three biases. Similarly, as the number of movers increases, both  $\text{tr}(\mathbf{F}'\mathbf{A}_{N^*}\mathbf{F})$  and  $\text{tr}(\mathbf{F}'\mathbf{A}_{N^*}\mathbf{P}_D\mathbf{F})$  decrease. However, this argument is only intuitive, as  $\text{tr}(\mathbf{B}^{-1}\mathbf{A}) \neq \text{tr}(\mathbf{A})/\text{tr}(\mathbf{B})$ .

The usefulness of having expressions for the bias in the three components of the estimated correlation between  $\hat{\psi}$  and  $\hat{\theta}$ ,

$$\text{EstCorr}(\hat{\theta}, \hat{\psi}) = \frac{\hat{\theta}'\mathbf{D}'\mathbf{A}_{N^*}\mathbf{F}\hat{\psi}}{\sqrt{\hat{\psi}'\mathbf{F}'\mathbf{A}_{N^*}\mathbf{F}\hat{\psi}}\sqrt{\hat{\theta}'\mathbf{D}'\mathbf{A}_{N^*}\mathbf{D}\hat{\theta}}},$$

means that one can assess the extent to which the true correlation between the estimated unobserved worker and firms effects differs, *on average*, from the corresponding estimate. One does this by adjusting the estimates of the three components by using estimates of the bias, and recomputing the correlation. The only potential computational problem with this is that each trace involves inverting  $J \times J$  matrices; as the software has already computed  $(\mathbf{F}'\mathbf{M}_D\mathbf{F})^{-1}$  to produce LSDV estimates, the number of firms presents no further computational problems.<sup>7</sup>

To conclude, any investigator who computes  $\text{EstCorr}(\hat{\theta}, \hat{\psi})$  can also compute these biases and therefore recompute the correlation to see how biased the original correlation is. The size of the bias depends on the number of movers to and from each firm, a property of the matrix  $\mathbf{F}'\mathbf{D}$ , and the variance of the idiosyncratic error,  $\sigma_\varepsilon^2$ . In the next section, we use simulated data to show how large these biases can be for datasets that are amenable to estimation by the estimation methods discussed

---

<sup>7</sup>Stata has commands for computing inner products of matrices that do not involve storing matrices with  $N^*$  rows or columns.

here. In particular, we attempt to uncover the non-linear relationship that links the bias in the correlation (or its components) to the number of movers and  $\sigma_\varepsilon^2$ . Investigators who have very large datasets, and therefore must use ACK’s genetic algorithms, should also be able to compute these biases using the formula reported in this section.<sup>8</sup>

## 5 The simulation design

The simulated data mimics the generic model outlined in Section 2.  $J$  firms are created indexed  $j = 1, \dots, J$ , each with a random number of employees drawn from a Uniform distribution. Each firm is given a realisation of  $w_{jt}$  and  $\psi_j$ ; each worker is given a realisation of  $x_{it}$  and  $\theta_i$ .<sup>9</sup> These realisations are drawn from a joint Normal distribution with the following means and covariance structure for any period  $t$ :

$$\begin{bmatrix} \psi_j \\ w_{jt} \\ \theta_i \\ x_{it} \end{bmatrix} \sim N \begin{bmatrix} 0 & \sigma_\psi^2 & & & \\ 0 & \sigma_{w\psi} & \sigma_w^2 & & \\ 0 & \sigma_{\theta\psi} & \sigma_{\theta w} & \sigma_\theta^2 & \\ 0 & \sigma_{x\psi} & 0 & \sigma_{x\theta} & \sigma_x^2 \end{bmatrix} \quad (9)$$

The structure above focuses on the correlation between the unobservables and the observables, and the correlation between the unobservables themselves.<sup>10</sup> We assume that the observed firm and worker effects ( $w_{jt}$  and  $x_{it}$ ) are uncorrelated with each other, but we allow for non-zero covariance between the unobserved components ( $\sigma_{\theta\psi} \neq 0$ ), as well as between the unobserved components and both firm and worker time-varying effects.

The draw of  $[\psi_j, w_{jt}, \theta_i, x_{it}]$  initially ensures that workers with certain characteristics are matched with firms with certain characteristics. For example, if  $\sigma_{\theta\psi} > 0$  then high wage workers tend, on average, to be matched with high wage firms. This gives the distribution of workers across firms in period  $t = 1$ .

We now generate the movement of workers between firms. As noted, this is crucial for the identification of the fixed effects. For each worker we draw a potential new

---

<sup>8</sup>This is work in progress: these formula need developing for covariates and unbalanced panels, however.

<sup>9</sup>We use one variable of each type, hence  $w_{jt}$  and  $x_{it}$  are scalars rather than vectors as in Equation (4).

<sup>10</sup>For clarity, we write out the correlation structure at time  $t$ . In addition, there are correlations across periods. Both variables  $x_{it}$  and  $w_{jt}$  are autocorrelated, with parameter 0.9. All  $x_{it}$  and  $w_{jt}$  pairs are uncorrelated.

firm  $j'$  from the list of currently existing firms. This new firm has its own set of characteristics  $[\psi_{j'}, w_{j't}]$ .<sup>11</sup>

The probability of movement from  $j$  to  $j'$ , denoted  $m_{it}^*$ , is determined by one of three rules.

$$m_{it}^* = u \tag{10}$$

$$m_{it}^* = a(\theta_i - \psi_j)^2 + u \tag{11}$$

$$m_{it}^* = a[(\theta_i - \psi_j)^2 - (\theta_i - \psi_{j'})^2] + u \tag{12}$$

where  $u$  is a realisation from  $U \sim N(0,1)$  and the parameter  $a$  is chosen to affect the correlation structure of  $\theta$  and  $\psi$ .

In (10) the probability of movement is a random draw from a Normal distribution.

In (11) the probability of movement is increasing in the distance between  $\theta$  and  $\psi$ . This is intended to capture the notion of matches being pure “experience” goods (Jovanovic 1979). The quality of the worker  $\theta_i$  is not observable to the firm until the match is made. Similarly,  $\psi_j$  is not observable to the worker until a match is made. Once a match occurs, both are observable, and the worker and firm decide whether or not to separate at the end of the period. If they do separate, the new partner is once again entirely random.

In (12) the probability of movement depends on the quality of the potential new firm,  $\psi'$ , relative to the current match. This is intended to capture the notion of matches being “search” goods.  $\psi'$  can be observed in potential new matches, and it is the arrival of this new information which causes current matches to dissolve.

A move occurs if  $m^*$  is greater than some critical percentile of the distribution of  $m^*$ . Altering this percentile allows us to alter the number of workers who move each period. If a move occurs, the value of  $j'$  is copied to  $j$  in that period and for all future periods, as are  $\psi_{j'}$ ,  $q_{j'}$  and  $w_{j't}$ . The potential matching of workers and firms occurs once per period  $t$ . The number of periods  $T$  can be varied to mimic real data. Typically  $T$  is small because linked data are recorded annually, and have become available only recently.

Once the identity of each firm is established for every individual in all  $T$  rows of the data, the dependent variable  $y_{it}$  is generated according to Equation (4). The resulting dataset is balanced for individuals, unlike real data. All individuals appear

---

<sup>11</sup>In order to ensure that a new match is drawn with a probability proportional to firm size, the list of new firms is weighted by the size of the firm.

$T$  times. It is not however necessarily balanced in terms of firms, because small firms who experience worker exits may disappear.

## 6 Results

### 6.1 Baseline simulation

We now repeatedly generate a synthetic dataset using the methods outlined in Section 5. Table 1 reports the baseline values chosen for the synthetic data. Choices of parameters were made with reference to the linked employer-employee data used in Andrews et al. (2004).

Table 1: Baseline parameter values and realisations: random mobility

|  | <i>Population</i> | <i>Realisation (100 reps.)</i> |             |
|--|-------------------|--------------------------------|-------------|
|  |                   | <i>Mean</i>                    | <i>S.D.</i> |
| Number of firms $J$                                    | 100               | 100                            | —           |
| Number of time periods $T$                             | 5                 | 5                              | —           |
| Average number of workers per firm $\bar{N}_j$         | 50                | 50.096                         | 2.761       |
| Total number of observations $N^*$                     | 25000             | 25029.800                      | 1382.385    |
| Probability of movement per period $m^*$               | 0.1               | 0.100                          | 0.002       |
| Total number of groups $G$                             |                   | 1.74                           | 0.848       |
| Number of observations in largest group                |                   | 25023.350                      | 1385.184    |
| $\beta$  | 0.5               | 0.5                            | —           |
| $\gamma$   | 0.3               | 0.3                            | —           |
| Variance of worker effects $\sigma_\theta^2$           | 0.3               | 0.309                          | 0.008       |
| Variance of firm effects $\sigma_\psi^2$               | 0.3               | 0.298                          | 0.049       |
| Variance of idiosyncratic error $\sigma_\varepsilon^2$ | 1                 | 0.999                          | 0.008       |
| $\text{Corr}(\theta, \psi)$                            | 0.3               | 0.239                          | 0.024       |
| $\text{Corr}(\theta, x)$                               | 0.3               | 0.295                          | 0.012       |
| $\text{Corr}(\theta, w)$                               | 0.2               | 0.160                          | 0.026       |
| $\text{Corr}(\psi, x)$                                 | 0.1               | 0.082                          | 0.014       |
| $\text{Corr}(\psi, w)$                                 | 0.3               | 0.299                          | 0.097       |

The number of workers per firm is drawn randomly from a Uniform distribution, and so varies across simulations, as does the exact number of workers who change firm each period. Each replication involves a completely new set of worker movements from firm to firm, and so the number of groups  $G$  (and hence the number of estimable effects) varies slightly between replications. In about half the replications there is only one group (all workers and firms are connected). Note that the size of the

largest group is only slightly smaller than the total sample size. This is the usual finding in real linked data (Abowd et al. 2002).

The crucial parameter is the correlation between  $\theta$  and  $\psi$ , which is chosen to be positive (0.3): unobservably high wage workers work for unobservably high wage firms. We also assume positive correlation between each unobservable and both time-varying observables. High wage workers work for firms with observably better characteristics, and high wage firms employ workers with observably better characteristics. The latter assumption is supported by much evidence from real linked employer-employee data (see the Introduction).

Note that after generating the data and allowing movement of workers between firms, the resulting average correlation between  $\theta$  and  $\psi$  is significantly lower than the chosen correlation. This is because in the baseline simulation worker mobility is random with respect to *all* the variables in the model, including  $\theta$  and  $\psi$ . In other words, good matches are just as likely to separate as bad matches, and good matches are just as likely to be consummated as bad matches. In period  $t = 1$  the correlation between  $\theta$  and  $\psi$  is approximately 0.3. This falls because workers who move have a zero correlation. For the same reason, the resulting correlation between  $\theta$  and  $w$ , and between  $\psi$  and  $x$  are also lower than their original period  $t = 1$  population values. The average correlations reported are the averages across all  $T$  periods.

For each dataset we estimate Equation (5), and then compute  $\hat{\psi}$  and  $\hat{\theta}$  using Equations (6) and (7). In Table 2 we report the baseline estimation results. We note first of all that the FEiLSDVj method produces consistent estimates of  $\beta$  and  $\gamma$ .

The most striking result is that the resulting estimate of the correlation of the worker and firm effects is significantly downwards biased, with a mean estimate of 0.115 compared to the true value of 0.239. The explanation for this bias was discussed earlier: any sampling variation in the estimates of  $\psi$  lead to the reverse variation in estimates of  $\theta$ . This sampling variation should be greater in firms with less worker turnover, because, as with any fixed effects model, estimates of the unobserved heterogeneity are functions only of the observed characteristics of workers who change firms.

To investigate this we calculate, for each firm, the number of workers who change firm. Call this  $M_j$ . We then divide the data into quintiles ordered by  $M_j$ , and report the correlation of  $\hat{\psi}$  and  $\hat{\theta}$  for each quintile. It is noticeable that the estimated correlation increases with every quintile except the last. The estimated correlation is particularly poor for those firms with the least amount of worker movement.

Table 2: Baseline results, 100 reps., random mobility

|  | <i>Population</i> | <i>Simulation</i> |             |
|--|-------------------|-------------------|-------------|
|  |                   | <i>Mean</i>       | <i>s.d.</i> |
| $\beta$                                      | 0.5               | 0.498             | 0.008       |
| $\gamma$                                     | 0.3               | 0.302             | 0.007       |
| Variance of worker effects $\sigma_\theta^2$ | 0.309             | 0.537             | 0.014       |
| Variance of firm effects $\sigma_\psi^2$     | 0.298             | 0.327             | 0.053       |
| Corr( $\theta, \psi$ )                       | 0.239             | 0.115             | 0.034       |
| Corr( $\theta, x$ )                          | 0.295             | 0.294             | 0.012       |
| Corr( $\theta, w$ )                          | 0.160             | 0.198             | 0.031       |
| Corr( $\psi, x$ )                            | 0.082             | 0.098             | 0.016       |
| Corr( $\psi, w$ )                            | 0.299             | 0.307             | 0.091       |
| Corr( $\psi, \theta$ ) by $M_j$ :            |                   |                   |             |
| Bottom quintile (few movers)                 | 0.240             | 0.039             | 0.061       |
| 2nd  | 0.232             | 0.124             | 0.060       |
| 3rd  | 0.226             | 0.123             | 0.058       |
| 4th  | 0.229             | 0.143             | 0.076       |
| Top quintile (many movers)                   | 0.220             | 0.131             | 0.073       |

## 6.2 Departures from the baseline simulation

We now vary the simulation in single dimensions away from the baseline. This provides more evidence that the bias in the estimated correlation of  $\theta$  and  $\psi$  is a result of sampling variation, but also helps to quantify the extent of the bias conditional on the characteristics of particular data.

1. Varying overall error variance. In Figure 1 we illustrate the effect of increasing the overall error variance of Equation (4). As  $\sigma_\varepsilon^2$  increases the sampling variability of  $\hat{\psi}$  increases, which decreases the estimated correlation of  $\psi$  and  $\theta$ . Note again that the true correlation of  $\theta$  and  $\psi$  is slightly less than the original chosen value of 0.3 because of the assumption of random mobility of workers between firms.
2. Varying the probability of movement. As noted earlier, the precision of the estimates of  $\hat{\psi}$  is determined by the extent of movement between firms. For this reason the probability of a match dissolving ( $m^*$ ) and the size of firms are important determinants of the estimated correlation. In Figure 2 we plot the estimated correlation of  $\psi$  and  $\theta$  as the probability of movement increases. Note that as  $m^*$  increases the “true” correlation decreases because more workers are being separated from their initial match. As predicted, the accuracy of the estimated correlation improves greatly as the amount of mobility increases.

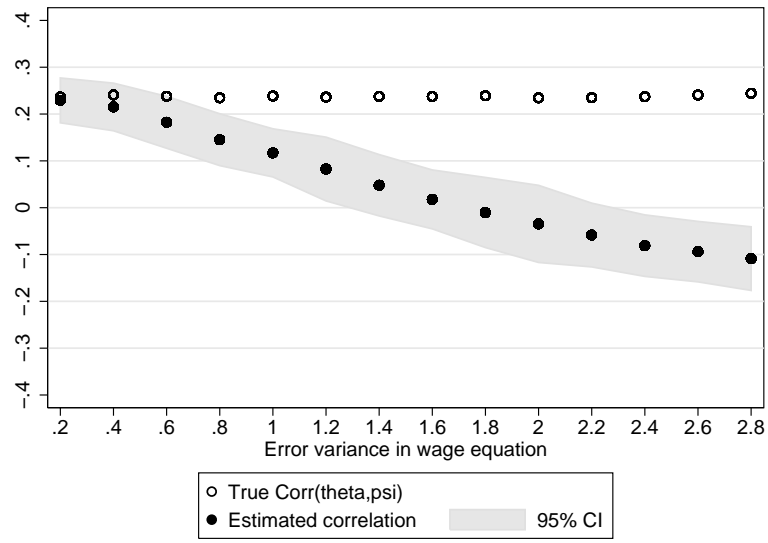


Figure 1: Varying  $\sigma_\epsilon^2$

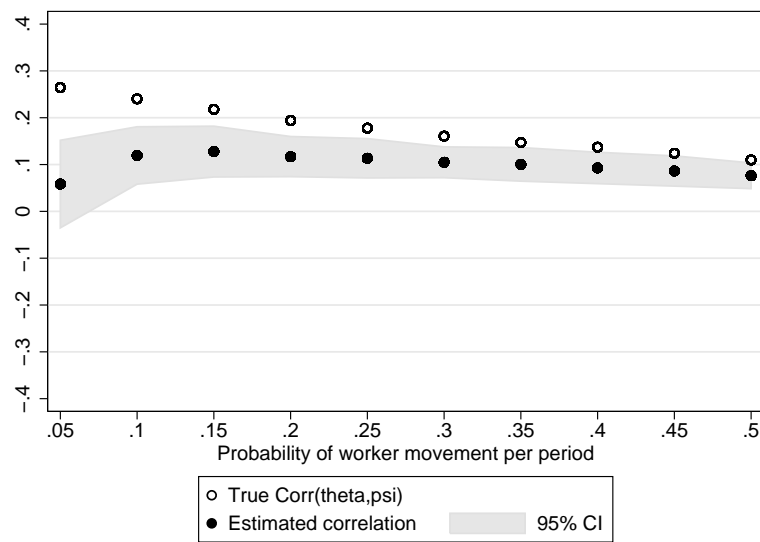


Figure 2: Varying  $m^*$

3. Varying average firm size. Larger firms tend to have more workers joining and leaving them, and therefore provide more accurate estimates of  $\psi$ , leading to more accurate estimates of  $\theta$ . This was consistent with the results shown in Table 2. In Figure 3 we show that as the size of firms increases in the simulated data the estimated correlation of  $\theta$  and  $\psi$  approaches the true value.

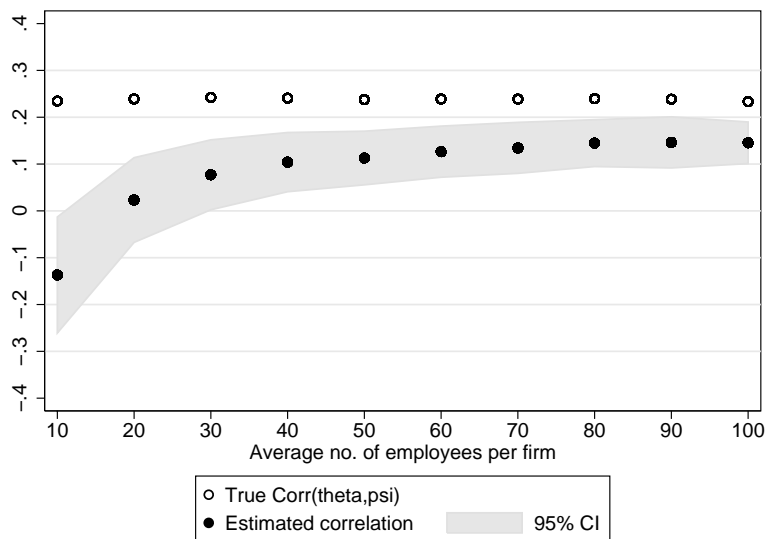


Figure 3: Varying  $\bar{N}_j$

4. Varying the number of time periods. The third dimension over which the number of movers per firm can be increased is simply the length of the panel. The longer the panel, the more accurately  $\psi$  can be identified because, once again, each firm has on average more movers. This is confirmed in Figure 4.
5. Varying the number of firms. In contrast, varying the number of firms has no effect on the bias of the estimated correlation. This is because every new firm requires a new estimated parameter  $\psi$ , and no improvement in sampling variability. Figure 5 illustrates this result.

### 6.3 Non-random mobility

A drawback with the simulation presented thus far is that it is actually logically inconsistent with the notion of positive assortative matching. Because workers and firms match and separate randomly, any positive association between worker and firm characteristics originally imposed by the covariance matrix (9) reduces over time. A better simulation design would allow for non-random mobility determined

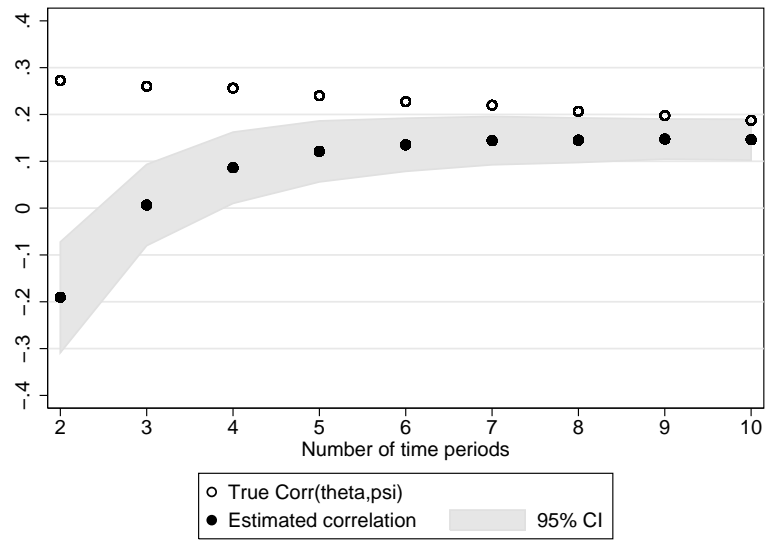


Figure 4: Varying  $T$

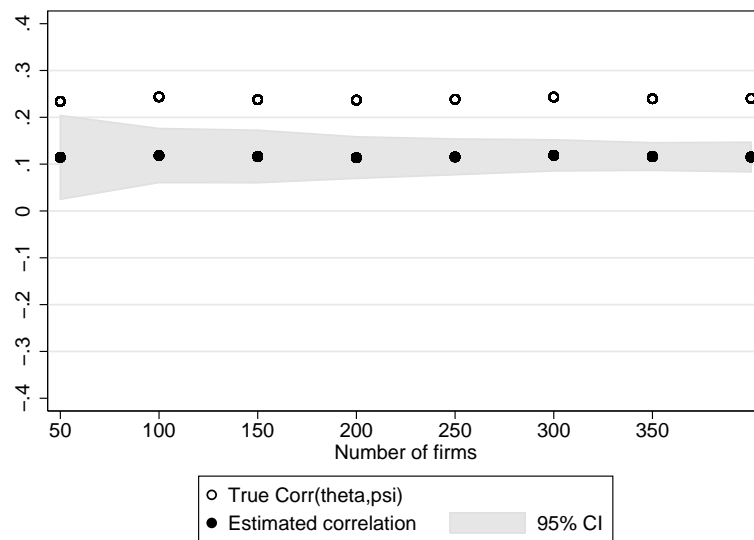


Figure 5: Varying  $J$

either by Equation (11) or (12). In Table 3 we report the baseline results in the case where mobility is determined by Equation (11). That is, where a match is modelled as a pure experience good.

Table 3: Baseline results, 100 reps., non-random (experience) mobility

|  | <i>Population</i> | <i>Simulation</i> |             |
|--|-------------------|-------------------|-------------|
|  |                   | <i>Mean</i>       | <i>s.d.</i> |
| $\beta$                                      | 0.5               | 0.500             | 0.008       |
| $\gamma$                                     | 0.3               | 0.300             | 0.008       |
| Variance of worker effects $\sigma_\theta^2$ | 0.310             | 0.537             | 0.014       |
| Variance of firm effects $\sigma_\psi^2$     | 0.274             | 0.303             | 0.050       |
| Corr( $\theta, \psi$ )                       | 0.360             | 0.196             | 0.037       |
| Corr( $\theta, x$ )                          | 0.297             | 0.225             | 0.022       |
| Corr( $\theta, w$ )                          | 0.183             | 0.139             | 0.033       |
| Corr( $\psi, x$ )                            | 0.117             | 0.112             | 0.015       |
| Corr( $\psi, w$ )                            | 0.298             | 0.285             | 0.092       |
| Corr( $\psi, \theta$ ) by $M_j$ :            |                   |                   |             |
| Bottom quintile (few movers)                 | 0.346             | 0.088             | 0.076       |
| 2nd  | 0.343             | 0.185             | 0.071       |
| 3rd  | 0.334             | 0.201             | 0.066       |
| 4th  | 0.349             | 0.221             | 0.083       |
| Top quintile (many movers)                   | 0.362             | 0.245             | 0.091       |

Because worker mobility is now non-random w.r.t  $\theta$  and  $\psi$ , the resulting correlation between them does not now decrease over time. In fact, setting  $a = 1$  in (11) actually causes  $\text{Corr}(\theta, \psi)$  to increase over time, resulting in an average correlation of 0.360 over 5 periods. For the same reasons,  $\text{Corr}(\theta, w)$  and  $\text{Corr}(\psi, x)$  are also now higher than in the original simulation.

Note that although worker mobility is “non-random”, consistent estimates of  $\beta$  and  $\gamma$  are still achieved, because mobility is still random w.r.t. the idiosyncratic error,  $\varepsilon$ .

As with the baseline simulation, however, estimates of  $\text{Corr}(\theta, \psi)$  are significantly downwards biased, and in fact in absolute terms the bias is larger ( $0.196 - 0.360$ ) compared to  $0.115 - 0.239$ . As before, we split the sample according to the amount of turnover which firms experience, and again the extent of the bias decreases sharply for low turnover firms.<sup>12</sup>

<sup>12</sup>Similar results (not reported here) are also obtained if matches are modelled as search goods as in Equation (12).

## 7 Conclusion

Even in the presence of true positive assortative matching between workers and firms, estimates of the correlation between firm- and worker-effects may actually be negative, or at least strongly downwards biased. The extent of the bias depends on how much worker mobility each firm experiences. We develop formulae for the biases for the components of the estimated correlation. Users of real linked employer-employee data should be able to use these formulae to assess the actual magnitude of this downwards bias.

## 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., Lengermann, P. & Perez-Duarte, S. (2004), Are good workers employed by good firms? A test of a simple assortative matching model for france and the united states, Mimeo, February.
- 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), Practical estimation methods for linked employer-employee data, Discussion Paper No. 29, University of Erlangen-Nürnberg, September.
- 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.
- Goux, D. & Maurin, E. (1999), ‘Persistence of interindustry wage differentials: a reexamination 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.
- Jovanovic, B. (1979), 'Job matching and the theory of turnover', *Journal of Political Economy* **87**, 972–990.
- Wooldridge, J. (2002), *Econometric analysis of cross section and panel data*, MIT Press.